A Colorectal Coordinate-Driven Method for Colorectum and Colorectal Cancer Segmentation in Conventional CT Scans

Article *in* IEEE Transactions on Neural Networks and Learning Systems · April 2024

DOI: 10.1109/TNNLS.2024.3386610

CITATIONS
0

READS
23

14 authors, including:

Yingda Xia
Johns Hopkins University
54 PUBLICATIONS **1,823** CITATIONS
SEE PROFILE

Chen Zhihong
Guangzhou University
9 PUBLICATIONS **34** CITATIONS
SEE PROFILE

Jiawen Yao
Alibaba Group
63 PUBLICATIONS **2,025** CITATIONS
SEE PROFILE

Dakai Jin
Alibaba Group
104 PUBLICATIONS **1,823** CITATIONS
SEE PROFILE

A Colorectal Coordinate-Driven Method for Colorectum and Colorectal Cancer Segmentation in Conventional CT Scans

Lisha Yao, Yingda Xia, Zhihong Chen, Suyun Li, Jiawen Yao, Dakai Jin, Yanting Liang, Jiatai Lin, Bingchao Zhao, Chu Han, *Member, IEEE*, Le Lu, *Fellow, IEEE*, Ling Zhang, Zaiyi Liu, and Xin Chen

***Abstract*— Automated colorectal cancer (CRC) segmentation in medical imaging is the key to achieving automation of CRC detection, staging, and treatment response monitoring. Compared with magnetic resonance imaging (MRI) and computed tomography colonography (CTC), conventional computed tomography (CT) has enormous potential because of its broad implementation, superiority for the hollow viscera (colon), and convenience without needing bowel preparation. However, the segmentation of CRC in conventional CT is more challenging due to the difficulties presenting with the unprepared bowel, such as distinguishing the colorectum from other structures with similar appearance and distinguishing the CRC from the contents of the colorectum. To tackle these challenges, we introduce DeepCRC-SL, the first automated segmentation algorithm for CRC and colorectum in conventional contrast-enhanced CT scans. We propose a topology-aware deep learning-based approach, which builds a novel 1-D colorectal coordinate system and encodes each voxel of the colorectum with a relative position along the coordinate system. We then induce an auxiliary regression task to predict the colorectal coordinate value of each voxel, aiming to integrate global topology into the segmentation network and thus improve the colorectum's continuity. Self-attention layers are utilized to capture global contexts for the coordinate regression task and enhance the ability to differentiate CRC and colorectum tissues. Moreover, a coordinate-driven self-learning (SL) strategy is introduced to leverage a large amount of unlabeled data to improve segmentation performance. We validate the proposed approach on a dataset including 227 labeled and 585 unlabeled CRC cases by fivefold cross-validation. Experimental results demonstrate that our method outperforms some recent related segmentation methods and achieves the segmentation accuracy in DSC for CRC of 0.669 and colorectum of 0.892, reaching to the performance (at 0.639 and 0.890, respectively) of a medical resident with two years of specialized CRC imaging fellowship.**

***Index Terms*— Colorectal cancer (CRC), computed tomography (CT), image segmentation, self-attention, self-learning (SL), topology information.**

Manuscript received 24 November 2022; revised 31 December 2023; accepted 30 March 2024. This work was supported in part by the National Science Fund for Distinguished Young Scholars of China under Grant 81925023, in part by the Regional Innovation and Development Joint Fund of National Natural Science Foundation of China under Grant U22A20345, in part by the National Natural Scientific Foundation of China under Grant 82072090 and Grant 82371954, in part by the Natural Science Foundation for Distinguished Young Scholars of Guangdong Province under Grant 2023B1515020043, and in part by Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application under Grant 2022B1212010011. *(Lisha Yao and Yingda Xia contribute equally to this work.) (Corresponding authors: Ling Zhang; Zaiyi Liu; Xin Chen.)*

Lisha Yao, Bingchao Zhao, and Zaiyi Liu are with the School of Medicine, South China University of Technology, Guangzhou 510006, China, also with the Department of Radiology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080, China, and also with Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou 510080, China (e-mail: yaols2021@163.com; zyliu@163.com).

Yingda Xia, Dakai Jin, Le Lu, and Ling Zhang are with the DAMO Academy, Alibaba Group, New York, NY 10014 USA (e-mail: yingda.xia@alibaba-inc.com; Tiger.lelu@gmail.com; ling.z@alibaba-inc.com).

Zhihong Chen is with the Institute of Computing Science and Technology, Guangzhou University, Guangzhou 510006, China.

Suyun Li, Yanting Liang, and Chu Han are with the Department of Radiology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080, China, and also with Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou 510080, China.

Jiawen Yao is with the DAMO Academy, Alibaba Group, Hangzhou 310024, China.

Jiatai Lin is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, also with the Department of Radiology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080, China, and also with Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou 510080, China.

Xin Chen is with the School of Medicine, South China University of Technology, Guangzhou 510006, China, and also with the Department of Radiology, Guangzhou First People's Hospital, Guangzhou 510080, China (e-mail: wolfchenxin@163.com).

Digital Object Identifier 10.1109/TNNLS.2024.3386610

I. Introduction

COLORECTAL cancer (CRC), originating in the mucosal lining of the colon or rectum within the digestive system, is one of the most common malignancies with high incidence and mortality [1]. Computed tomography (CT) can perform cross-sectional imaging of the abdomen and pelvis, which not only provides a clear perspective of the colorectum and its surrounding organs, but also depicts the location, morphology, and infiltration depth of CRCs. These imaging characteristics play a crucial role in the diagnosis and treatment planning for CRCs [2]. With the rapid development of deep learning algorithms, automated CRC segmentation offers a great opportunity to achieving automation of CRC detection, staging, and treatment response monitoring [3]. Previous studies for CRC segmentation mainly focus on magnetic resonance imaging (MRI) and computed tomography colonography (CTC) scanning [4], [5], [6], [7], [8], [9], [10]. However, MRI usually

scans the rectum region only and mainly contributes to rectal cancer staging in clinical practice because it is generally considered inferior to CT for the hollow viscera (colon). For CTC scanning, a patient needs to take strong medication or drink contrast medium called gastrografin before a couple of days of the scanning, which is time-consuming and may lead to some cramping pains and other adverse reactions. Compared with these imaging modalities, automated segmentation in conventional contrast-enhanced CT scans without bowel preparation has broader benefits in CRC-related clinical practice.

In this work, we aim to develop a method for automated 3-D colorectum and CRC segmentation in routine contrast-enhanced CT scans. Despite the success of deep learning in various organ segmentation and tumor segmentation tasks, segmenting colorectum and CRCs in routine CT scans poses more significant challenges as follows.

1) The colorectum is intricately entangled with various adjacent organs and soft tissues within the abdomen. In CT images without bowel preparation, the colorectal segments may exhibit features resembling either a lumen-shaped or shriveled configuration, sharing imaging characteristics with the small intestine and soft tissues, respectively. Traditional convolutional neural networks, constrained by limitations in kernel size, typically capture localized structural information, hindering the acquisition of the topological information essential for accurate tracking and continuous segmentation of the colorectum. The discontinuous segmentation of colorectum may further cause the misdetection of CRCs.
2) In addition to the challenges posed by the variable shapes, sizes, and locations similar to most tumor types, CRCs may be obscured by collapsed segments and the contents within the colorectum, potentially resulting in detection failures. Moreover, detecting CRCs with attenuations similar to normal colorectal tissues also presents a considerable obstacle.
3) Well-annotated data are expensive to obtain in the field of medical image analysis. Especially for our task, annotating the colorectum is tedious and time-consuming, given its substantial voxel occupancy in CT images. Moreover, the annotation of CRCs demands a high level of expertise of the annotator.

To resolve these three challenges, we design a colorectal coordinate-driven network for colorectum and CRC segmentation. Inspired by the topological structure of the colorectum with a single-path and continuity, we propose the concept of *Colorectal Coordinate Transform*, which transforms the colorectum to a topology-aware coordinate map. To be specific, we establish a 1-D colorectal coordinate system based on the extracted centerline of the colorectum label, and then project each point in the colorectum label into the colorectal coordinate system, thus obtaining a colorectal coordinate map related to the coordinate values. The coordinate map is modeled as an auxiliary task by introducing a regression loss to improve the segmentation continuity of the colorectum. We also incorporate a self-attention layer into the proposed model to offer more global context for the coordinate regression task and improve the ability to distinguish tumor and normal soft tissues in 3-D volumes. Motivated by the success of semisupervised learning and especially self-learning (SL) approaches [11], [12], we further extend our framework with a coordinate-driven SL strategy that combines large-scale unlabeled data and limited labeled data to boost the segmentation performance. In addition to pseudo segmentation masks, we also simultaneously generate the pseudo coordinate map on the unlabeled data for the self-training under both segmentation loss and regression loss.

