

# Conquering Data Variations in Resolution: A Slice-Aware Multi-Branch Decoder Network

Shuxin Wang, Shilei Cao, Zhizhong Chai, Dong Wei, Kai Ma, Liansheng Wang, *Member, IEEE*,  
and Yefeng Zheng, *Senior Member, IEEE*

**Abstract**—Fully convolutional neural networks have made promising progress in joint liver and liver tumor segmentation. Instead of following the debates over 2D versus 3D networks (for example, pursuing the balance between large-scale 2D pretraining and 3D context), in this paper, we novelly identify the wide variation in the ratio between intra- and inter-slice resolutions as a crucial obstacle to the performance. To tackle the mismatch between the intra- and inter-slice information, we propose a slice-aware 2.5D network that emphasizes extracting discriminative features utilizing not only in-plane semantics but also out-of-plane coherence for each separate slice. Specifically, we present a slice-wise multi-input multi-output architecture to instantiate such a design paradigm, which contains a Multi-Branch Decoder (MD) with a Slice-centric Attention Block (SAB) for learning slice-specific features and a Densely Connected Dice (DCD) loss to regularize the inter-slice predictions to be coherent and continuous. Based on the aforementioned innovations, we achieve state-of-the-art results on the MICCAI 2017 Liver Tumor Segmentation (LiTS) dataset. Besides, we also test our model on the ISBI 2019 Segmentation of Thoracic Organs at Risk (SegTHOR) dataset, and the result proves the robustness and generalizability of the proposed method in other segmentation tasks.

**Index Terms**—Liver and liver tumor segmentation, 2.5D convolutional neural network, slice-aware design, deep learning.

## I. INTRODUCTION

THE liver is a vital organ in the human body as it is essential for bile secretion and detoxifying harmful substances into urea. According to the global cancer statistics reported in 2018 [1], liver cancer is the sixth most frequently diagnosed cancer and the fourth leading cause of cancer death worldwide. The liver is also a common site for other metastatic cancer because of the rich blood supply [2]. In the current clinical routine, CT is the most frequently used

This work was supported by National Natural Science Foundation of China (Grant No. 61671399), the Fundamental Research Funds for the Central Universities (Grant No. 20720190012), the Key Area Research and Development Program of Guangdong Province, China (Grant No. 2018B010111001) and Science and Technology Program of Shenzhen, China (No. ZDSYS201802021814180). (Corresponding authors: Liansheng Wang; Yefeng Zheng.)

Shuxin Wang, Zhizhong Chai and Liansheng Wang are with Department of Computer Science, Fujian Key Laboratory of Sensing and Computing for Smart City, Xiamen University, Siming South Road, Xiamen 361005, China. Shuxin Wang and Zhizhong Chai contributed to this work when they were interns at Tencent (e-mail: shuxin\_icey@163.com, chai\_zhizhong1995@163.com, lswang@xmu.edu.cn).

Shilei Cao, Dong Wei, Kai Ma, and Yefeng Zheng are with Tencent Jarvis Lab, Malata Building, Kejizhongyi Road, Nanshan District, Shenzhen 518075, China (e-mail: eliaslcao@tencent.com, kylekma@tencent.com, donwei@tencent.com, yefengzheng@tencent.com).

Shuxin Wang and Shilei Cao contribute equally.

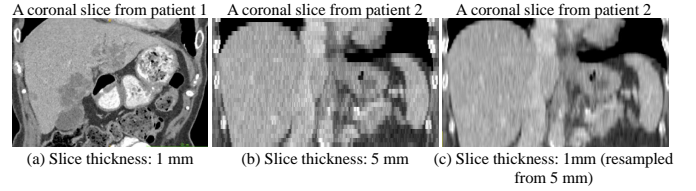


Fig. 1. An illustration of the wide variation in out-of-plane slice thickness with a fixed in-plane pixel spacing. A CT liver image with slice thickness of 1 mm (a); another image with slice thickness of 5 mm (b) and the corresponding resampled image of 1 mm (c). We can observe that, comparing (a) and (b), CT images show apparent visual difference with different slice thicknesses; and comparing (a) and (c), CT images still show apparent visual difference (e.g., blurring) even after being resampled to the same slice thickness.

imaging modality for radiologists and oncologists to make accurate hepatocellular carcinoma evaluation and treatment planning [3]. Nevertheless, outlining the liver and liver tumor in CT slice-by-slice is a time-consuming task and prone to annotator variations. Therefore, a standardized and automatic segmentation method is highly desirable to enable efficient delineation of liver and liver tumor contours in practice.

While remarkable performance on liver and liver tumor segmentation was recently reported with the development of deep learning [4]–[10], a few challenges deserve wide attention of the community. Firstly, to achieve a good generalization of the established model, data are usually collected from various clinical sites. However, due to the variations in equipment manufacturers, physical parameters, scanning protocols, and reconstruction methods, the voxel resolution of CT images from multiple centers suffers from wide variations, especially along the out-of-plane direction. Taking the Liver Tumor Segmentation (LiTS) dataset [11] as an example, its in-plane pixel spacing ranges from 0.56 mm to 1 mm, whereas its out-of-plane slice thickness ranges from 0.45 mm to 6 mm. Secondly, liver and liver tumor segmentation in CT images is a task to assign coherent semantic masks to the full volume, rather than to individual slices, which implies that objects in adjacent slices of a volume usually have intrinsic correlations in terms of context, shape, and location. In conclusion, handling the information asymmetry as a result of the inconsistent in- and out-of-plane resolutions and ensuring the inter-slice segmentation consistency are two critical issues to the liver and liver tumor segmentation.

There were some remarkable deep learning works for liver and liver tumor segmentation, which can be roughly classified as 2D [4], [5], 2.5D [10] and 3D [7], [9] methods. Standard

2D methods can learn abundant deep semantics by employing deeper neural networks with large quantities of training samples. However, losing inter-slice information makes it hard to learn a smooth segmentation map along the out-of-plane direction. On the contrary, 2.5D and 3D methods can exploit 3D context information and thus learn more meaningful feature representations to maintain 3D coherence. However, most of the existing methods underestimate the impact of the differences between the in-plane pixel spacing and out-of-plane slice thickness, and simply resample the input to a fixed in-plane pixel spacing and out-of-plane slice thickness. Although the resampling process can relieve the resolution anisotropy problem, it does not bring extra information to the dimension(s) being finely *interpolated*, while some additional artifacts may be introduced (see Fig. 1). Besides, 3D methods usually suffer from high computational cost and GPU memory consumption, which may hinder application in practice.

In this paper, we propose a novel 2.5D Slice-Aware Multi-Branch Decoder (SAMBD) network which utilizes not only the large-scale 2D pre-training but also 3D contextual information based on the observation of variations in data. SAMBD is focused on learning discriminative slice-specific features, and instantiates a slice-wise multi-input multi-output architecture for this goal. The core components of SAMBD can be summarized in three parts:

(i) **A mutual encoder.** We employ the Xception [12] model with its weights pre-trained on natural images (except for the first layer) to simultaneously extract deep semantics for multiple input slices with discriminative feature initialization, which captures the local volumetric information at different semantic levels. Motivated by DeepLabv3+ [13], an Atrous Spatial Pyramid Pooling (ASPP) [13] module is adopted to fuse features of different scales.

(ii) **A Multi-branch Decoder (MD) with Slice-centric Attention Block (SAB).** Since the multiple input slices are indiscriminately processed by the encoder and the slice-specific information is thus scrambled, we design an MD to explicitly re-establish discriminative features for each slice by fully exploiting the intra- and inter-slice information learned by the encoder. To further strengthen the discriminative power of each slice, we propose and embed an SAB into the MD, which is implemented with the widely adopted attention mechanism [14], [15]. All these designs are centered around learning the best features for each slice, thus avoiding directly processing asymmetric intra- and inter-slice information.

(iii) **A Densely Connected Dice (DCD) loss.** Based on the assumption that target objects lying in successive slices should have consistent labels, we propose a DCD loss to regularize the inter-slice predictions to be more coherent in the label space, where the intra- and inter-slice constraints can be jointly optimized.

In summary, the contributions of this work can be summarized as four-fold:

- Instead of considering the debates over 2D versus 3D networks, in this paper, we identify the wide variation in the ratio between the intra- and inter-slice resolutions as an important obstacle to the performance.

- Observing the variations in data, we propose a 2.5D encoder-decoder network with a multi-input and multi-output structure, featuring a novel slice-aware multi-branch decoder with a slice-centric attention block which not only utilizes the large-scale 2D pre-training but also 3D contextual information for learning discriminative features for each separate slice.
- An auxiliary loss function is proposed to strengthen the inter-slice correlations and regularize the inter-slice predictions to be more coherent.
- We mainly evaluate our method on the CT volumes for liver and liver tumor segmentation provided by LiTS and the result outperforms other methods. Besides, extended validation is conducted on the ISBI 2019 Segmentation of THoracic Organs at Risk (SegTHOR) dataset and the result is competitive.

The remainders of this paper are organized as follows. We review the related work in Section II and elaborate on the proposed method in Section III. We present experiments and results in Section IV, a discussion in Section V, and finally draw the conclusions in Section VI.

