3D Image Segmentation With Sparse Annotation by Self-Training and Internal Registration

Adeleh Bitarafan, Mahdi Nikdan, and Mahdieh Soleymani Baghshah

***Abstract*—Anatomical image segmentation is one of the foundations for medical planning. Recently, convolutional neural networks (CNN) have achieved much success in segmenting volumetric (3D) images when a large number of fully annotated 3D samples are available. However, rarely a volumetric medical image dataset containing a sufficient number of segmented 3D images is accessible since providing manual segmentation masks is monotonous and time-consuming. Thus, to alleviate the burden of manual annotation, we attempt to effectively train a 3D CNN using a sparse annotation where ground truth on just one 2D slice of the axial axis of each training 3D image is available. To tackle this problem, we propose a self-training framework that alternates between two steps consisting of assigning pseudo annotations to unlabeled voxels and updating the 3D segmentation network by employing both the labeled and pseudo labeled voxels. To produce pseudo labels more accurately, we benefit from both propagation of labels (or pseudo-labels) between adjacent slices and 3D processing of voxels. More precisely, a 2D registration-based method is proposed to gradually propagate labels between consecutive 2D slices and a 3D U-Net is employed to utilize volumetric information. Ablation studies on benchmarks show that cooperation between the 2D registration and the 3D segmentation provides accurate pseudo-labels that enable the segmentation network to be trained effectively when for each training sample only even one segmented slice by an expert is available. Our method is assessed on the CHAOS and Visceral datasets to segment abdominal organs. Results demonstrate that despite utilizing just one segmented slice for each 3D image (that is weaker supervision in comparison with the compared weakly supervised methods) can result in higher performance and also achieve closer results to the fully supervised manner.**

***Index Terms*—Deep learning, inter-slice registration, medical 3d image segmentation, self-training, sparse annotation, weakly supervised learning.**

I. Introduction

DEEP learning models for analyzing medical images have been recently popularized, and a broad range of biomedical tasks have been developed accordingly. Nowadays, there are different architectures and training procedures on various biomedical tasks such as image classification [1], object detection [2], image reconstruction [3], image registration [4], and image segmentation [5], [6]. Among them, the image segmentation plays an essential role in many biomedical applications involved in computer-aided diagnosis [7], simulation computed tomography imaging [8], and radiation therapy [9]. In recent years, many works have been presented for the segmentation of different anatomical organs including the brain [10], liver [11], lung [5] and cardiac image segmentation [12]. Nevertheless, the high performance of these works strongly relies on the availability of a dataset containing a sufficient amount of fully annotated images. However, manual segmentation of training data, particularly for 3D medical images like CT and MR, is extremely laborious and expensive since it is a boring and time-consuming task for annotators. Therefore, a large full annotated dataset rarely is accessible for the 3D image segmentation task. To tackle this challenge, self-supervised [6], semi-supervised [13]–[15], and weakly supervised [16]–[19] techniques have been proposed to address the pixel-level segmentation given sparse annotations.

All existing self-supervised and semi-supervised methods have been developed to tackle the problem of the limited annotation where the ground truth of few 3D images is available for training. These models first assign an initial annotation to unlabeled 3D samples. Then, the segmentation model follows an iterative process to refine synthetic labels. Studies [13] and [14] propose an iterative framework to segment the heart chambers in MR images, and the segmentation model benefits from the dense connected conditional random field (CRF) to refine the segmentation masks in each iteration. Besides, [14] trains three segmentation models for the axial, sagittal, and coronal axes. To tackle the weakness of propagating errors in the synthetic labels, the recent study [15] has been proposed to estimate the uncertainty of the synthetic labels during training and avoid the unreliable regions in synthetic segmentation masks. The recent approach in [10] utilizes a bulk of unlabeled MR images along with only one labeled 3D image as the source to leverage the limited annotation challenge. It learns transformations to align the source 3D image to that of unlabeled 3D images in order to generate synthetic labeled samples. However, all of these approaches require some fully segmented volumes as strong supervision for training.