Preliminary results of this work was reported in [13]. The current paper introduces a new coordinate-driven SL strategy to further improve the segmentation performance; and validates the proposed method on a larger and partially annotated dataset ($n = 812$ versus $n = 107$ in [13]) by explicitly comparing with five supervised segmentation algorithms and five semisupervised segmentation algorithms as well. The contributions of this work are summarized as follows.

1) We propose a topology-aware model that fully leverages the topology information of the colorectum and utilizes the self-attention mechanism to capture global contexts for both colorectum and CRC segmentation.
2) We design a colorectal coordinate-driven SL strategy that utilizes pseudo colorectal coordinate maps of unlabeled data for model training, resulting in a significant improvement in the accuracy of CRC segmentation
3) Comparison results of the proposed method with five supervised methods on 227 labeled data, as well as the comparison with five semisupervised methods on 812 partially labeled data, demonstrate the superior performance of our approach in both colorectum and CRC segmentation.

II. Related Work

A. CRC Image Segmentation

Recently, various deep learning-based methods have been proposed for CRC image segmentation. In terms of rectal tumor segmentation in MRIs, 2-D-based segmentation networks that incorporated multiscale features are developed to accomplish this task [4], [5]. To fully leverage 3-D spatial information and interslice correlation, 3-D CNNs are further designed to improve the segmentation performance [6], [7], [8]. For colorectal (colon and rectum) cancer segmentation in CTC, a CNN with an attention mechanism is developed [9]. In [10], a generative adversarial network is employed as a postprocessing to refine the segmentation results of traditional deep learning networks. Besides CTC, a conventional CT dataset for CRC segmentation is provided by the medical segmentation decathlon [14] and arouses some interest [15], [16], but bowel preparation and contrast agent injection are preperformed for better visibility of the CRC. As mentioned before, routine/conventional CT scan without bowel preparation has a wider range of application compared to MRI and CTC, and is the targeted modality of this article.

B. Deep Segmentation With Medical Knowledge

Medical image segmentation is a widely investigated task in the field of medical image analysis, which can be mainly categorized into 2-D slice-wise approaches [17], [18], 2.5-D approaches [19], [20], and 3-D approaches [5], [21]. Recently, a self-configured framework nnUNet is developed, which achieves robust performances on various medical image datasets and sets up new state-of-the-arts [15].

Anatomical structural information has been proved useful to improve the performance of medical image segmentation [22]. To help small bowel segmentation, a topological loss is developed to predict the inner cylinder of the small bowel [23]. An elastic boundary projection is incorporated into a 3-D segmentation network to model the segmentation boundary [24]. By combining the deep neural network and the statistical shape model, a novel Bayesian model is proposed for pancreas segmentation [25]. Distance transform maps are introduced as additional knowledge for tubular structure segmentation [26]. In [27], a mesh representation is utilized to model the 3-D geometry of the pancreas-tumor to improve segmentation performance. The 3-D shapes of the pancreas-tumor is learned by deforming a surface mesh starting from an average shape to guide pancreas segmentation [28].

Our approach is a novel 3-D topology-aware deep learning approach for colorectum and CRC segmentation. Different from these previous studies, we leverage the single-path and continuous structure of the colorectum and design an auxiliary coordinate regression task to improve the colorectum and CRC segmentation performance.

C. Long-Range Dependency Modeling

Long-range dependency modeling is a heated topic in computer vision and medical vision, which aims to enhance the global contexts in deep learning models [29], [30]. In medical image segmentation, recent studies either directly utilize vision transformers as the backbone, or adopt a hybrid CNN and transformer architecture. The UNETR method employs a transformer as the encoder followed by a CNN decoder [31]. This model is further extended by a shifted windows (SWINs) operation [32] [33]. In addition, the pretraining of transformers are demonstrated effective [34].

In terms of hybrid CNN-transformer models, the TransUNet method [35] encodes the CNN features using transformer and decodes the transformer and CNN features for multiorgan segmentation. A segmentation model with two parallel branches combining CNN and transformer is provided to capture long-range dependency and local features at the same time [36]. In addition, attention modules are integrated into both the encoder and decoder of U-shaped CNN for MRI image segmentation [37]. These hybrid models take advantage of both the CNN and the self-attention layers and achieve remarkable segmentation performance.

The global context might benefit the model's perception of CRC and colorectum. In our approach, we add self-attention layers into our network architecture and find it most beneficial to the segmentation of CRC while preserving the advantage of CNNs to distinguish local textual patterns for the CRC.

D. SL Strategy

As a common approach in semisupervised learning, the SL strategy leverages a large amount of unlabeled data to improve the performance of a fully supervised model that is trained on a limited number of labeled data [38]. The vanilla SL strategy involves a teacher–student model, which directly utilizes the knowledge of the teacher model to train the student model [12], [39], [40]. Other SL approaches integrate uncertainty estimation with the Bayesian model [41], [42] to add reliability to the pseudolabels. Instead of giving noisy pseudolabels, high-confidence unlabeled samples are selected to train the student model. In addition, a confidence-regularized SL strategy is proposed to mitigate the overconfident effect [43], [44].

In this work, we propose a coordinate-driven SL strategy that integrates the "pseudo coordinate map" into the entire regime, adding more flexibility to our framework on unlabeled data to improve the overall performance.

III. Methodology

We aim to do both colorectum and CRC segmentation in 3-D venous-phase CT scans. Fig. 1 depicts an overview of the proposed colorectal coordinate-driven segmentation framework. This framework includes a teacher model and a student model, which share the same network architecture and output branches. We first introduce our key approach *Colorectal Coordinate Transform* that transforms the colorectum labels to colorectal coordinate maps in Section III-A. We then describe our network architecture with self-attention layers and how to utilize colorectal coordinate maps in an auxiliary regression task in Section III-B. Finally, we introduce the coordinate-driven SL strategy by utilizing both labeled and unlabeled data in Section III-C.

The general math notations are introduced as follows. For each labeled CRC case, the 3-D CT volume in the venous phase and corresponding voxel-wise label can be described as $\mathbf{X}$ and $\mathbf{Y}$, respectively. We denote the labeled sample pairs as $\mathbf{M} = \{(\mathbf{X}_i \in \mathbb{R}^{H_i \times W_i \times D_i}, \mathbf{Y}_i \in \mathbb{N}^{H_i \times W_i \times D_i}) | i = 1, 2, \ldots, M\}$, where $\mathbf{X}_i$ and $\mathbf{Y}_i$ have the same dimensional size (H_i, W_i, D_i). $\mathbb{N} = \{0, 1, 2\}$ denotes the segmentation targets, i.e., background, colorectum, and tumor (CRC). M is the number of the labeled CRCs cases. The unlabeled dataset is denoted as $\mathbf{U} = \{\mathbf{X}_j \in \mathbb{R}^{H_j \times W_j \times D_j} | j = 1, 2, \ldots, N\}$ with N unlabeled cases.

A. Colorectal Coordinate Transform

To improve the continuity of colorectum segmentation and consequently improve the detection of CRC, we develop a *Colorectal Coordinate Transform* that transforms a colorectum label into a colorectal coordinate map. Fig. 1 depicts the diagram of the *Colorectal Coordinate Transform*, which is described as follows.

1) Colorectum Centerline: We employ an automatic centerline extraction algorithm [45] to extract the centerline $\mathbf{C}$ of each colorectum label $\mathbf{Y}$. First, a set of fuzzy centers of maximal balls (fCMBs) $P = \{p_1, p_2, \ldots, p_n\}$ is obtained,