## II. RELATED WORKS

### A. Debates over 2D versus 3D Networks

Prior studies chose either 2D networks for the benefits of 2D pretraining and large-scale slice-wise training sets or alternatively 3D networks for native 3D representation learning [16], [17].

Recently, the LiTS challenge was organized to benchmark the performance of different automatic algorithms for liver and liver tumor segmentation, where the top-scoring methods were dominated by Fully Convolutional Networks (FCNs) [18]. (i) **2D/2.5D networks.** Vorontsov *et al.* [5] segmented liver and liver tumor with two FCNs, which were connected in tandem and trained together end-to-end, using a 2D axial slice as input. To capitalize on the complementary information between a few adjacent slices, Han [10] proposed a 2.5D model, which combined the long-range connection of U-Net [19] and the short-range connection of ResNet [20]. Other noteworthy works [4], [21], [22] attempted to use triplanar networks to learn generalized features from the axial, coronal, and sagittal planes. (ii) **3D networks.** There were some works employing 3D convolution to mine 3D context information. For example, Liu *et al.* [7] implemented an improved 3D U-Net equipped with dilated convolutions and separable convolutions to segment livers. Deng *et al.* [9] proposed dynamic regulation of level-set parameters using 3D CNN for liver tumor segmentation. (iii) **Hybrid approaches.** To simultaneously take advantage of the merits of 2D and 3D networks, Li *et al.* [6] proposed a novel hybrid densely connected U-Net named H-DenseUNet, which consists of a 2D DenseUNet for efficiently extracting intra-slice features and a 3D counterpart for hierarchically aggregating volumetric contexts for better liver and tumor segmentation. With similar motivation, Zhang *et al.* [23] proposed a light-weight hybrid convolutional network to segment the liver and liver tumors with an encoder-decoder structure, in which 2D convolutions

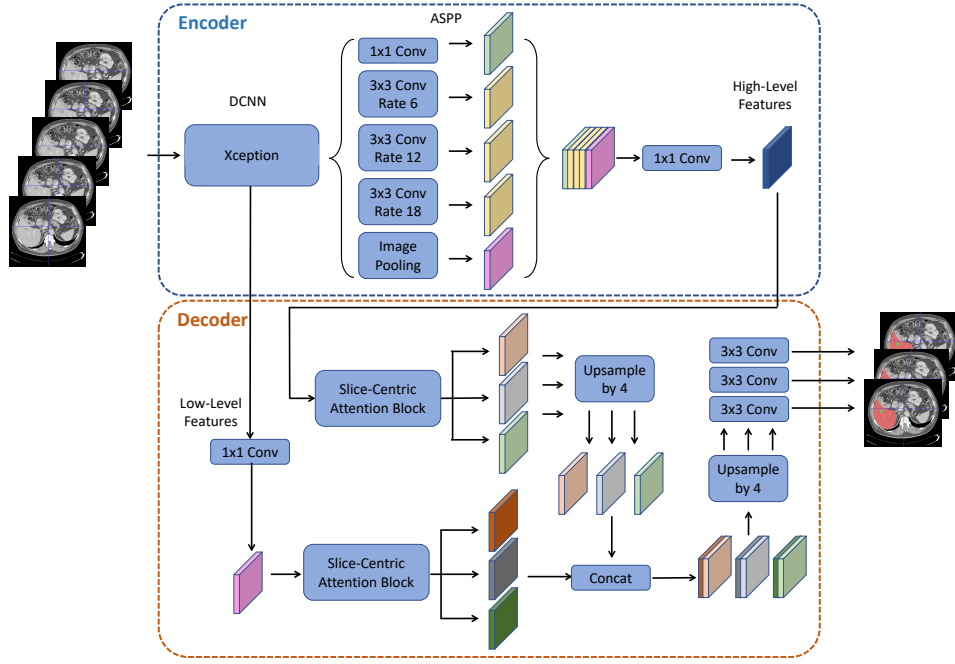


Fig. 2. Schematic view of our proposed framework with a set of five input slices ( $C_{in}=5$ ) and the corresponding three central slice predictions ( $C_{out}=3$ ). A multi-branch decoder with a slice-centric attention block is proposed to gradually and explicitly re-establish discriminative features for each slice by fully exploiting intra- and inter-slice information learned by the encoder.

used at the bottom of the encoder decreases the complexity and 3D convolutions used in other layers explore both in- and out-of-plane information.

Despite the remarkable performance achieved by the aforementioned methods, they underestimated the impact of the inconsistency between the pixel spacing and slice thickness of 3D volumetric data. Our work begins with the basic observation of the data variations which motivates us to extract discriminative features for each slice, based on the fact that the intra-slice information is more coherent due to the uniform in-plane pixel spacing than inter-slice information.

### B. Approaches to Conquering Data Variations in Resolution

Intuitively, resampling to a unified resolution may be a solution to the data variation problem; however, the resampled images suffer from different information densities along the dimensions, and the resampling operation cannot guarantee the validity of the interpolated information as also mentioned in [6].

It is known that thin slice thickness (less than 2.5 mm) results in better performance both for the human reader and computer-aided diagnosis (CAD) systems; however, CT scans with thicker slice thickness (greater than 2.5 mm) are widely used in clinical setting mainly because of the efficiency in terms of reading time and storage [24]. Some works [24], [25] explored to reduce the slice thickness of CT scans from thick to thin. For example, Bae *et al.* [24] proposed a 2.5D image super-resolution (SR) network based on fully residual convolutional neural networks (CNN) for dense slice reconstruction. Ge *et al.* [25] proposed a residual voxel-wise generative adversarial network, which densely reconstructed

slices into a thin thickness (1 mm) and meanwhile denoised the CT images into the more readable pattern, from the widely accessible low-dose thick CT. Although we can adopt such techniques to reconstruct CT images into thin slices of a unified resolution, extra computational costs would be inevitably incurred.

## III. METHODOLOGY

In this section, we present the details of the proposed Slice-Aware Multi-Branch Decoder (SAMBd) network. The network architecture is depicted in Fig. 2. In general, the network design adopts the standard encoder-decoder structure, where the encoder takes a stack of adjacent slices as input and outputs compact feature representations. Meanwhile, the decoder restores the feature maps to the original resolution by fusing features from different levels of the encoder and outputs the label predictions for the central slices. Here, we use  $C_{in}$  and  $C_{out}$  to denote the numbers of input slices and output slice predictions ( $C_{out} = C_{in} - 2$  in this work), respectively. No segmentation output is generated for the top or bottom slice since there is not enough context for these boundary slices.

### A. Encoder

Inspired by the success of the design of DeepLabv3+ [13], the encoder consists of a modified Xception [12] structure as the backbone and an ASPP [26] module. By adopting depthwise separable convolution, the Xception model [12] achieves improvement in terms of both speed and accuracy in semantic segmentation. We employ it as our network backbone for its strong feature representation power and small model

size. Since the original Xception model processes a three-channel color image, we modify the input channel number to  $C_{in}$  to jointly process adjacent slices. The ASPP module can potentially improve the segmentation performance by involving different sampling rates and enlarging effective field-of-views, thus capturing target objects as well as context information at different scales. We adopt it here to cover the tumors of various sizes. We initialize the encoder with the weights pre-trained on PASCAL VOC 2012 [27] provided by the official implementation of DeepLabv3+ [13].

### B. Multi-branch Decoder

In the decoder design, a multi-branch structure is proposed to distill the slice-specific information from the encoded volumetric features. Formally speaking, the decoder structure used in 2D or 2.5D FCNs from previous studies can be seen as a single-branch decoder (as illustrated in Fig. 3(a)), where only one central slice is predicted. This single-branch decoder balances the inherent tension between semantics and location, enables precise localization, and produces semantically meaningful predictions from the rich context. However, there are three problems that deserve attention. Firstly, anisotropic volumes have inconsistent information densities along different dimensions. When the slice thickness is much larger than pixel spacing, the single-branch decoder would learn mismatched information along different axes. Secondly, the encoder extracts features by simply fusing intra- and inter-slice information together using isotropic operators, and it fails to extract slice-specific features and loses the inter-slice structural information. Thirdly, the segmentation maps predicted by standard 2D or 2.5D approaches suffer from semantic inconsistency in neighboring slices, since each segmentation map of a slice is separately predicted by one forward inference.

To address the above problems, we design the decoder to have the same number of branches as the number of output slice predictions, and each branch shares the same structure with the single-branch decoder introduced in [13]. The design of our multi-branch decoder is motivated by the fact that the slice-specific information is scrambled through the encoder and we should re-establish them in the decoder with the rich volumetric information provided by the encoder. In this sense, we can explicitly associate one particular branch with one slice, thus bringing more room for improvement by exploiting structure prior between slices.