On the other hand, weakly supervised methods have been presented that can be categorized according to the type of annotation they utilize, including image-level, box-level, scribble, and points (or clicks). Recently, numerous approaches have

Manuscript received June 10, 2020; revised September 26, 2020 and November 2, 2020; accepted November 15, 2020. Date of publication November 19, 2020; date of current version July 20, 2021. *(Corresponding author: Mahdieh Soleymani Baghshah.)*

The authors are with the Department of Computer Engineering, Sharif University of Technology, Tehran 1459973941, Iran (e-mail: bitarafan@ce.sharif.edu; nikdan@ce.sharif.edu; soleymani@sharif.edu).

Digital Object Identifier 10.1109/JBHI.2020.3038847

been proposed for coping with these types of annotations. The most prevalent ones focus on image-level annotation, which are generally based on multiple instance learning [20], [21] or class activation maps [22]. However, all these techniques rely on 2D medical images and can not leverage image-level annotation in 3D images. Recently, some weakly supervised segmentation techniques with annotations of scribble [16], [17], point [23], and bounding box [18], [19] are presented in association with 3D medical images. Most of them generally follow an iterative procedure that alternates between generating pseudo labels for unlabelled voxels and training a 3D segmentation network using synthesized ground-truths. The proposed methods in [16], [17] tackle scribble annotations that specify a small ratio of the target object's voxels in all slices of a 3D image. The method [17] suggests an iterative framework based on the GrabCut algorithm to complete the missing voxels and refine pseudo masks. Contrary to [17], [16] shows that adding a prior knowledge like object size and image tags as inequality constraints into the cross-entropy loss function can improve the performance of the segmentation networks in the presence of scribble-level annotations. The method introduced in [23] addresses 3D point annotations located on the organ's surface of interest. By these 3D points annotations, initial pseudo masks are generated using a random walker algorithm. To deal with bounding-box supervision in 3D biomedical images, [18], [19] assume that in all slices of a 3D image, the bounding box around the target object is determined by four coordinates. DeepCut [18] presents a patch-based classification network to segment the fetal brain and lung from MR images. It produces initial pseudo masks using the GrabCut segmentation algorithm given bounding-box labels. Then, a dense CRF as post-processing is utilized to refine the segmentation output. The recent method proposed in [19] attempts to segment renal tumors from CT angiography images. It proposes a 3-stage framework consisting of generating initial pseudo masks by convolutional CRFs, training several segmentation networks using initial pseudo masks, and finally, generating synthesized ground-truths by combining predictions of these networks.

Since segmenting a 2D image is much more convenient than a volumetric image, some works focus on supervision in which a fraction of slices of 3D images is annotated [24], [25]. The 3D U-Net [25] learns to provide dense volumetric segmentation from sparsely annotated 3D images where just some slices are segmented. The 3D U-Net architecture is similar to the 2D U-Net [26] but all 2D convolutions are changed into 3D ones. To train on a few numbers of manually annotated slices, 3D U-Net utilizes the weighted softmax loss function wherein only labeled voxels contribute. Recently, [24] has showed that 30% to 35% fully annotated consecutive slices for a 3D segmentation network training are required to achieve comparable results to those of fully supervised training. It presents an iterative framework based on attention-guided active learning and attempts to select the most informative slices for training. This challenge inspired us to suggest and work on the sparse supervision in which the ground truth for only one slice of a 3D image is available aiming to alleviate the burden of more manual annotation. To the best of our knowledge, it is less time-consuming for experts to segment one slice than to provide a 3D segmentation mask or even determine scribble or bounding box annotations in all slices of a 3D image [16], [19]. To overcome this challenge, we propose a novel self-training framework that follows an iterative process to propagate labels from the labeled slice to unlabeled ones gradually. The training process iteratively performs two steps: 1) generating pseudo labels for unlabeled voxels incrementally by applying both the 3D U-Net to benefit from volumetric information and a 2D auxiliary network to exploit label propagation by inter-slice registration, 2) updating the 3D segmentation network using both initially annotated voxels by experts and automatically annotated ones. Our proposed model has the potential to be trained using only one slice which is randomly selected and annotated by an expert before training and also requires no export during the training process to strengthen supervision as opposed to [24]. Moreover, our evaluation shows possessing the segmentation mask of one slice in each 3D image seems more informative than the other weak annotation scenarios like scribbles and bounding boxes.