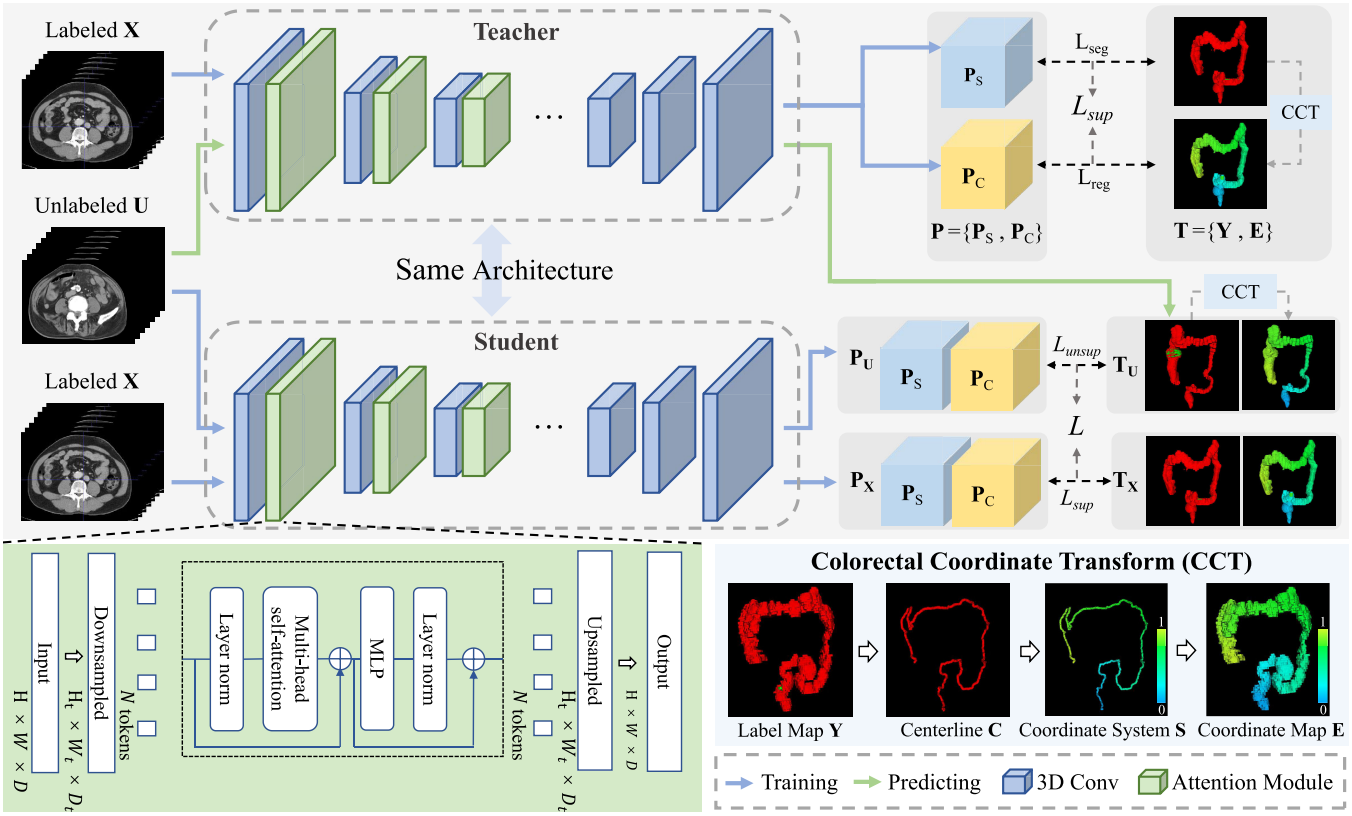

Fig. 1. Overview of the proposed colorectal coordinate-driven framework for colorectum and CRC segmentation. The teacher and student models share the same architecture with self-attention layers during downsampling, and include two outputs: the segmentation prediction $\mathbf{P}_s$ and the coordinate map prediction $\mathbf{P}_c$. The teacher model is trained on labeled data using the supervised loss, which is a weighted sum of the segmentation loss $\mathcal{L}_{\text{seg}}$ and the regression loss $\mathcal{L}_{\text{reg}}$. Subsequently, pseudolabels for unlabeled data are predicted by the teacher model, and corresponding coordinate maps are constructed from the pseudo labels. The student model is trained on both labeled and pseudolabeled data, optimizing by minimizing a combined loss comprising both the supervised loss $\mathcal{L}_{\text{sup}}$ and unsupervised loss $\mathcal{L}_{\text{unsup}}$.

where a voxel p is a fCMB if it holds the following condition for each 26-neighbor q:

$$\text{FDT}(q) - \text{FDT}(p) < \frac{1}{2}(\mu_\delta(p) + \mu_\delta(q))|p - q| \tag{1}$$

where $\text{FDT}(p)$ denotes the fuzzy distance transform [46] at voxel p. $\mu_\delta(p)$ is a membership function [45]. The algorithm starts with a root voxel r as the initial seed point, and connects the seed point to the farthest fCMB from it as the centerline. To find the farthest fCMB, the geodesic distance (GD) from r to each fCMB voxel p is computed as

$$\text{GD}(p) = \min_{\pi \in \Pi_p} \sum_{i}^{l-1} |p_i - p_{i-1}| \tag{2}$$

where $\pi = \langle r, p_1, p_2, \ldots, p_{l-1} \rangle$ is an ordered sequence of voxels. Π_p is the set of the paths from r to each fCMB voxel p. p_{i-1} and p_i are 26-adjacent. The farthest fCMB is selected as follows:

$$v = \arg\max_{p \in P} \text{GD}(p). \tag{3}$$

To ensure medialness of the centerline, the centerline is computed as a minimum cost path from v to r, which can be described as follows:

$$\mathbf{C} = \arg\min_{\pi \in \Pi_{v,r}} f(\pi) \tag{4}$$

where $\pi = \langle r, p_1, p_2, \ldots, p_{l-1} \rangle$ is one of the geodesic paths from v to r. $f(\pi)$ is a path cost function, which is defined as

$$f(\pi) = \sum_{i}^{l-1} \text{SC}(p_i - p_{i-1}) \tag{5}$$

where $\text{SC}(p_i - p_{i-1})$ is a step-cost between two 26-adjacent voxels, which is defined as follows:

$$\text{SC}(p_i - p_{i-1}) = \frac{|p_i - p_{i-1}|}{\varepsilon + ((\text{LSF}(p_i) - \text{LSF}(p_{i-1}))/2)^2} \tag{6}$$

where LSF is a local significance factor [47]. ε is set to 0.01 to overcome numerical computational difficulties.

The automatic algorithm ensures $\mathbf{C}$ to be one-voxel thick, 26-connected, and in the format of 3-D volume with the same shape as $\mathbf{X}$ and $\mathbf{Y}$.

2) Colorectal Coordinate System: We transform the extracted centerline $\mathbf{C}$ into a colorectal coordinate system $\mathbf{S}$. Since the correct centerline $\mathbf{C}$ is one-voxel thick and 26-connected, we exploit a minimum cost path algorithm to track $\mathbf{C}$ from the lowest foreground position t and set up a 1-D colorectal coordinate system, which is normalized to range from 0 to 1.

3) Colorectal Coordinate Map: We establish a colorectal coordinate map $\mathbf{E}$ by projecting each foreground voxel in $\mathbf{Y}$ into the colorectal coordinate system $\mathbf{S}$, where $\mathbf{E}$ is initialized

Algorithm 1 Colorectal Coordinate Transform

Require: Ground truth label map $\mathbf{Y} \in \mathbb{N}^{H\times W\times D}$
Ensure: Coordinate map $\mathbf{E} \in [0, 1]^{H\times W\times D}$
Initialize a root voxel r on $\mathbf{Y}$
Find the fCMB voxel v that is the farthest from r
Extract the centerline (1-voxel thick) $\mathbf{C}$ by connecting r and v using a minimum cost path
Find the lowest foreground position t on $\mathbf{C}$
Initialize zero map $\mathbf{E}$ with the same shape of $\mathbf{Y}$ and set $\mathbf{E}^t \leftarrow 1$
while $\exists$ unvisited position k in the 26-connectivity of t **do**
    $\mathbf{E}^k \leftarrow \mathbf{E}^t + 1$
    $t \leftarrow k$
Normalize $\mathbf{E}$ to the range of $[0, 1]$: $\mathbf{E} \leftarrow \frac{\mathbf{E}}{\max_i(\mathbf{E}^i)}$.
for each foreground position p on the label map $\mathbf{Y}$ **do**
    Find its nearest point q on the centerline $\mathbf{C}$
    $\mathbf{E}^p \leftarrow \mathbf{E}^q$
return $\mathbf{E}$

as an all-zero matrix with the same shape as $\mathbf{Y}$. For each foreground point p, we find the nearest point q on $\mathbf{S}$, and update $\mathbf{E}^p$ with the corresponding coordinate value of q. After this step, a colorectal coordinate map $\mathbf{E}$ of the ground-truth label map $\mathbf{Y}$ is obtained.

The overall algorithm of the *Colorectal Coordinate Transform* is summarized in Algorithm 1.

B. Topology-Aware Segmentation Framework

1) Network Architecture: Our network employs the nnUNet as the backbone, which includes an encoder and a decoder with five stages, respectively. Each stage has a stacked convolution layer with two convolution blocks. Each block includes a convolution operation, an instance norm operation, and a rectified linear unit operation. In the encoder, each stacked convolution layer is followed by a downsampling operation. Symmetrically, an upsampling operation is employed after each stacked convolution layer in the decoder to restore the original resolution of feature maps. The skip-connections deliver precision localization information of images to the decoder to refine the segmentation performance. The kernel sizes of the convolution operations are $1 \times 3 \times 3$ and $3 \times 3 \times 3$, and the strides of downsampling operations are $1 \times 2 \times 2$ and $2 \times 2 \times 2$. This network produces the segmentation prediction $\mathbf{P}_s$ and the coordinate map prediction $\mathbf{P}_c$ simultaneously.