As shown in Fig. 3(b), we take the low-level features (from the first residual block of Xception) and high-level features (from the outputs of the ASPP module) as the input of the multi-branch decoder (a design similar to DeepLabv3+ [13]). For low-level features, we first employ a  $1 \times 1$  convolution on them to reduce the number of channels, since too many channels in low-level features would outweigh the importance of the rich encoder features and make the training harder [13]. Then,  $C_{out}$  number of  $1 \times 1$  convolutions are conducted on the outputs to explicitly associate one particular branch with one slice. For high-level features, we similarly adopt  $C_{out}$  number of  $1 \times 1$  convolutions on them and upsample their outputs by a bilinear upsampling layer (with a factor of four) to make

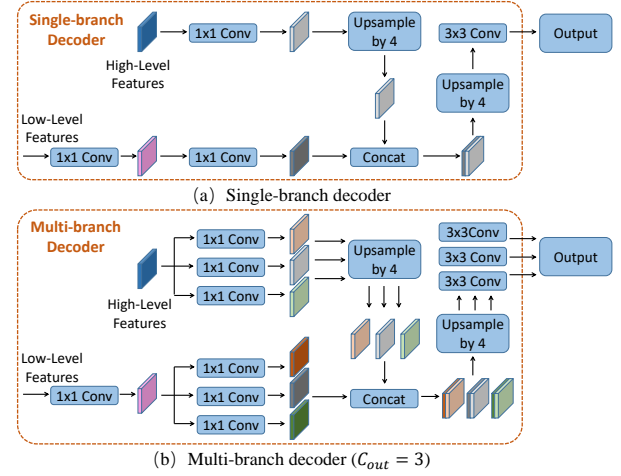


Fig. 3. Single-branch decoder vs. our proposed multi-branch decoder. The single-branch decoder only generates one-slice output while the multi-branch decoder simultaneously generates  $C_{out}$  (equals three in this figure) outputs. Note that the low-level features are from the first residual block of Xception [12], while the high-level features are the output of the ASPP module [13].

the scale consistent with the low-level branch. After that, we get the refined slice-specific features that cover information from both low- and high-level semantics of the encoder. A concatenation operation is then slice-wisely conducted on them to merge the multi-scale semantics. Finally, we adopt another  $C_{out}$  number of bilinear upsampling operations with a rate of four and  $3 \times 3$  convolutions to form the final segmentation outputs.

### C. Slice-Centric Attention Block (SAB)

Instead of directly splitting the encoded feature maps into slice-specific parts as shown in Fig. 3 (b), we propose a novel SAB to strengthen the discrimination of slice-specific features by considering the inter-slice correlations with an attention mechanism. The motivation of this block is based on the observation that the  $1 \times 1$  convolutions employed in the multi-branch decoder to extract slice-specific information are overly simplistic, which is hard to effectively extract discriminative information for individual slices. In contrast, the attention mechanism steers the allocation of slice-specific semantic features towards the most informative components for each output slice and explores the inter-slice correlations, hence improving the performance in learning slice-specific features. For implementation, we embed the proposed SAB into the multi-branch decoder in both the low- and high-level decoding paths as shown in Fig. 2. Note that the two SABs do not share weights since they face different contexts in two different scales.

The attention mechanism is widely adopted in medical image applications, such as pancreas segmentation [14] and universal lesion detection [15]. Our work innovatively explores its usage on the problem of extracting discriminative features for each slice. Fig. 4 shows the technical implementation of the proposed SAB. The volumetric features first pass through a  $3 \times 3$  convolution layer, with one-eighth of channel numbers



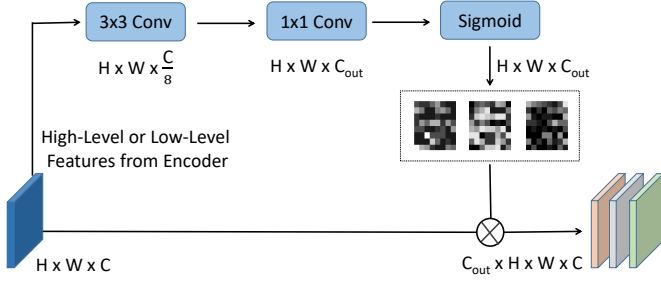


Fig. 4. Illustration of the proposed slice-centric attention block ( $C_{out} = 3$ ). It strengthens the discriminative power of the multi-branch decoder in learning slice-specific features with an attention mechanism.  $C$  is the channel number of the high-level or low-level features, which is the same as the original DeepLabv3+.

of the input feature map. Then,  $C_{out}$   $1 \times 1$  convolution layers and sigmoid functions are applied to generate  $C_{out}$  weight maps of size  $H \times W \times C_{out}$ . These weight maps can be seen as an attention mechanism for each branch attending to the key slice-specific information from the abundant features. The learned weight maps are then separately multiplied by the input features to extract slice-specific features. As we will demonstrate in the experiments, although conceptually simple, the proposed SAB is effective in strengthening the discriminative power for each slice in feature learning.

#### D. Loss Function

The Dice loss introduced in [28] is commonly used to address the class imbalance problem between foreground and background classes. As defined below in Eq. (1), the Dice loss with three classes (*i.e.*, background, liver, and liver tumor) is formulated as:

$$L_{Dice} = - \sum_{c=1}^3 \sum_{m=1}^{C_{out}} \frac{2 \sum_{i=1}^V p_{m,i}^c g_{m,i}^c}{\sum_{i=1}^V (p_{m,i}^c)^2 + \sum_{i=1}^V (g_{m,i}^c)^2}, \quad (1)$$

where  $p_{m,i}^c$  denotes the predicted probability of voxel  $i$  in the  $m^{th}$  slice belonging to class  $c$ ;  $g_{m,i}^c$  denotes the corresponding ground truth; and  $V$  is the number of voxels in each slice.

As aforementioned, slice-aware design brings room for improvement by exploiting structure prior between slices. We thus propose a regularization term as an additional loss to improve the coherence between neighboring slices in the label space. Specifically, we use the union of the two slices in the prediction results and the union of the corresponding two slices in the ground truth to calculate the pairwise Dice loss, which is denoted as

$$P_{m,n} = - \sum_{c=1}^3 \frac{2 \sum_{i=1}^V (p_{m,i}^c + p_{n,i}^c)(g_{m,i}^c + g_{n,i}^c)}{\sum_{i=1}^V (p_{m,i}^c + p_{n,i}^c)^2 + \sum_{i=1}^V (g_{m,i}^c + g_{n,i}^c)^2}, \quad (2)$$

where  $P_{m,n}$  denotes the pairwise Dice loss of the  $m^{th}$  slice and  $n^{th}$  slice along out-of-plane direction;  $p_{m,i}^c$  and  $p_{n,i}^c$  denote the predicted probability of voxel  $i$  belonging to class  $c$  in the  $m^{th}$  slice and  $n^{th}$  slice, respectively;  $g_{m,i}^c$  and  $g_{n,i}^c$  denote the corresponding ground truth. To further supplement inter-slice information flow and improve inter-slice coherence,

TABLE I  
COMPARISON OF OUR PROPOSED SAMBD ( $C_{in} = 7$ ,  $C_{out} = 5$ ) WITH DEEPLABV3+ [13] AND H-DENSEUNET [6] IN TERMS OF PARAMETERS AND FLOPS.

	DeepLabv3+	SAMBD	H-DenseUNet
Parameters (M)	41.06	41.26	61.44
FLOPs (G)	0.83	0.84	2841.6

we calculate the pairwise Dice loss in a dense way, where each slice is coupled with multiple nearby slices to calculate multiple Dice losses. We name the new loss as the Densely Connected Dice (DCD) loss. Since the interaction of two slices decreases with increasing distance, for each paired slices  $m$  and  $n$  ( $n > m$ ), we add a weight  $w_{m,n} = 1/(n - m)$ . The DCD loss is defined as:

$$L_{DCD} = \sum_{m=1}^{C_{out}-1} \sum_{n=m+1}^{C_{out}} w_{m,n} P_{m,n}. \quad (3)$$

The final loss function is composed of a weighted combination of the Dice loss and the proposed DCD loss:

$$L = L_{Dice} + \lambda \times L_{DCD}, \quad (4)$$

where we define  $\lambda = C_{out} / (\sum_{m=1}^{C_{out}-1} \sum_{n=m+1}^{C_{out}} w_{m,n})$  to balance the importance of the intra-slice semantic constraint and inter-slice smoothness.  $\lambda$  is such designed that its denominator normalizes the sum of the weights  $w_{m,n}$  in Eq. (3) to one, whereas its numerator strengthens the supervision with more output slices. Empirical parameter tuning is likely to yield better results; in this paper, however, we would like to present a generally useful regularizer that can be safely applied to other segmentation tasks without any parameter tuning. We find that the presented design of  $\lambda$  performs well as validated by the superior performance on two publicly available datasets in the experiments.

#### E. Model Complexity

Due to the effective design of the multi-branch decoder and slice-centric attention block, the parameters and FLOPs (undefined) of our proposed SAMBD with  $C_{in} = 7$ ,  $C_{out} = 5$  are very competitive with DeepLabv3+ [13], and markedly superior to H-DenseUNet [6]. As shown in Table I, our method only brings 0.49% extra parameters and 1.2% extra FLOPs compared to DeepLabv3+, and only incurs 67.2% and 0.03% of the parameters and FLOPs of H-DenseUNet, respectively.