Regular UNet-based segmentation networks are susceptible to local context changes and are inherently limited to receptive fields, which would be an obstacle to capturing global contexts for colorectum and CRC segmentation. To alleviate this problem, we integrate a self-attention layer to help the network understand the colorectum's topology globally.

As displayed in Fig. 1, we add a self-attention block after each downsampling block in the encoder. In each self-attention block, we first downsample the feature map to a fixed spatial size (H_0, W_0, D_0) and reshape it to a sequence of tokens $\mathbf{F}_{\text{in}}$ with a length of $\mathcal{T} = H_0 \times W_0 \times D_0$. We then project $\mathbf{F}_{\text{in}}$ to $\mathcal{D}$ dimensions with a trainable linear projection f_{proj} and enrich the projected tokens with a learnable positional embedding $\mathbf{F}_{\text{pos}}$. This process is denoted as follows:

$$\mathbf{Z}_0 = f_{\text{proj}}(\mathbf{F}_{\text{in}}) + \mathbf{F}_{\text{pos}} \tag{7}$$

where $\mathbf{F}_{\text{pos}} \in \mathbb{R}^{\mathcal{T}\times\mathcal{D}}$. We then forward $\mathbf{Z}_0$ to the multihead self-attention (MSA) block, which includes L layers of MSA and multilayer perceptron (MLP) as follows:

$$\mathbf{Z}'_l = \text{MSA}\big(\text{LN}\big(\mathbf{Z}'_{l-1}\big)\big) + \mathbf{Z}'_{l-1}, \quad l = 1, \ldots, L \tag{8}$$

$$\mathbf{Z}_l = \text{MLP}\big(\text{LN}\big(\mathbf{Z}'_l\big)\big) + \mathbf{Z}'_l, \quad l = 1, \ldots, L \tag{9}$$

$$\mathbf{F}_{\text{out}} = \text{LN}\big(\mathbf{Z}^0_L\big) \tag{10}$$

where LN represents layer normalization.

After the MSA calculation, we resample the output $\mathbf{F}_{\text{out}}$ back to the size of (H_0, W_0, D_0) and then upsample it with the original size of the feature map. Finally, we employ a convolution operation to merge the convolution and the self-attention feature maps, and feed the merged feature map into the next resolution stage. This model is named "DeepCRC."

2) Loss Function: As shown in Fig. 1, the segmentation model includes two outputs, i.e., the segmentation prediction $\mathbf{P}_s$ and the coordinate map prediction $\mathbf{P}_c$. To evaluate the difference between the ground-truth label and the predicted label, a segmentation loss $\mathcal{L}_{\text{seg}}$ consisting of cross-entropy loss $\mathcal{L}_{\text{CE}}$ and Dice loss $\mathcal{L}_{\text{dc}}$ is used, which can be written as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{n}^{\mathbb{N}}\sum_{t}^{\text{HWD}} \mathbb{1}\big(\mathbf{Y}^t = n\big)\log\big(\mathbf{P}_s^{t,n}\big) \tag{11}$$

$$\mathcal{L}_{\text{dc}} = -2\sum_{n}^{\mathbb{N}} \frac{\sum_t \mathbb{1}(\mathbf{Y}^t = n)\mathbf{P}_s^{t,n}}{\sum_t \mathbb{1}(\mathbf{Y}^t = n) + \sum_t \mathbf{P}_s^{t,n}} \tag{12}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, and $\mathbf{P}_s^{t,n}$ is the segmentation probability output of the nth class at the tth voxel

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{dc}}. \tag{13}$$

In addition, we introduce a regression loss $\mathcal{L}_{\text{reg}}$ that minimizes the difference between the predicted coordinate map $\mathbf{P}_c$ and the generated colorectal coordinate map $\mathbf{E}$, which is defined as

$$\mathcal{L}_{\text{reg}} = \sum_{t}^{\text{HWD}} \left\|\mathbf{P}_c{}^t - \mathbf{E}^t\right\|^2. \tag{14}$$

Finally, the loss function of the teacher model trained with the labeled data can be given as follows:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{seg}} + \alpha\mathcal{L}_{\text{reg}} \tag{15}$$

where α is a hyperparameters to control the weights of the segmentation loss and the regression loss.

C. Coordinate-Driven SL Strategy

1) Training Strategy: Compared to previous SL approaches that rely on plain pseudolabels or confidence-regularized labels, we introduce a coordinate-driven SL strategy to benefit from the colorectal topology knowledge in the large unlabeled data. As illustrated in Fig. 1, we first train a teacher model with labeled data. We then infer the large unlabeled data using

TABLE I
CLINICAL CHARACTERISTICS OF THE CRC PATIENTS IN THIS STUDY

		Labeled data (n=227)	Unlabeled data (n=585)
Gender, No.	Male	154	350
	Female	73	235
Age, years	Mean±Std	59±19	55±21
	Range	5-89	3-89
	Median	63	61
Tumor stage, No.	T1	7	18
	T2	35	81
	T3	149	427
	T4	31	47
	Missing data	5	12

the teacher model to generate pseudolabels of colorectum and CRC. To provide the topology knowledge of the colorectum, we build the colorectum coordinate maps from these pseudolabels as described in Section III-A. Taken together, the unlabeled data are labeled with pseudo colorectum and CRC labels, as well as the generated colorectal coordinate maps. As such, the student model is subsequently forced to learn from the additional topology knowledge provided by the coordinate maps. Moreover, the student's knowledge is expanded through learning on the large data that contains more colorectum and CRC variations, allowing the student to learn beyond his teacher to be capable of segmenting more challenging cases.

Finally, we train the student DeepCRC model from scratch on the combination of labeled and pseudolabeled data. Note that such a training strategy is found to be more effective than initializing the student model with the teacher model or first pretraining on unannotated data and then finetuning on annotated data [11]. We name this strategy "DeepCRC-SL."

2) Loss Function: The loss function of the student model is the sum of a supervised loss $\mathcal{L}_{\text{sup}}$ and an unsupervised loss $\mathcal{L}_{\text{unsup}}$ with a balance weight β as follows:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \beta \mathcal{L}_{\text{unsup}} \tag{16}$$

where $\mathcal{L}_{\text{unsup}}$ can be written as follows:

$$\mathcal{L}_{\text{unsup}} = \begin{cases} \mathcal{L}_{\text{seg}} + \alpha \mathcal{L}_{\text{reg}}, & \text{if } \mathbf{P}_s \neq \text{ None} \\ \mathcal{L}_{\text{seg}}, & \text{otherwise.} \end{cases} \tag{17}$$

For the cases with significant adhesive or clustered colorectum segmentation, the generation of the colorectal coordinate maps $\mathbf{P_s}$ is unavailable, and only the segmentation loss $\mathcal{L}_{\text{seg}}$ is used as $\mathcal{L}_{\text{unsup}}$.

IV. EXPERIMENTS AND RESULTS

A. Experimental Data and Annotation

To validate the segmentation performance of our proposed method, we collect 3-D CT volumes from 812 patients with CRC at Guangdong Provincial People's Hospital, Guangzhou, China. Table I lists the detailed clinical information of the 812 CRC patients. These patients are scanned by a GE, Philips, or Siemens scanner with the tube voltage at 120 kVp and tube current at 130–250 mAs. Before the scanning, the patients are injected with contrast agents at the rate of 2.5–3.5 mL/s. All CT scans are reconstructed with a shape of $512 \times 512 \times D$. D is varied from 76 to 672. The median voxel spacing of these CT volumes is $0.75 \times 0.75 \times 1.83$ mm^3. A radiologist with ten years of experience in diagnostic radiology annotates the colorectum and CRC of 227 CT volumes using ITK-SNAP software [48]. For colorectum annotation, the radiologist reviews the 3-D CT images slice-by-slice and carefully traces the correct shape of the colorectum from scratch manually. The CRC is annotated by referring to the corresponding clinical and pathological reports, as well as other CT phases if necessary, and further confirmed by a senior radiologist with 23-year experience in CRC imaging. In addition, for the interobserver experiment, a medical resident after two years of specialized CRC imaging fellowship annotates the colorectum in 30 cases and CRC in 107 cases with only venous phase CT provided.

B. Implementation Details

We conduct fivefold cross-validation on the labeled 227 cases. For the DeepCRC network training, all images are resampled to the spacing of $2 \times 2 \times 5$ mm^3 (in $[H, W, D]$) and randomly cropped into patches of $160 \times 160 \times 80$ voxels before being fed into the network. As for the self-attention layers, we resize the feature map from each encoder block to a size of (10, 10, 10) and acquire 1000 tokens for multihead attention. The dimension $\mathcal{D}$ of the embedded tokens is set to 256. In practice, we increase the colorectal coordinate map **E** to **E+1** and thus **E** varies in the range $[1, 2]$ to distinguish between the coordinate starting point (i.e., 1) and the background (i.e., 0). For SL, the pseudolabels of the unlabeled data utilized in each fold are predicted by the teacher model of the same fold. Note that the pseudolabeled data are only used in model training (not in validation). In the testing phase, all testing images are first resampled to the training spacing before being input into the network. We then resample the network predictions back to the original spacing before evaluating the performance. Following nnUNet and for fair comparison with nnUNet, we used the same data augmentation strategy (random flip, random rotation, and 3-D elastic transformation), the same batch size (a batch size of 2), and the same length of training (1000 epochs). As for the teacher model, we found that 300 epochs are enough for model convergence, and further training does not benefit the model performance. In addition, the model is optimized by RAdam optimizer with an initial learning rate of 0.001, which is decayed following the polynomial decay policy with $(1 - \text{epoch}/\text{epoch}_{\text{max}})^{0.9}$.

Using 227 labeled data, we compare our DeepCRC model against five state-of-the-art supervised methods, i.e., nnUNet [15], DDT [26], Swin UNETR [33], 3-D-UXNet [50], and AttnUNet [51]. Furthermore, we compare the DeepCRC-SL method with five semisupervised segmentation methods using both the 227 labeled data and 585 unlabeled data. These comparison methods include SL [11], MT [52], UAMT [53], and EM [54]. We also use the nnUNet as the teacher model in the proposed coordinate-driven SL strategy to

TABLE II

ABLATION STUDY ASSOCIATED WITH DIFFERENT COMPONENTS BY FIVEFOLD CROSS-VALIDATION. nnUNet IS USED AS THE BASELINE. THE ABLATION COMPONENTS INCLUDE THE SELF-ATTENTION BLOCKS (DEFINED AS "ATTN."), THE COORDINATE-REGRESSION TASK WITH REGRESSION LOSS $\mathcal{L}_{\text{REG}}$, AND THE COORDINATE-DRIVEN SL STRATEGY (DEFINED AS "SL"). RESULTS OF THE DSC, HD95, AND MSD ARE REPORTED AS MEAN ± STD

				Tumor				Colorectum			
Baseline	Attn.	$\mathcal{L}_{reg}$	SL	DSC↑	HD95(mm)↓	MSD(mm)↓	p value	DSC↑	HD95(mm)↓	MSD(mm)↓	p value
✓				0.570±0.275	26.25±47.84	11.91±36.17	Reference	0.874±0.077	6.33±7.88	1.21±1.64	Reference
✓	✓			0.619±0.255	20.45±40.39	8.10±27.92	<0.001	0.861±0.088	8.71±11.22	1.63±2.46	<0.001
✓		✓		0.591±0.271	24.85±46.62	10.32±32.55	0.060	0.882±0.070	6.31±10.51	1.25±2.46	<0.001
✓	✓	✓		0.619±0.249	19.92±38.58	7.42±26.15	<0.001	0.883±0.069	**4.54±5.39**	**0.84±0.83**	<0.001
✓	✓		✓	0.662±0.220	18.18±37.62	6.33±20.01	<0.001	0.879±0.063	6.09±6.77	1.33±1.16	<0.001
✓		✓	✓	0.659±0.227	18.99±38.46	6.46±20.49	<0.001	0.893±0.050	4.69±4.70	1.07±0.75	<0.001
✓	✓	✓	✓	**0.669±0.217**	**17.60±36.34**	**6.13±23.93**	<0.001	**0.892±0.055**	5.02±8.46	1.04±2.14	<0.001

TABLE III

ABLATION STUDY ASSOCIATED WITH HYPERPARAMETER SELECTION BY THE FIRST-FOLD VALIDATION. RESULTS OF THE DSC VALUES ARE REPORTED AS MEAN ± STD

		Colorectum	Tumor
α	5	0.892±0.057	0.618±0.239
	10	0.893±0.051	0.619±0.241
	30	0.896±0.051	0.625±0.250
	50	0.899±0.045	**0.633±0.236**
	80	**0.902±0.042**	0.626±0.231
	100	0.898±0.051	0.597±0.621
β	0.1	0.894±0.048	0.652±0.209
	0.5	0.898±0.042	0.658±0.220
	0.8	0.900±0.046	0.649±0.225
	1.0	**0.900±0.046**	**0.659±0.216**
	2.0	0.898±0.047	0.647±0.228
	5.0	0.899±0.045	0.646±0.232

evaluate its performance, which is referred to as "nnDeepCRC-SL." Except for Swin UNETR, all models are implemented using the nnUNet as the backbone for a fair comparison. The Swin UNETR network is initialized using a pretrained model [49] provided in their official release and implemented with the official code base. The batch size, initial learning rate, and the number of epochs are set to 1, 0.0001, and 1000, respectively, for the best performance on our data. All of the experiments are implemented in Pytorch and conducted on GeForce RTX 3090 with 24 GB memory.

C. Evaluation Metrics

To quantitatively evaluate the segmentation accuracy, we employ the Dice-Sørensen coefficient (DSC) [55], 95% Hausdorff distance (HD95) [56], mean surface distance (MSD) [57], truth detection rate (TDR), and false-positive rate (FPR) of tumors. DSC is used to compare the overlap between the segmentation and the ground truth, which ranges from 0 to 1. A greater DSC indicates a higher overlap. 95% HD is a matrix to compute the 95th percentile of the distance between boundary points in segmentation and the ground truth. MSD measures the mean of the distances between all surface pixels in the segmentation and corresponding pixels in the ground truth. Smaller HD95 and MSD values represent more accurate segmentation. TDR is defined as the proportion of successful detection of CRCs, where we set the DSC of CRC > 0.01 as the threshold for success. FPR is defined as the ratio of the number of detected false-positive tumors to the total number of cases. Moreover, the Wilcoxon signed rank test is used to evaluate the statistical significance (determined by p value <0.05) of the DSC difference between compared methods.

D. Results and Discussion

1) Ablation Study: As mentioned above, our method integrates self-attention blocks, an auxiliary coordinate-regression task with a regression loss $\mathcal{L}_{\text{reg}}$, and a coordinate-driven SL strategy for colorectum and CRC segmentation. In this section, we conduct an ablation study to assess the contribution of each component and evaluate the impact of hyperparameter selection. In ablation experiments when some models do not use the regression loss, the hyperparameter α is set to 0. Table II lists the quantitative results of the ablation experiments on different component combination by fivefold cross-validation. For colorectum segmentation, since the auxiliary regression task provides additional topology and location information of the colorectum, the methods with regression-loss $\mathcal{L}_{\text{reg}}$ outperform the methods with the segmentation-loss-only counterpart. In terms of tumor segmentation, both the model with self-attention blocks and the model with $\mathcal{L}_{\text{reg}}$ outperform the original nnUNet by 4.9% and 2.1% in DSC, respectively. By adding additional unlabeled data, both the self-attention blocks and the auxiliary regression task combining with a SL strategy yield further improvements for colorectum and tumor segmentation, demonstrating the robustness of the self-attention blocks and the auxiliary regression task, as well as the effectiveness of a SL strategy. Specifically, the methods with regression-loss $\mathcal{L}_{\text{reg}}$ still obtains better performance on colorectum segmentation than the method with segmentation-loss-only $\mathcal{L}_{\text{seg}}$, which shows the effectiveness of the proposed coordinate-driven SL strategy. Furthermore, the DeepCRC-SL method leveraging all the proposed components achieves the most significant improvements over nnUNet, with DSC improvements of 9.9% ($p < 0.001$) and 1.8% ($p < 0.001$) for tumor and colorectum segmentation, respectively.