### IV. EXPERIMENTS

In this section, we evaluate our approach on the LiTS [11] dataset to demonstrate the robustness and generalization capability, compared to several state-of-the-art segmentation methods. Extended experiments have also been performed on the ISBI 2019 SegTHOR dataset [29] to validate the generalization capability to other human organs. We implement all the experiments with Keras [30] using three NVIDIA GeForce GTX 1080 GPUs. Stochastic gradient descent with momentum (0.9) is used to update the weights of the network. The initial learning rate is set to 0.001 and multiplied by 0.9 after each epoch. We train the network for 80 epochs.

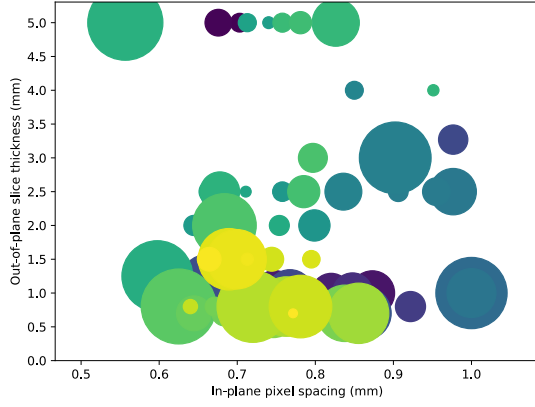


Fig. 5. The high variations of the LiTS training set. We can observe the large inconsistency between the in-plane pixel spacing (horizontal axis) and out-of-plane slice thickness (vertical axis). Besides, the variations in tumor size (shown in circles with different sizes) are also high.

### A. Experiments on LiTS

1) *LiTS*: The LiTS dataset<sup>1</sup> [11] is a publicly available liver tumor dataset consisting of 201 contrast-enhanced abdominal CT scans collected from various clinical sites over the world. The dataset was originally split into a training set (131 scans) and a test set (70 scans), where only the training set was publicly released with accurate liver and tumor masks. As aforementioned, the LiTS dataset suffers from apparent inconsistency between the in-plane pixel spacing and out-of-plane slice thickness (see Fig. 5). Furthermore, high varieties and complexities exist for livers and liver tumors, including the location, size, and shape. Besides, the heterogeneity in liver and liver tumor contrast is very large between subjects, as shown in Fig. 6.

2) *Implementation Details and Evaluation Metrics*: For image pre-processing, we unify the volume orientations and truncate the image intensity values of all scans outside the range of  $[-200, 250]$  Hounsfield Unit (HU) to ignore irrelevant image details. Since the slice thickness varies greatly between subjects, we resample scans with slice thickness greater than  $1\text{ mm}$  to  $1\text{ mm}$  in both training and inference phases. We preserve the original slice thickness for patients whose slice thickness is less than  $1\text{ mm}$  to leverage the original high-resolution spatial information. We do not unify the in-plane pixel spacing by the resample operation since the variation is relatively small and interpolation often introduces artifacts, which may offset the performance gain from resolution normalization. To alleviate the overfitting problem, we conduct data augmentation in the training phase; concretely, we first apply random scaling (from 0.8 to 1.2) to all training data, and then randomly crop a  $256 \times 256 \times C_{in}$  subregion as the input to the network. For post-processing during the inference phase, we take the largest connected component as the liver segmentation and remove liver tumor predictions outside the liver region.

In the test phase, a sliding-window approach is employed to predict the segmentation mask for an input volume. Con-

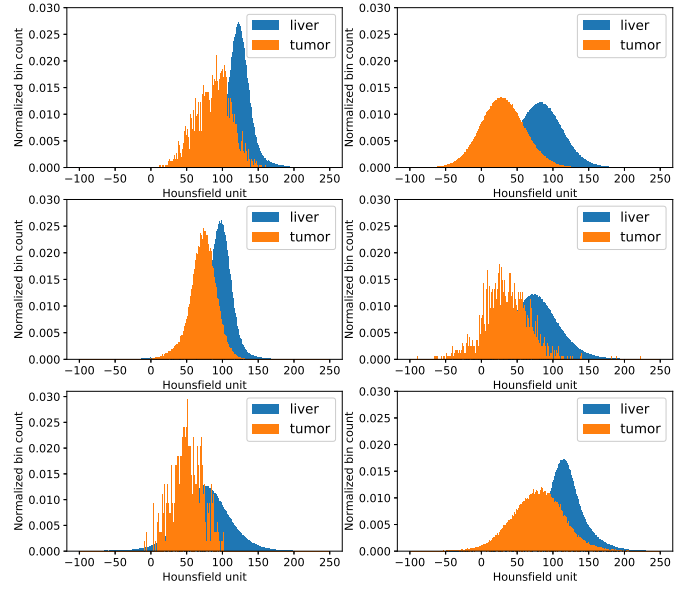


Fig. 6. Examples depicting the heterogeneity of CT scan contrast in liver and liver tumor areas. The horizontal axis represents the Hounsfield unit values of CT scans, and the vertical axis represents the proportion of voxels falling into intervals with different Hounsfield unit values.

cretely, we extract consecutive multi-slice inputs from the volume by moving along the out-of-plane direction with a stride of one, and predict the segmentation mask for each of these multi-slice inputs. Therefore, each slice may appear in multiple multi-slice inputs due to the overlap and be predicted multiple times. The final segmentation mask of a slice is then obtained by averaging its multiple predictions. Finally, the segmentation masks of all the slices are stacked in sequence to form the segmentation result of the entire volume, which is then resampled to the original resolution, if necessary.

The Dice-per-case score and Dice-global score are adopted in the LiTS challenge as the evaluation metrics to measure the liver and liver tumor segmentation performance. The Dice-per-case score reflects the averaged Dice score across all patients, whereas the Dice-global score is the Dice score evaluated by stacking all volumes into one long volume. For comparison with other methods, we also use Volume Overlap Error (VOE), Relative Volume Difference (RVD), Average Symmetric Surface Distance (ASSD), Maximum Surface Distance (MSD), and Root Means Square Symmetric Surface Distance (RMSD) as metrics for complementary evaluation. For interpretation of these evaluation metrics, readers can refer to [11].

3) *Ablation Study on LiTS*: To analyze the effectiveness of our approach, we conduct ablation studies on a validation dataset consisting of 25 volumes, which are randomly selected from the training set. Since our network backbone is derived from DeepLabv3+ [13], we take it as the baseline benchmarking method. We notice that the meta-information about slice thickness and pixel spacing of some cases in the training and validation sets provided by the LiTS organizers is wrong, which makes the values of surface-based metrics in the ablation study weird compared to those presented in Table VI. Therefore, we do not present surface-based metrics in Table

<sup>1</sup><https://competitions.codalab.org/competitions/17094>

TABLE II  
LIVER AND TUMOR SEGMENTATION RESULTS OF AN ABLATION STUDY ON THE LiTS VALIDATION DATASET.

$C_{in}^a$	$C_{out}^a$	Method	MD <sup>b</sup>	SAB <sup>b</sup>	DCD <sup>b</sup>	Liver			Tumor		
						Dice per case [%]	Dice global [%]	VOE	Dice per case [%]	Dice global [%]	VOE
1	1	(a) Baseline				95.27 ± 1.90	95.39	0.097 ± 0.032	57.29 ± 33.03	67.79	0.608 ± 0.311
5	1	(b) Baseline				95.34 ± 1.78	95.56	0.087 ± 0.033	60.02 ± 29.78	69.67	0.564 ± 0.289
7	1	(c) Baseline				95.39 ± 1.76	95.59	0.084 ± 0.032	60.97 ± 28.97	70.15	0.561 ± 0.279
5	3	(d) Baseline (1×)				96.08 ± 1.89	96.56	0.077 ± 0.034	62.54 ± 28.33	72.81	0.530 ± 0.285
		(e) Baseline (3×)				96.15 ± 1.75	96.59	0.076 ± 0.032	62.67 ± 27.80	71.25	0.531 ± 0.276
		(f) MD	✓			<b>96.27 ± 1.78</b>	<b>96.74</b>	0.074 ± 0.032	64.62 ± 27.29	72.48	0.510 ± 0.269
		(g) MD+SAB	✓	✓		96.12 ± 1.79	96.58	0.077 ± 0.032	66.04 ± 23.90	75.06	0.503 ± 0.253
		(h) MD+DCD	✓		✓	96.13 ± 1.92	96.61	0.076 ± 0.035	65.71 ± 23.60	71.98	0.508 ± 0.253
		(k) MD+SAB+DCD	✓	✓	✓	96.09 ± 1.80	96.55	0.077 ± 0.033	67.07 ± 23.79	73.82	0.491 ± 0.247
		(l) Baseline (1×)				95.69 ± 3.08	96.51	0.084 ± 0.034	63.04 ± 25.00	72.25	0.534 ± 0.266
7	5	(m) Baseline (5×)				95.97 ± 1.82	96.44	0.079 ± 0.033	64.60 ± 26.33	73.66	0.513 ± 0.258
		(n) MD	✓			96.15 ± 1.88	96.64	0.076 ± 0.034	65.69 ± 22.65	70.67	0.511 ± 0.245
		(o) MD+SAB	✓	✓		96.20 ± 1.67	96.60	<b>0.074 ± 0.030</b>	66.73 ± 24.34	74.40	0.494 ± 0.241
		(p) MD+DCD	✓		✓	96.15 ± 1.92	96.65	0.076 ± 0.035	67.65 ± 21.00	72.82	0.492 ± 0.232
		(q) MD+SAB+DCD	✓	✓	✓	95.95 ± 1.96	96.50	0.080 ± 0.035	<b>70.17 ± 18.06</b>	<b>75.84</b>	<b>0.467 ± 0.212</b>

<sup>a</sup>  $C_{in}$  and  $C_{out}$  represent the number of input slices and output slice predictions, respectively.