Table III shows the impact of different α and β values on the segmentation performance. The α is a critical factor for controlling the contribution of the proposed colorectal coordinate regression task to the segmentation task. The β is used to balance the weight between the labeled data and pseudolabeled data for model training. Specifically, the α is modified in the DeepCRC model. As the α increases

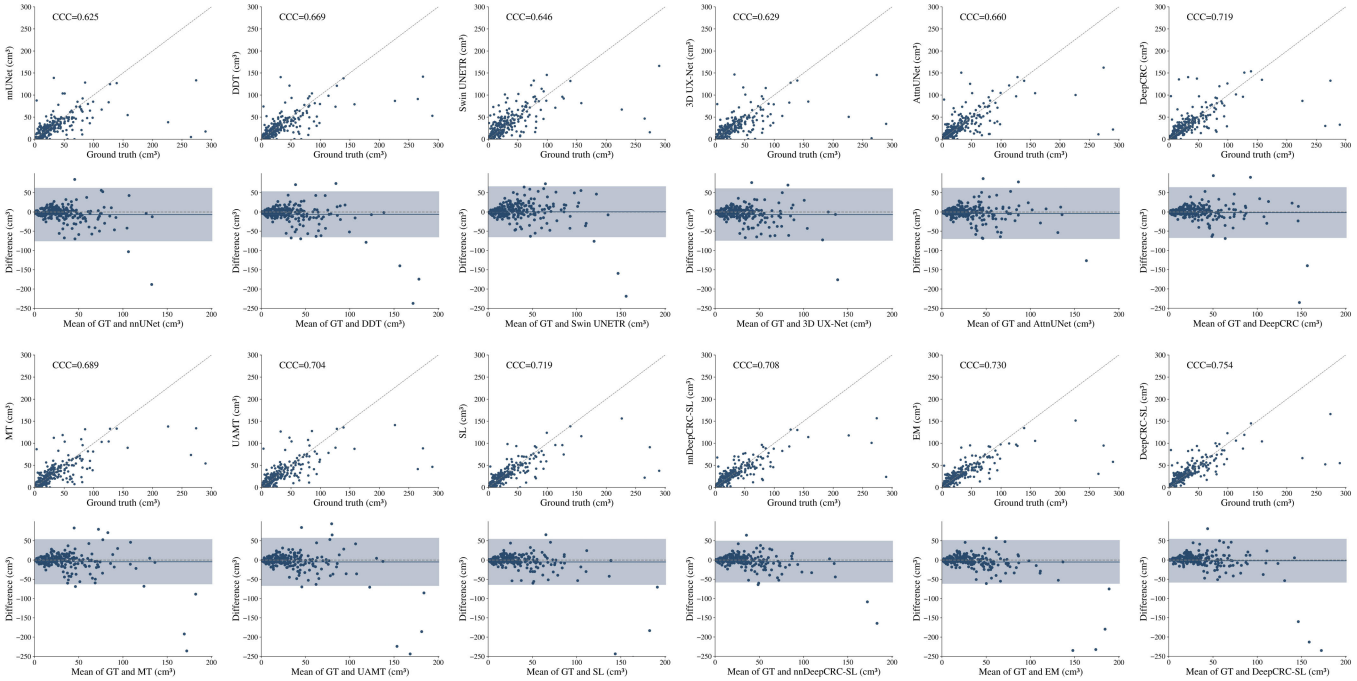

Fig. 2. Agreement between the tumor volumes of ground truth and those predicted by the different methods. The first row shows the concordance between the tumor volumes of ground truth and the different methods. Concordance correlation coefficients (CCCs) are also displayed. The second row shows the Bland–Altman plots. GT = ground truth.

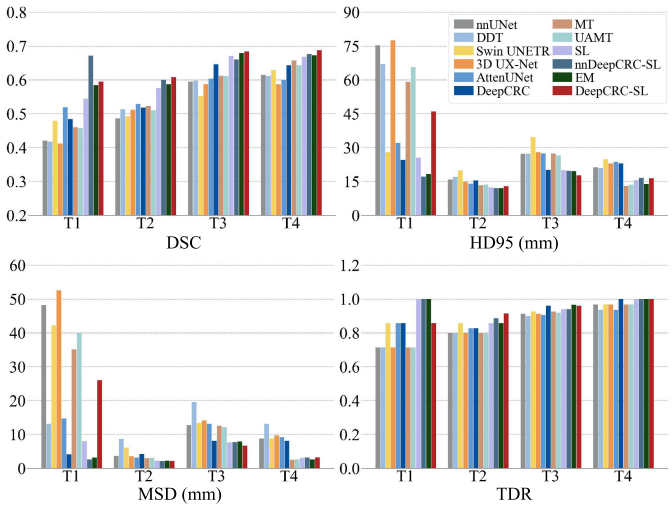

Fig. 3. Quantitative analysis of the segmentation performance on four tumor (T) stages. For the metric of DSC and TDR, the higher the better. For the metric of HD95 and MSD, the lower the better.

from 5 to 80, the DSC for the colorectum gradually increases, reaching its maximum when the α is set to 50. However, when the α is increased to 100, the DSC values for both colorectum and tumor segmentation become worsen. In addition, setting the β to 1 yields the best segmentation performance for both the colorectum and CRC. Therefore, we set the α to 50 and the β to 1 for all related experiments.

2) Comparison With State-of-the-Art Methods: We compare the performance of the DeepCRC against five supervised methods and evaluate the performance of DeepCRC-SL against five semisupervised methods. Quantitative results of these methods by fivefold cross-validation are summarized in Table IV. Among the supervised methods, the DeepCRC obtains the highest DSC values of 0.619 and 0.883 for tumor and colorectum segmentation, respectively. Compared to the nnUNet, both 3-D-UXNet and AttnUNet demonstrate no significant differences in the DSC values for colorectum segmentation ($p = 0.689$ and $p = 0.661$). With additional information of the distance transform map, DDT performs slightly better than nnUNet in colorectum segmentation ($p < 0.001$). Moreover, DeepCRC produces a DSC improvement of 0.8% on colorectum segmentation over DDT, mainly because DeepCRC is capable of capturing global topology information of colorectum instead of local topology information obtained by DDT. On the other hand, owing to the self-attention layers that leverage global contexts, the AttnUNet and DeepCRC achieve better performance than the nnUNet with 1.7% DSC improvement ($p = 0.030$) and 4.9% DSC improvement ($p < 0.001$) on tumor segmentation. In addition, the 3-D-UXNet yields similar DSC value on the tumor segmentation to the nnUNet ($p = 0.352$). The Swin UNETR is significantly worse for both tumor and colorectum segmentation and produces the lowest metrics.

By adding the unlabeled data, all the semisupervised methods perform similar or better tumor and colorectum segmentation performance than the nnUNet baseline. With the proposed coordinate-driven SL strategy, DeepCRC-SL further improves the segmentation performance, especially for the tumor, e.g., 5% absolute DSC value improvement compared to DeepCRC, and 9.9% improvement compared to the nnUNet baseline ($p < 0.001$). The SL and EM methods yield 7.8% and 8.6% absolute DSC value improvements on

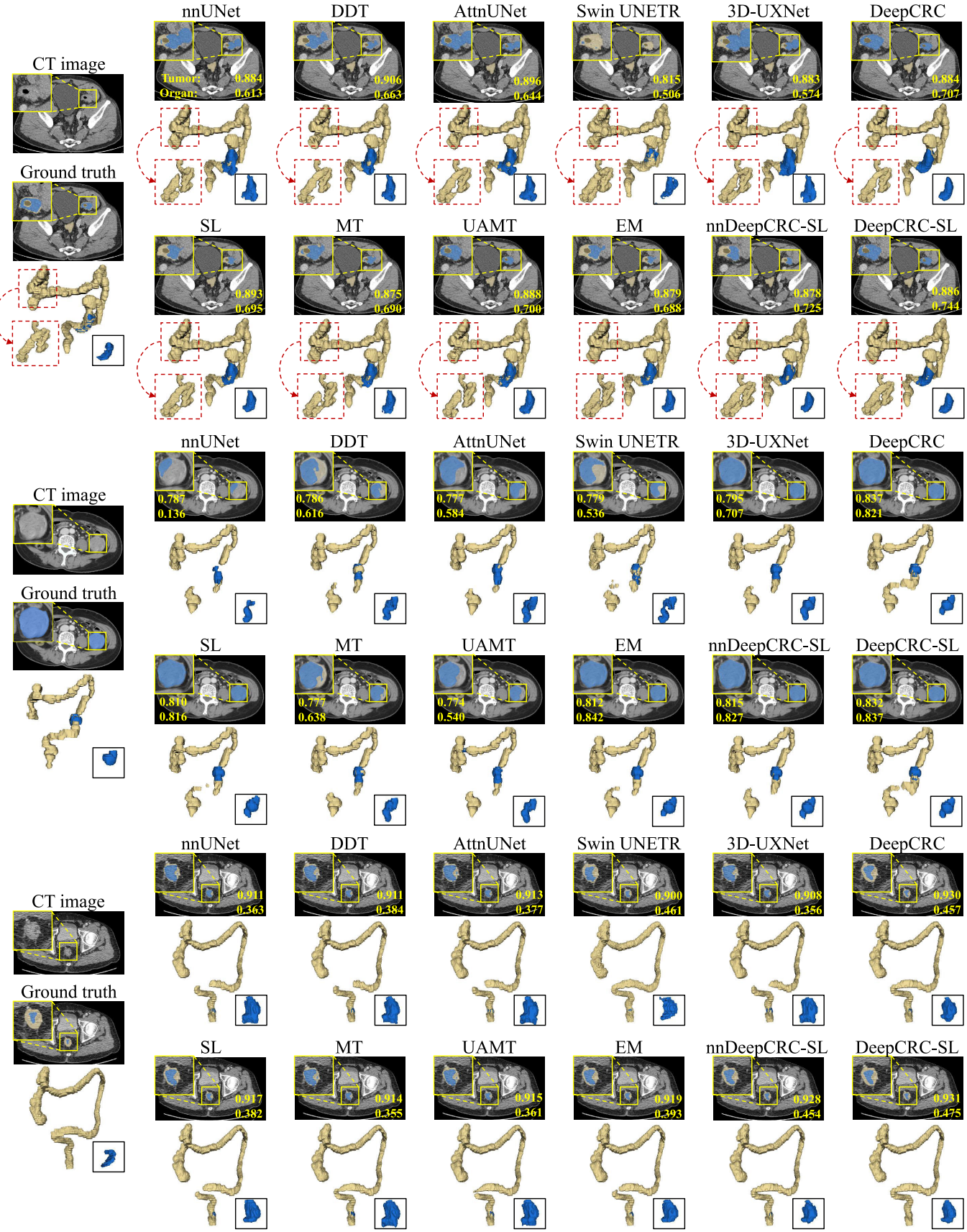

Fig. 4. Qualitative results of the colorectum (yellow) and colorectal tumor (blue) on three cases. The DSC values of the colorectum and colorectal tumor are included. For each case, we display the selected 2-D slice and 3-D segmentation results. Zoomed-in colorectum and tumor regions in 2-D or 3-D views are also displayed, where the 3-D view only displays the maximum component of tumors.