<sup>b</sup> MD, SAB and DCD represent the Multi-branch Decoder, the Slice-centric Attention Block and the Densely Connected Dice loss, respectively.

II for the ablation study.

*a) The number of input slices:* Objects in adjacent slices usually have intrinsic relations in various properties, such as shape and location. In this sense, we employ a 2.5D network, which takes a few adjacent slices as the input to capture the inter-slice information. To verify that the inter-slice information is useful in the segmentation task, we conduct experiments with different numbers of input slices. In Table II, rows (a), (b), (c) show the results of DeepLabv3+, with 1, 5, 7 adjacent slices as input, respectively, and the corresponding central slice as output. We can observe consistent improvement when the number of input slices increases, confirming our assumption that more adjacent slices can provide more inter-slice information for achieving higher segmentation performance.

*b) Effectiveness of the multi-branch decoder:* To verify the effectiveness of the proposed multi-branch decoder, a straightforward baseline is a single-branch decoder that outputs the same number of channels as the multi-branch decoder. The multi-branch decoder consists of  $C_{out}$  parallel branches, each having the same structure with the single-branch decoder in the baseline, as shown in Fig. 3(b). Here, we denote the multi-branch structure as MD. We present two different settings of the numbers of input slices and output predictions ( $C_{in} = 5, C_{out} = 3$  and  $C_{in} = 7, C_{out} = 5$ ) to verify that our multi-branch design can bring consistent improvement under different inter-slice context. The results are shown in Table II. As we can see, compared to the baseline with the same input and output settings (rows (d) and (l)), our proposed multi-input multi-output architecture prominently enhances the segmentation of liver and liver tumor, with improvements in both Dice-per-case and Dice-global (rows (f) and (n)). Besides, to further demonstrate that the improvements upon the single-branch decoder are not due to the increased parameters, we also present results of the single-branch decoder with 3× and 5× channels in rows (e) and (m). We find that even with more parameters, the single-branch decoder still performs worse than our multi-branch decoder in both liver and tumor segmentation. The multi-branch decoder focuses on modeling the slice-specific information from the mixed volumetric semantic information, which is of great significance for network training with anisotropic data.

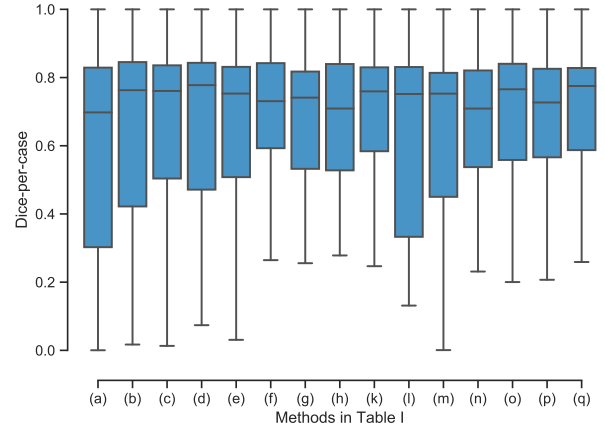


Fig. 7. Box-plots of each method in Table II on tumor segmentation. Horizontal axis represents the methods listed in Table II, and vertical axis represents the Dice-per-case score.

*c) Effectiveness of the slice-centric attention block:* Next, we verify the effectiveness of the slice-centric attention block, which is abbreviated as SAB in Table II. Rows (g) and (o) show the detailed results with the multi-branch decoder and SAB for liver and liver tumor segmentation. Compared to (f) and (n), which are equipped only with the multi-branch decoder, the SAB achieves 1.42% and 1.04% improvements in Dice-per-case for tumor segmentation. The improvements demonstrate that the deep semantics learned by the encoder contains vast amounts of redundant information, and directly decoding with convolution operators cannot effectively extract slice-specific features. With the proposed attention mechanism, the network can steer the allocation of available semantic features towards the most informative components for different slices, thus bringing improvements in terms of accuracy.

*d) Effectiveness of the DCD loss:* Besides the regular Dice loss which measures the difference between the segmentation prediction and given ground truth mask, the method is further equipped with a DCD loss as shown in Eq. (4), which is proposed to regularize the inter-slice predictions to be more coherent in the label space. To investigate the effectiveness of

TABLE III  
THE SEGMENTATION RESULTS OF BASELINE (d) AND BASELINE (l) IN TABLE II WITH UNIFIED RESOLUTION ON THE LiTS VALIDATION DATASET.

$C_{in}^a$	$C_{out}^a$	Unified Resolution [mm]	Liver			Tumor		
			Dice per case [%]	Dice global [%]	VOE	Dice per case [%]	Dice global [%]	VOE
5	3	2	91.02 $\pm$ 7.21	92.43	0.158 $\pm$ 0.103	46.42 $\pm$ 32.67	64.18	0.636 $\pm$ 0.297
		1	95.41 $\pm$ 1.80	95.74	0.088 $\pm$ 0.032	63.80 $\pm$ 24.27	72.14	0.440 $\pm$ 0.260
		0.75	95.25 $\pm$ 2.07	95.79	0.090 $\pm$ 0.037	63.42 $\pm$ 24.89	68.02	0.489 $\pm$ 0.263
7	5	2	95.11 $\pm$ 1.85	95.45	0.093 $\pm$ 0.033	55.89 $\pm$ 32.17	72.05	0.503 $\pm$ 0.310
		1	95.66 $\pm$ 1.95	96.17	0.083 $\pm$ 0.035	63.75 $\pm$ 25.36	72.53	0.486 $\pm$ 0.255
		0.75	95.27 $\pm$ 1.95	95.78	0.090 $\pm$ 0.035	65.52 $\pm$ 24.06	69.87	0.468 $\pm$ 0.254

<sup>a</sup>  $C_{in}$  and  $C_{out}$  represent the number of input slices and output slice predictions, respectively.

TABLE IV  
THE SEGMENTATION RESULTS OF 3D U-NET [31] WITH UNIFIED RESOLUTION ON THE LiTS VALIDATION DATASET.

Unified Resolution [mm]	Liver			Tumor		
	Dice per case [%]	Dice global [%]	VOE	Dice per case [%]	Dice global [%]	VOE
2	90.95 $\pm$ 7.05	92.34	0.160 $\pm$ 0.102	40.43 $\pm$ 35.03	63.34	0.433 $\pm$ 0.364
1	93.55 $\pm$ 3.19	94.27	0.120 $\pm$ 0.054	59.06 $\pm$ 27.79	65.29	0.529 $\pm$ 0.272
0.75	95.41 $\pm$ 1.79	95.77	0.087 $\pm$ 0.0319	62.49 $\pm$ 27.27	71.80	0.492 $\pm$ 0.275

TABLE V  
P-VALUES FOR THE PAIRED T-TESTS BETWEEN OUR METHOD AND THE SINGLE-BRANCH DECODER THAT OUTPUTS THE SAME NUMBER OF CHANNELS AS THE MULTI-BRANCH DECODER.

$C_{in} = 5, C_{out} = 3$	Dice-per-case	$C_{in} = 7, C_{out} = 5$	Dice-per-case
p-value	0.034	p-value	0.010

the DCD loss, we conduct several ablation studies. First, we train the network with the multi-branch decoder using only the DCD loss, as shown in Table II, rows (h) and (p). Compared to training with the Dice loss (rows (f) and (n) in Table II), the DCD loss improves the overall Dice-per-case score for tumor by 1.09% and 1.96%, respectively. Then, we add the slice-centric attention block to the network (rows (k) and (q) in Table II) and train it with the Dice loss and the DCD loss, respectively. We get a 1.03% improvement in Dice-per-case score for tumor with 5-slice input and 3-slice output, while 3.44% improvement with 7-slice input and 5-slice output. It is worth mentioning that compared to the baseline, the proposed method performs preferably in segmenting tumor with an improvement of about 7.13% in Dice-per-case score by comparing rows (l) and (q), which proves the effectiveness of the proposed network with all modules enabled.

*e) Statistical analysis:* To analyze whether the performance improvement of the proposed SAMBD is statistically significant, we conduct the paired t-test as [32] on tumor segmentation between our method and the single-branch decoder that outputs the same number of channels as the multi-branch decoder with two configurations of  $C_{out}$  and  $C_{in}$ . We evaluate the significance for Dice-per-case with a significance level of 0.05. The p-values are shown in Table V. As we can see, all the p-values are below 0.05, demonstrating that our improvements upon the single-branch decoder are statistically significant.

*f) Box-plots:* In Fig. 7, we show the box-plots of each method listed in Table II in terms of Dice-per-case on tumor segmentation. Compared to other methods (rows (a) – (h) and

(l) – (p)), the proposed SAMBD presents results that not only have higher median accuracy but also show less dispersion, indicating consistently better performance in general.

*4) Does a unified resolution help address the data variability problem?:* In the medical image segmentation, a widely adopted approach to the data variability problem is resampling the data into a unified resolution. We present the results of the single-branch decoder that outputs the same number of channels as the multi-branch decoder with different unified resolutions in Table III. Compared to the results of SAMBD in rows (k) and (q) of Table II, the results with the single-branch decoder still present a large performance gap in Dice-per-case, Dice-global, and VOE.

Since 3D networks are good at extracting contextual information and the resampling operation is widely adopted to process data suffering large variations, we present extra experimental results with 3D U-Net [31] to see if it can produce sound results with a unified resolution. The results are shown in Table IV. From the table, we can observe that 3D U-Net produces significantly different results with different input resolutions. The best performance is achieved at the highest resolution ( $0.75 \times 0.75 \times 0.75$  in mm), yet substantially lower than the results of SAMBD in row (q) of Table II (e.g., 62.49% vs. 70.17% in Dice-per-case for tumor segmentation).

The superior performance of the proposed SAMBD towards 2D/3D networks with a unified resolution verifies the effectiveness of our slice-aware design in processing anisotropic data.

*5) Comparison with Other Methods:* Based on the results from the ablation study, we pick the best network architecture and hyper-parameters, and train the model on the whole LiTS training set with 131 CT scans. We evaluate the model on 70 test cases and submit the results to the challenge website. Table VI tabulates the quantitative comparison results of our proposed SAMBD and several state-of-the-art methods already published. All of these top-ranking methods employ deep learning based approaches, demonstrating the effectiveness of



TABLE VI  
COMPARISON WITH STATE-OF-THE-ART METHODS FOR LIVER AND TUMOR SEGMENTATION ON THE LiTS TEST DATASET.

Method	Liver								Tumor							
	Dice per case [%]	Dice global [%]	VOE	RVD	ASSD	MSD	RMSD		Dice per case [%]	Dice global [%]	VOE	RVD	ASSD	MSD	RMSD	
Han [10]	96.0	96.5	0.077	-0.004	1.15	24.499	2.421		67.6	79.6	0.383	0.464	1.143	7.322	1.728	
3D DenseUNet [6]	93.6	92.9	-	-	-	-	-		59.4	78.8	-	-	-	-	-	
H-DenseUNet [6]	96.1	96.5	0.074	-0.018	1.45	27.118	3.15		72.2	82.4	0.366	4.272	<b>1.102</b>	6.228	<b>1.595</b>	
AH-Net [8]	96.3	97.0	0.07	-0.004	1.099	23.992	2.398		63.4	<b>83.4</b>	<b>0.353</b>	0.365	1.185	6.482	1.667	
DeepX [33]	96.3	96.7	0.071	-0.01	1.104	23.847	2.303		65.7	82.0	0.378	0.288	1.151	6.269	1.678	
SAMBD	96.5	97.0	<b>0.065</b>	0.004	0.971	21.997	2.034		72.8	81.0	0.405	-0.208	1.258	6.582	1.796	
SAMBD (ensemble)	<b>96.6</b>	<b>97.1</b>	<b>0.065</b>	<b>0.002</b>	<b>0.953</b>	<b>21.933</b>	<b>1.998</b>		<b>73.6</b>	81.2	0.401	<b>-0.196</b>	1.174	<b>6.18</b>	1.675	

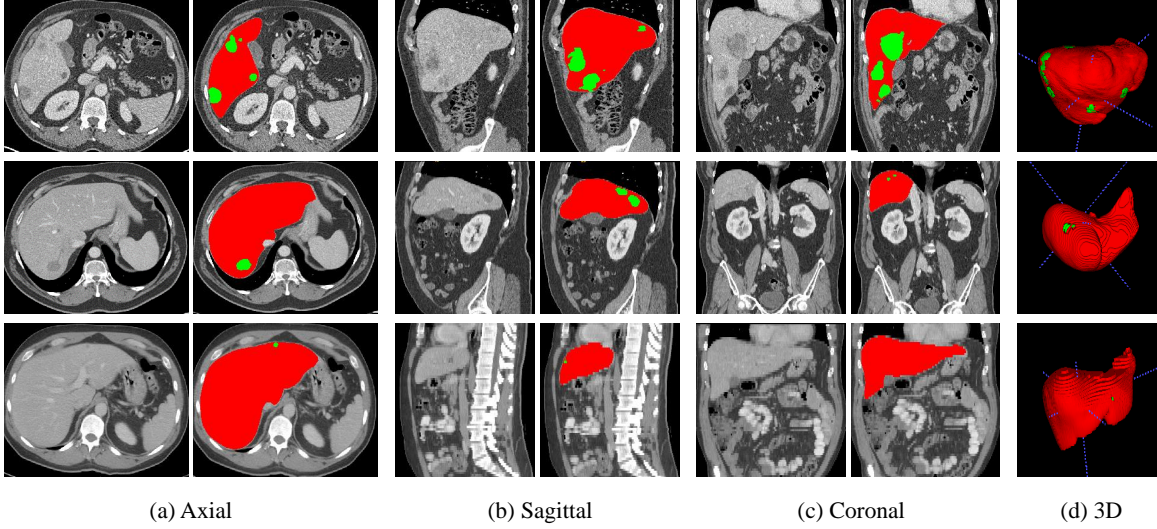


Fig. 8. Example segmentation results on the LiTS test set. Each row shows a CT scan acquired from an individual subject with different slice thickness (0.7 mm, 2.5 mm, and 5 mm from top to bottom, respectively). The first six columns show the original CT scans and corresponding segmentation results in the axial, sagittal, and coronal planes, respectively. The last column shows segmentation results in a 3D view. Red represents the liver and green the liver tumor.

CNNs in the field of medical image analysis. Han [10] adopted a 32-layer U-Net-alike architecture, where adjacent slices were employed as input and produced segmentation maps through a single-branch decoder. Li *et al.* [6] and Liu *et al.* [8] transferred convolutional features learned from 2D images to 3D volumes and then applied 3D convolutional kernels to extract 3D context. Yuan [33] developed a hierarchical framework to segment the liver and tumor in three steps. Compared to these methods, we are the first who pay attention to the large inconsistency between the in-plane pixel spacing and out-of-plane slice thickness.

As shown in Table VI, our method achieves the best segmentation accuracy for both liver and tumor in the Dice-per-case score even with a single model. Using an ensemble version, we achieve state-of-the-art performances in three of the four main evaluation metrics, including the Dice-per-case score of the tumor, and Dice-global and Dice-per-case scores of the liver, demonstrating the effectiveness of our method. Besides, our methods show very competitive results on the complementary evaluation metrics, achieving the best results in VOE, RVD, ASSD, MSD, and RMSD for liver, as well as best results in RVD and MSD for tumor. Fig. 8 shows several examples of the segmentation results with different slice thickness (0.7 mm, 2.5 mm, and 5 mm, respectively). By taking into consideration the information asymmetry along with the in- and out-of-plane directions into our network

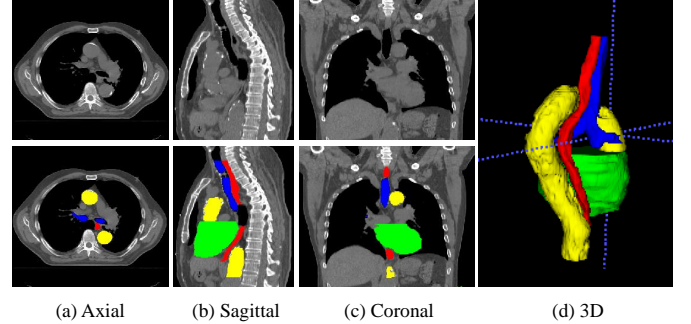


Fig. 9. A typical CT scan in the SegTHOR dataset and the corresponding segmentation ground truth with red for the esophagus, green for the heart, blue for the trachea, and yellow for the aorta.

design, our method presents decent ability in segmenting liver and liver tumors across a wide range of resolution settings.

## B. Experiments on SegTHOR

1) *SegTHOR*: The dataset<sup>2</sup> [29] is provided by ISBI 2019, with the purpose of addressing the problem of Organs at Risk (OAR) segmentation in CT images. SegTHOR focuses on four OARs: heart, aorta, trachea, and esophagus (refer to Fig. 9 for

<sup>2</sup><https://competitions.codalab.org/competitions/21145>

TABLE VII  
SEGMENTATION RESULTS ON THE TEST DATASET IN LEADERBOARD OF 2019 SEGTHOR CHALLENGE (UNTIL MAY 10, 2019 WHEN THE LIVE CHALLENGE OF ISBI ENDED.)