TABLE IV

COMPARISON OF DIFFERENT STATE-OF-THE-ART SEGMENTATION METHODS BY FIVEFOLD CROSS-VALIDATION. RESULTS OF THE DSC, HD95, AND MSD ARE SHOWN AS MEAN±STD

Methods	Tumor				Colorectum			
	DSC↑	HD95(mm)↓	MSD(mm)↓	p value	DSC↑	HD95(mm)↓	MSD(mm)↓	p value
nnUNet [15]	0.570±0.275	26.25±47.84	11.91±36.17	Reference	0.874±0.077	6.33±7.88	1.21±1.64	Reference
DDT [26]	0.576±0.274	25.97±49.31	12.48±37.96	0.944	0.877±0.078	5.96±8.03	1.16±1.90	0.001
Swin UNETR [49]	0.550±0.248	30.67±61.50	16.78±50.44	0.008	0.835±0.080	27.00±57.83	7.76±15.69	<0.001
3D-UXNet [50]	0.566±0.277	26.90±48.10	13.10±34.10	0.352	0.874±0.077	6.56±8.28	1.39±1.46	0.689
AttnUNet [51]	0.587±0.272	25.07±45.85	11.20±29.52	0.030	0.873±0.078	6.44±7.41	1.37±1.32	0.661
DeepCRC	0.619±0.249	19.92±38.58	7.42±26.15	<0.001	0.883±0.069	4.54±5.39	0.84±0.83	<0.001
MT [52]	0.598±0.265	24.17±44.75	10.43±33.45	<0.001	0.874±0.071	6.42±7.23	1.36±1.22	0.063
UAMT [53]	0.594±0.267	23.96±45.30	10.29±33.73	0.002	0.875±0.075	6.51±8.00	1.36±1.28	0.919
SL [11]	0.648±0.239	18.38±36.56	6.25±23.24	<0.001	0.892±0.053	4.77±5.30	1.07±0.78	<0.001
nnDeepCRC-SL	0.652±0.228	18.07±35.77	6.19±22.74	<0.001	0.890±0.053	5.08±5.64	1.13±0.88	<0.001
EM [54]	0.656±0.226	17.66±36.08	6.24±24.41	<0.001	**0.893±0.052**	4.63±5.37	1.05±0.79	<0.001
DeepCRC-SL	**0.669±0.217**	**17.60±36.34**	**6.13±23.93**	<0.001	0.892±0.055	5.02±8.46	1.04±2.14	<0.001
Inter-observer	0.639±0.280	26.12±48.25	15.64±40.47	-	0.890±0.041	**2.19±1.23**	**0.57±0.25**	-

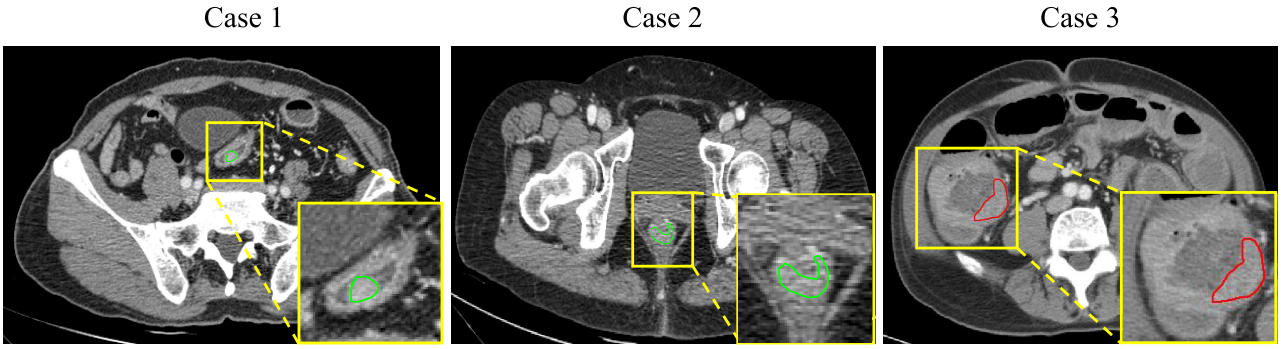

Fig. 5. Failure cases on the tumor segmentation. The green line indicates the annotation of ground truth, and the red line indicates the prediction of false-positive tumor.

tumor segmentation compared to nnUNet ($p < 0.001$ and $p < 0.001$). While the MT and UAMT obtain relative improvements in tumor segmentation performance compared to the nnUNet model ($p < 0.001$ and $p = 0.002$), the performance for colorectum segmentation does not demonstrate significant enhancements ($p = 0.063$ and $p = 0.919$). In addition, the nnDeepCRC-SL model achieves 0.4% absolute DSC value improvement than the SL model, which further validate the effectiveness of the proposed coordinate-driven SL strategy. Moreover, despite yielding larger HD95 and MSD values in colorectum segmentation, the SL, nnDeepCRC-SL, EM, and DeepCRC-SL methods reaches the human interobserver variability in both tumor and colorectum segmentation. Specifically, the DeepCRC-SL achieves the largest gains in tumor segmentation, and slightly worse than the EM model in colorectum segmentation.

In Fig. 2, correlation analysis (upper panel) shows that the DeepCRC-SL yields the highest agreement with the ground truth tumor volume (CCC = 0.754), followed by EM (CCC = 0.730). Bland–Altman plots (Fig. 2, lower panel) show that DeepCRC-SL (versus ground truth tumor volume) has the lowest bias among all methods. Fig. 3 illustrates the segmentation results of different methods stratified by the four tumor (T) stages. In general, as the T stage increases, the segmentation accuracy of each method becomes higher. Specifically, DeepCRC-SL exhibit superior performance over the other methods consistently in all T stages.

3) Qualitative Results: Fig. 4 shows qualitative results of three CRC cases to further analyze the segmentation performance of different methods. For each case, we displayed the selected 2-D image slices and 3-D segmentation results. Zoomed-in tumor regions in 2-D and 3-D views are also displayed. As shown, owing to introducing the proposed topology-aware colorectal coordinate-regression branch or the coordinate-driven SL strategy, the DeepCRC and DeepCRC-SL methods predict more correct topology or more continuous segmentation of the colorectum. In addition, more continuous segmentation of the colorectum yielded by the nnDeepCRC-SL in the third case also demonstrates the effectiveness of the coordinate-driven SL strategy. As for the tumor segmentation, some comparative methods may predict a small portion of the cancerous regions or misclassify the surrounding tissues as the tumor. The DeepCRC and DeepCRC-SL methods achieve the best performance in terms of ground-truth overlap and edge preservation among all comparison methods.

V. CONCLUSION

In this study, we propose a deep learning-based solution for 3-D colorectum and CRC segmentation in conventional contrast-enhanced CT scans. First, a novel auxiliary regression task is designed to help the segmentation network understand the continuity of the colorectum. Second, self-attention layers are integrated to provide global contexts for the regression task and help distinguish between the tumor and normal

TABLE V
TDR AND FPR OF TUMORS AT INSTANCE LEVEL BY DIFFERENT METHODS ON FIVEFOLD CROSS-VALIDATION. TDR: TRUTH DETECTION RATE. FPR: FALSE-POSITIVE RATE

		TDR	FDR
Supervised	nnUNet [15]	0.863	**0.313**
	DDT [26]	0.859	0.330
	Swin UNETR [49]	0.899	1.511
	3D-UXNet [50]	0.863	0.361
	AttnUNet [51]	0.894	**0.313**
	DeepCRC	**0.925**	0.357
Semi-supervised	MT [52]	0.855	0.352
	UAMT [53]	0.881	0.352
	SL [11]	0.934	0.225
	nnDeepCRC-SL	0.938	**0.216**
	EM [54]	0.943	0.247
	DeepCRC-SL	**0.947**	0.295

tissues. Third, a coordinate-driven SL strategy is introduced to utilize unlabeled data to further improve the segmentation performance. We conduct an ablation study to analyze the contributions of the above three components. Moreover, we compare the proposed DeepCRC and DeepCRC-SL methods with five supervised segmentation methods and five semisupervised segmentation methods. Experimental results show that our methods achieve the best performance for both colorectum and CRC segmentation, and DeepCRC-SL even reaches the human interobserver variability.