Rank	User/Method	Esophagus		Heart		Trachea		Aorta	
		Dice [%]	Hausdorff	Dice [%]	Hausdorff	Dice [%]	Hausdorff	Dice [%]	Hausdorff
1	gaoking132	86.51	0.2590	95.36	0.1272	<b>92.76</b>	<b>0.1453</b>	94.64	0.1209
2	MILab [34]	85.94	0.2743	95.00	0.1383	92.01	0.1824	<b>94.84</b>	<b>0.1129</b>
3	Jone	85.91	0.3185	94.89	0.1435	92.19	0.1590	94.73	0.1251
4	hyang	83.81	0.3534	<b>95.42</b>	<b>0.1208</b>	92.33	0.1973	94.43	0.1290
5	dp	83.39	0.3351	95.19	0.1325	91.57	0.2041	93.51	0.1980
6	grr	85.82	0.2928	94.56	0.1867	91.53	0.2090	93.91	0.2010
7	dlachinov	85.72	0.2686	94.36	0.1761	91.55	0.1736	91.82	0.2214
8	ZWB	82.03	0.3838	94.58	0.1594	92.17	0.2045	94.33	0.1551
9	svesal [35]	85.79	0.3303	94.15	0.2263	92.57	0.1929	93.75	0.2971
10	Louisvh	83.61	0.3399	94.02	0.1973	90.68	0.2091	93.32	0.2732
-	Chen <i>et al.</i> [36]	81.66	0.4914	93.29	0.2417	89.10	0.2746	92.32	0.3081
-	Zhang <i>et al.</i> [37]	77.32	1.6774	93.84	0.2089	89.39	0.2741	92.32	0.3081
2	SAMBD	<b>86.57</b>	<b>0.2478</b>	95.28	0.1296	92.56	0.1817	94.43	0.1498

data visualization), and pays particular attention to esophagus segmentation (the esophagus is the most challenging to segment due to its variable location relative to neighboring organs and low-intensity contrast to background). All CT images are obtained from 60 patients (11,084 slices) with non-small cell lung cancer and have already been randomly split into a training set with 40 volumes (7,390 slices) and a test set with 20 volumes (3,694 slices). The planar size is  $512 \times 512$  in pixels and the in-plane pixel spacing varies from 0.9 *mm* to 1.37 *mm*, while the out-of-plane slice thickness varies from 2 *mm* to 3.7 *mm*. All the organs are delineated by an experienced radiation oncologist.

2) *Implementation Details and Evaluation Metrics*: We set the HU value range to  $[-1000, 500]$  to exclude irrelevant organs and objects. The slice thicknesses of all subjects are resampled to 2.5 *mm*. Hyperparameters and data augmentation are the same as those used in the experiments on the LiTS dataset. In addition, there are great variations of shape and position among four organs: some extend along the out-of-the plane direction (*e.g.*, esophagus) and some have a large volume (*e.g.*, heart). To address these issues, we cut the 3D CT scans into slices along the axial, sagittal, and coronal plane, respectively, and then feed them into the network together. The size of the axial slices is  $512 \times 512$  pixels, whereas those of the sagittal and coronal slices are  $512 \times z$  and  $z \times 512$  pixels ( $z$  is the number of axial slices), respectively. To facilitate the segmentation network, the sagittal and coronal slices are then resized to  $512 \times 512$  pixels. In the test phase, we separately predict the segmentation masks in three orientations (*i.e.*, the axial, sagittal, and coronal) for each volume in a way similar to the experiments on the LiTS dataset, and ensemble them with the majority voting to produce the final segmentation mask.

The overlap Dice metric and the Hausdorff distance [35] are used by the SegTHOR challenge as the official evaluation metrics. The Dice score and Hausdorff distance are complementary metrics and our aim is to make the Hausdorff distance close to 0 while the Dice score close to 1.

3) *Comparison with Other Methods*: In this experiment, we pay special attention to the segmentation of the esophagus, because of its hard-to-distinguish boundary and low contrast.

Table VII lists the performance of the top 10 teams on the leaderboard. Our method outperforms other methods on the segmentation of the esophagus and achieves very competitive performance for heart and trachea segmentation. We finally achieve the tied second place by the overall rank and the first place on esophagus segmentation. It is worth mentioning that our method is not specially tailored for the tasks of SegTHOR—unlike the contrasting methods in Table VII which were intended for the specific challenge—yet produces such competitive results, demonstrating its generalizability.

Among all the methods listed in Table VII, He *et al.* [34] developed a uniform U-like encoder-decoder architecture for the segmentation of thoracic organs, which combined the major task of local pixel-wise segmentation and an auxiliary task of global slice classification. Vesal *et al.* [35] employed a 2D U-Net combined with dilated convolutions using only one slice as input. Chen *et al.* [36] and Zhang *et al.* [37] segmented thoracic organs through a two-stage strategy, where four organs were first localized and then the precise segmentation stage was applied based on the location. Our method concentrates on making use of the relationships between adjacent slices and learning slice-specific information, which helps identify inconspicuous objects. The segmentation result of the esophagus in Table VII confirms this. We conclude that our method can conquer the challenge and well segment the OARs from CT scans.

## V. DISCUSSION

Accurate segmentation of liver and liver tumors in CT images facilitates the quantitative assessment of the tumor burden, treatment planning, and prognosis. There have been considerable debates over 2D versus 3D networks on 3D medical images—choosing 2D networks for the benefit of 2D pretraining and large-scale training sets or alternatively 3D networks for native 3D representation learning [16], [17]. This paper innovatively identifies the wide variation in the ratio between intra- and inter-slice resolutions as a crucial obstacle to the performance, which may in turn affect decision-making in choosing 2D or 3D networks. This argument is also supported by the five-fold cross-validation results of nnU-Net

[38], in which 2D U-Net produces inferior results on the brain tumour, heart, hippocampus, lung, and pancreas tasks than 3D U-Net, but better results on prostate and liver tasks. In this sense, we hope this work would provide a different perspective of deciding 2D or 3D networks.

It is common to take multiple slices as input and output a single slice, either for making up a three-channel image for pre-trained weights or pursuing a 3D context [10]. Our work takes a step further to output multiple slices, which enables to explore extra design space for the decoder and loss function for supervision to maintain 3D coherence between slices. Our focus is then on how to learn the most discriminative features for each individual slice, which avoids directly processing anisotropic information. To this end, we propose a multi-branch decoder to explicitly re-establish discriminative features for each separate slice by associating each slice with a separate branch. As far as we know, only few existing works emphasize learning slice-specific features in a 2.5D network. As verified in the experiment, such a slice-aware design greatly boosts the performance compared to the multi-output single-branch decoder that does not distinguish different slices in the feature learning. Although in clinical practice, some data may not show large variations, we believe that the slice-aware design would still bring extra help to the segmentation performance.

The limitation of such an explicit design of the multi-branch decoder is that it is computationally prohibitive when the number of output slices becomes large. In this work, we leverage and embed an attention mechanism into the multi-branch decoder, which is expected to steer the allocation of slice-specific semantic features towards the most informative components for each output slice by fully-exploiting the inter-slice correlations. We believe that it is novel to adopt such an attention mechanism to strengthen the discriminative power of each slice, yet we acknowledge that there should also be other design choices of the attention block. We plan to integrate our multi-branch decoder with other well-designed attention blocks (e.g., considering multi-scale features) in future work, seeking more accurate segmentation of the images with slice-aware modeling. Besides, in this work we only study the proposed SAMBD with  $C_{in} = 5$ ,  $C_{out} = 3$  and  $C_{in} = 7$ ,  $C_{out} = 5$ . Future work should investigate an optimal way to determine the configuration that is general enough for satisfactory results in most cases. Moreover, although our framework demonstrates competitive computational complexity with the DeepLabV3+ [13] and H-DenseUNet [6] (Table I), methodologies for more compact deep neural network designs (e.g., [39]) can be considered in future work.

A potential concern of the DCD loss is that it may potentially affect the proper training when the slice thickness is large. Actually, the motivation of this regularizer is to supplement inter-slice information and thus improve inter-slice coherence. There should be a balance between the main loss function and the regularizer just as traditional machine learning algorithms (such as Lasso). In Fig. 10, we group the segmentation accuracy of different methods (rows (n), (p), and (q) in Table II) by slice thickness into three groups: less than, equal to, and greater than 1 mm. From the figure,

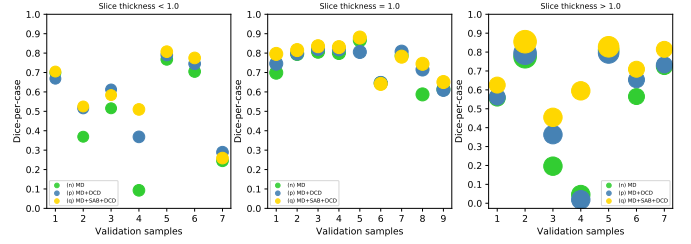


Fig. 10. The segmentation accuracy of different methods in rows (n), (p), and (q) of Table II grouped by slice thickness less than, equal to, greater than 1 mm. We find consistent performance improvement brought by the DCD loss irrespective of different slice thicknesses.

we find consistent performance improvement brought by that DCD loss with different slice thicknesses. Besides, the range (from 0.69 to 5 in mm) of slice thickness in LiTS is large enough to cover most cases in clinical practice. Future works should consider the effectiveness of such a regularizer in other segmentation tasks.