A major limitation of our approach is the generation of the colorectal coordinate maps. To ensure correct centerline extraction during *Colorectal Coordinate Transform*, the adhesion parts in the colorectum mask need to be manually erased. To make the method more practical, for the student model training, colorectal coordinate maps are not generated in the cases with severe discontinuous and adhesive colorectum segmentation. Therefore, the segmentation performance of the proposed model may be influenced by the proportion of coordinate maps in the training set. Nevertheless, the proposed method still yields more accurate segmentation compared to various state-of-the-art segmentation methods, especially for CRCs. Fig. 5 shows three segmentation failure cases by our proposed method or all comparison methods. Results show that CRCs with small size (Case 1) or low-contrast to surrounding tissues (Case 2) remain the challenges for detection and segmentation in conventional CT scans. As shown in Case 3, due to the partial volume effect induced by localized intestinal folding, the proposed model misidentifies the uneven enhancement of normal tissue in the ascending colon as a tumor. In addition, we acknowledge that, while our model achieves the highest tumor detection rate among all methods, its performance in suppressing false-positive tumor segmentation is relatively moderate, as listed in Table V. Future work will focus on resolving the above three problems and will further explore the potential of DeepCRC in the opportunistic screening of CRCs.

REFERENCES

[1] H. Sung et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, Feb. 2021.

[2] R. Labianca et al., "Early colon cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up," *Ann. Oncol.*, vol. 24, pp. VI64–VI72, Oct. 2013.

[3] G. Argilés et al., "Localised colon cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up," *Ann. Oncol.*, vol. 31, no. 10, pp. 1291–1305, Oct. 2020.

[4] J. Jian et al., "Fully convolutional networks (FCNs)-based segmentation method for colorectal tumors on T2-weighted magnetic resonance images," *Australas. Phys. Eng. Sci. Med.*, vol. 41, no. 2, pp. 393–401, Jun. 2018.

[5] S. Zheng et al., "MDCC-Net: Multiscale double-channel convolution U-Net framework for colorectal tumor segmentation," *Comput. Biol. Med.*, vol. 130, Mar. 2021, Art. no. 104183.

[6] Y.-J. Huang et al., "3-D RoI-aware U-Net for accurate and efficient colorectal tumor segmentation," *IEEE Trans. Cybern.*, vol. 51, no. 11, pp. 5397–5408, Nov. 2021.

[7] M. H. Soomro et al., "Automated segmentation of colorectal tumor in 3D MRI using 3D multiscale densely connected convolutional neural network," *J. Healthcare Eng.*, vol. 2019, pp. 1–11, Jan. 2019.

[8] Y. Jiang et al., "ALA-Net: Adaptive lesion-aware attention network for 3D colorectal tumor segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3627–3640, Dec. 2021.

[9] Y. Pei et al., "Colorectal tumor segmentation of CT scans based on a convolutional neural network with an attention mechanism," *IEEE Access*, vol. 8, pp. 64131–64138, 2020.

[10] X. Liu et al., "Accurate colorectal tumor segmentation for CT scans based on the label assignment generative adversarial network," *Med. Phys.*, vol. 46, no. 8, pp. 3532–3542, Aug. 2019.

[11] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 10687–10698.

[12] L. Zhang et al., "Robust pancreatic ductal adenocarcinoma segmentation with multi-institutional multi-phase partially-annotated CT scans," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 491–500.

[13] L. Yao et al., "DeepCRC: Colorectum and colorectal cancer segmentation in CT scans via deep colorectal coordinate transform," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2022, pp. 564–573.

[14] M. Antonelli et al., "The medical segmentation decathlon," 2021, *arXiv:2106.05735*.

[15] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," in *Proc. Nature Methods*, 2021, vol. 18, no. 2, pp. 203–211.

[16] Y. Tang et al., "Self-supervised pre-training of Swin transformers for 3D medical image analysis," 2021, *arXiv:2111.14791*.

[17] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8280–8289.

[18] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 5, pp. 1116–1126, May 2018.

[19] Y. Xia, F. Liu, Z. Zhu, E. K. Fishman, and A. L. Yuille, "Bridging the gap between 2D and 3D organ segmentation with volumetric fusion net," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.* Cham, Switzerland: Springer, 2018, pp. 445–453.

[20] 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. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 424–432.

[21] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[22] L. Liu, J. M. Wolterink, C. Brune, and R. N. J. Veldhuis, "Anatomy-aided deep learning for medical image segmentation: A review," *Phys. Med. Biol.*, vol. 66, no. 11, Jun. 2021, Art. no. 11TR01.

[23] S. Y. Shin, S. Lee, D. Elton, J. L. Gulley, and R. M. Summers, "Deep small bowel segmentation with cylindrical topological constraints," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 207–215.

[24] T. Ni, L. Xie, H. Zheng, E. K. Fishman, and A. L. Yuille, "Elastic boundary projection for 3D medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2104–2113.

[25] J. Ma, F. Lin, S. Wesarg, and M. Erdt, "A novel Bayesian model incorporating deep neural network and statistical shape model for pancreas segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 480–487.

[26] Y. Wang et al., "Deep distance transform for tubular structure segmentation in CT scans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3833–3842.

[27] J. Yao, J. Cai, D. Yang, D. Xu, and J. Huang, "Integrating 3D geometry of organ for improving medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 318–326.

[28] T. Zhao et al., "3D graph anatomy geometry-integrated network for pancreatic mass segmentation, diagnosis, and quantitative patient management," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13738–13747.

[29] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[30] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[31] A. Hatamizadeh et al., "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 574–584.

[32] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[33] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," 2022, *arXiv:2201.01266*.

[34] H. Cao et al., "Swin-Unet: Unet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*.

[35] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[36] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 14–24.

[37] Y. Gao, M. Zhou, and D. N. Metaxas, "UTNet: A hybrid transformer architecture for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 61–71.

[38] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, vol. 3, no. 2, p. 896.

[39] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "ST++: Make self-training work better for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4268–4277.

[40] W. Bai et al., "Self-supervised learning for cardiac MR image segmentation by anatomical position prediction," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 541–549.

[41] K. Wang et al., "Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 450–460.

[42] S. Mukherjee and A. Awadallah, "Uncertainty-aware self-training for few-shot text classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21199–21212.

[43] Y. Zou, Z. Yu, X. Liu, B. Kumar, and J. Wang, "Confidence regularized self-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2019, pp. 5982–5991.

[44] Z. Xie, E. Tu, H. Zheng, Y. Gu, and J. Yang, "Semi-supervised skin lesion segmentation with learning model confidence," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1135–1139.

[45] D. Jin, K. S. Iyer, C. Chen, E. A. Hoffman, and P. K. Saha, "A robust and efficient curve skeletonization algorithm for tree-like objects using minimum cost paths," *Pattern Recognit. Lett.*, vol. 76, pp. 32–40, Jun. 2016.

[46] P. K. Saha, F. W. Wehrli, and B. R. Gomberg, "Fuzzy distance transform: Theory, algorithms, and applications," *Comput. Vis. Image Understand.*, vol. 86, no. 3, pp. 171–190, Jun. 2002.

[47] D. Jin and P. K. Saha, "A new fuzzy skeletonization algorithm and its applications to medical imaging," in *Proc. Int. Conf. Image Anal. Process.*, 2013, pp. 662–671.

[48] P. A. Yushkevich et al., "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *NeuroImage*, vol. 31, no. 3, pp. 1116–1128, Jul. 2006.

[49] Y. Tang et al., "Self-supervised pre-training of Swin transformers for 3D medical image analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 20730–20740.

[50] H. H. Lee, S. Bao, Y. Huo, and B. A. Landman, "3D UX-Net: A large kernel volumetric Convnet modernizing hierarchical transformer for medical image segmentation," 2022, *arXiv:2209.15076*.

[51] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[52] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[53] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*. Cham, Switzerland: Springer, 2019, pp. 605–613.

[54] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2512–2521.

[55] W. R. Crum, O. Camara, and D. L. G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1451–1461, Nov. 2006.

[56] A. A. Taha and A. Hanbury, "An efficient algorithm for calculating the exact Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2153–2163, Nov. 2015.

[57] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, pp. 1–28, Dec. 2015.