Considering that the problem of data variations in resolution is prevalent in medical imaging and we believe that the proposed SAMBD can generalize well to other segmentation tasks, we also evaluated SAMBD on SegTHOR to verify this in Section IV-B. It is noteworthy that, here, we re-trained all over again on the SegTHOR dataset; it should be a potential direction to explore the transfer learning techniques [40] (for example, fine-tuning with pre-trained weights) to speed up the training process. Besides, to further verify the effectiveness of our proposed method in conquering data variations in resolution, we plan to test our method on more organs/tumors in other body parts and imaging modalities.

## VI. CONCLUSION

In this paper, we rethought the debates over 2D versus 3D networks from a data viewpoint, where we identified the wide variation in the ratio between intra- and inter-slice resolutions as an important obstacle to the performance. To circumvent this, we proposed a slice-aware multi-input multi-output structure to emphasize the importance of feature learning for each slice. A multi-branch decoder with a slice-centric attention block was proposed to gradually and explicitly re-establish discriminative features for each slice by fully-exploiting intra- and inter-slice information learned by the encoder with the widely adopted attention mechanism. To further enhance the correlation between slices and enable coherent segmentation, we proposed a densely connected Dice loss as a regularization term. Quantitative evaluations on the LiTS and SegTHOR datasets demonstrated that our approach could significantly improve segmentation accuracy for anisotropic data.

## REFERENCES

- [1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [2] Z. Meng, Y. Yang, K. Cheng, G. Chen, L. Wang, and W. Li, "A huge malignant peripheral nerve sheath tumor with hepatic metastasis arising from retroperitoneal ganglioneuroma," *Oncology Letters*, vol. 5, no. 1, pp. 123–126, 2013.

- [3] S. Gaiani, N. Celli, F. Piscaglia, L. Cecilion, F. Losinno, F. Giangregorio *et al.*, “Usefulness of contrast-enhanced perfusional sonography in the assessment of hepatocellular carcinoma hypervascular at spiral computed tomography,” *Journal of Hepatology*, vol. 41, no. 3, pp. 421–426, 2004.
- [4] G. Chlebus, H. Meine, J. H. Moltz, and A. Schenk, “Neural network-based automatic liver tumor segmentation with random forest-based candidate filtering,” *arXiv preprint arXiv:1706.00842*, 2017.
- [5] E. Vorontsov, A. Tang, C. Pal, and S. Kadoury, “Liver lesion segmentation informed by joint liver segmentation,” in *Proc. IEEE Int’l Sym. Biomedical Imaging*. IEEE, 2018, pp. 1332–1335.
- [6] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, “H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes,” *IEEE Trans. Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.
- [7] C. Liu, D. Cui, D. Shi, Z. Hu, Y. Qin, and J. Lang, “Automatic liver segmentation in CT volumes with improved 3D U-Net,” in *Proceedings of the 2nd International Symposium on Image Computing and Digital Medicine*. ACM, 2018, pp. 78–82.
- [8] S. Liu, D. Xu, S. K. Zhou, O. Pauly, S. Grbic, T. Mertelmeier *et al.*, “3D anisotropic hybrid network: Transferring convolutional features from 2D images to 3D anisotropic volumes,” in *Proc. Int’l Conf. Medical Image Computing and Computer Assisted Intervention*. Springer, 2018, pp. 851–858.
- [9] Z. Deng, Q. Guo, and Z. Zhu, “Dynamic regulation of level set parameters using 3D convolutional neural network for liver tumor segmentation,” *Journal of Healthcare Engineering*, pp. 1–17, Feb. 2019.
- [10] X. Han, “Automatic liver lesion segmentation using a deep convolutional neural network method,” *arXiv preprint arXiv:1704.07239*, 2017.
- [11] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou *et al.*, “The liver tumor segmentation benchmark (LiTS),” *arXiv preprint arXiv:1901.04056*, 2019.
- [12] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.
- [13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. European Conf. Computer Vision*, 2018, pp. 801–818.
- [14] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa *et al.*, “Attention U-Net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
- [15] Q. Shao, L. Gong, K. Ma, H. Liu, and Y. Zheng, “Attentive CT lesion detection using deep pyramid inference with multi-scale booster,” in *Proc. Int’l Conf. Medical Image Computing and Computer Assisted Intervention*. Springer, 2019, pp. 301–309.
- [16] J. Yang, X. Huang, B. Ni, J. Xu, C. Yang, and G. Xu, “Reinventing 2D convolutions for 3D medical images,” *arXiv preprint arXiv:1911.10477*, 2019.
- [17] Q. Yu, Y. Xia, L. Xie, E. K. Fishman, and A. L. Yuille, “Thickened 2D networks for efficient 3D medical image segmentation,” *arXiv preprint arXiv:1904.01150*, 2019.
- [18] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [19] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int’l Conf. Medical Image Computing and Computer Assisted Intervention*. Springer, 2015, pp. 234–241.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [21] Z. Wang and G. Wang, “Triplanar convolutional neural network for automatic liver and tumor image segmentation,” *International Journal of Production Economics*, vol. 14, no. 12, pp. 3151–3158, 2018.
- [22] S. H. Chung, K. H. Gan, A. Achuthan, and R. Mandava, “Liver tumor segmentation using triplanar convolutional neural network: A pilot study,” in *10th International Conference on Robotics, Vision, Signal Processing and Power Applications*. Springer, 2019, pp. 607–614.
- [23] J. Zhang, Y. Xie, P. Zhang, H. Chen, Y. Xia, and C. Shen, “Light-weight hybrid convolutional network for liver tumour segmentation,” in *Proc. Int’l Joint Conf. on Artificial Intelligence*, 2019, pp. 10–16.
- [24] W. Bae, S. Lee, G. Park, H. Park, and K. Jung, “Residual CNN-based image super-resolution for CT slice thickness reduction using paired CT scans: Preliminary clinical validation,” *Medical Imaging with Deep Learning*, 2018.
- [25] R. Ge, G. Yang, C. Xu, Y. Chen, L. Luo, and S. Li, “Stereo-correlation and noise-distribution aware ResVoxGAN for dense slices reconstruction and noise reduction in thick low-dose CT,” in *Proc. Int’l Conf. Medical Image Computing and Computer Assisted Intervention*. Springer, 2019, pp. 328–338.
- [26] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *arXiv preprint arXiv:1706.05587*, 2017.
- [27] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results,” <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [28] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Fourth International Conference on 3D Vision*. IEEE, 2016, pp. 565–571.
- [29] R. Trullo, C. Petitjean, B. Dubray, and S. Ruan, “Multiorgan segmentation using distance-aware adversarial networks,” *Journal of Medical Imaging*, vol. 6, no. 1, p. 014001, 2019.
- [30] F. Chollet, “Keras,” 2015. [Online]. Available: <https://github.com/keras-team/keras>
- [31] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-Net: Learning dense volumetric segmentation from sparse annotation,” in *Proc. Int’l Conf. Medical Image Computing and Computer Assisted Intervention*. Springer, 2016, pp. 424–432.
- [32] Q. Liu, Q. Dou, L. Yu, and P. A. Heng, “MS-Net: Multi-site network for improving prostate segmentation with heterogeneous MRI data,” *IEEE Transactions on Medical Imaging*, 2020.
- [33] Y. Yuan, “Hierarchical convolutional-deconvolutional neural networks for automatic liver and tumor segmentation,” *arXiv preprint arXiv:1710.04540*, 2017.
- [34] T. He, J. Guo, J. Wang, X. Xu, and Z. Yi, “Multi-task learning for the segmentation of thoracic organs at risk in CT images,” in *Proceedings of the 2019 Challenge on Segmentation of Thoracic Organs at Risk in CT Images*, 2019.
- [35] S. Vesal, N. Ravikumar, and A. Maier, “A 2D dilated residual U-Net for multi-organ segmentation in thoracic CT,” *arXiv preprint arXiv:1905.07710*, 2019.
- [36] P. Chen, C. Xu, X. Li, Y. Ma, and F. Sun, “Two-stage network for OAR segmentation,” in *Proceedings of the 2019 Challenge on Segmentation of Thoracic Organs at Risk in CT Images*, 2019.
- [37] L. Zhang, L. Wang, Y. Huang, and H. Chen, “Segmentation of thoracic organs at risk in CT images combining coarse and fine network,” in *Proceedings of the 2019 Challenge on Segmentation of Thoracic Organs at Risk in CT Images*, 2019.
- [38] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl *et al.*, “nnU-Net: Self-adapting framework for U-Net-based medical image segmentation,” *arXiv preprint arXiv:1809.10486*, 2018.
- [39] S. Ghosh, A. Pal, S. Jaiswal, K. Santosh, N. Das, and M. Nasipuri, “SegFast-V2: Semantic image segmentation with less parameters in deep learning for autonomous driving,” *Int. J. Machine Learning and Cybernetics*, vol. 10, no. 11, pp. 3145–3154, 2019.
- [40] S. Chen, K. Ma, and Y. Zheng, “Med3D: Transfer learning for 3D medical image analysis,” *arXiv preprint arXiv:1904.00625*, 2019.