II. Method

A. Problem Formulation

We purpose to develop a method for 3D image segmentation in which a sparsely annotated training set is available, i.e. only one slice of each 3D image is annotated by experts. Let $X = \{X^{(1)}, X^{(2)}, \ldots, X^{(N)}\}$ denote the training 3D images consisting of N volumetric images, where each $X^{(n)}$ contains D slices (i.e. 2D images), $X^{(n)} = (X_1^{(n)}, X_2^{(n)}, \ldots, X_D^{(n)})$ in which $X_d^{(n)} \in \mathbb{R}^{H \times W}$ (i.e. H and W show the height and width of slices respectively). In the introduced novel scenario, we assume that for each $X^{(n)}$ only one of its slices, s_n, has been annotated by experts. We call this slice the source slice. Thus, the set of available ground truth masks for the training set is $Y = \{Y_{s_1}^{(1)}, Y_{s_2}^{(2)}, \ldots, Y_{s_N}^{(N)}\}$, where $Y_{s_n}^{(n)} \in \mathbb{R}^{(H \times W)}$ has the same size as that of its corresponding source slice, $X_{s_n}^{(n)}$. The objective is to learn a 3D segmentation model $M_{3DSeg}(.)$ from the partially annotated training set.

B. Overview of the Proposed Model

In this work, we suggest a novel 3D image segmentation algorithm called 3D-SegReg that learns to produce dense segmentation masks given training samples (with sparse annotation). To this end, we propose an iterative framework based on a self-training model aiming to propagate labels from the annotated slice to unannotated ones gradually. The training process of our proposed method, summarized in Algorithm 1, alternates between the two steps: (1) predicting the segmentation label of unannotated voxels by aggregating the results of the 3D segmentation network trained so far (the block (a) of Fig. 1) and a 2D registration network (the block (b) of Fig. 1) that transforms the available segmentation of a slice to that of its neighboring slices (via inter-slice registration). (2) training and updating the 3D segmentation network by incorporating both labeled data and unlabeled data with their pseudo-labels predicted in the first step. Indeed, our purpose is to show that applying internal

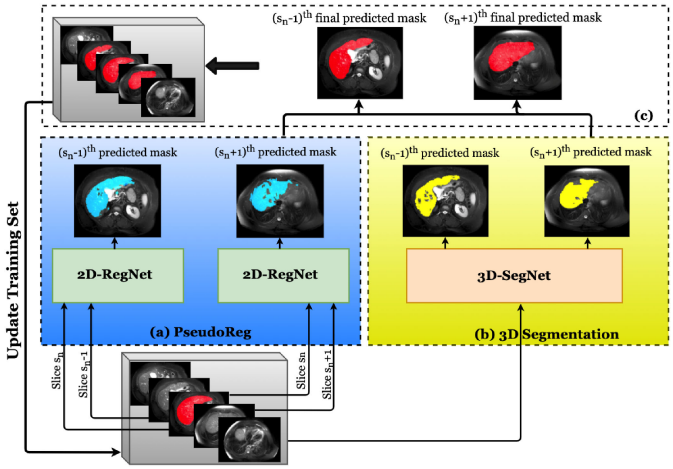

Fig. 1. A graphical presentation of the first iteration of the proposed 3D-SegReg method for dense volumetric segmentation. The segmentation mask with red color is ground truth. The segmentation masks in the (a) block with blue, (b) with yellow, and (c) with green colors are generated by PseudoReg, 3D-SegNet, and a combination of these two, respectively. Dataset is updated according to the pseudo-labels generated for further iterations.

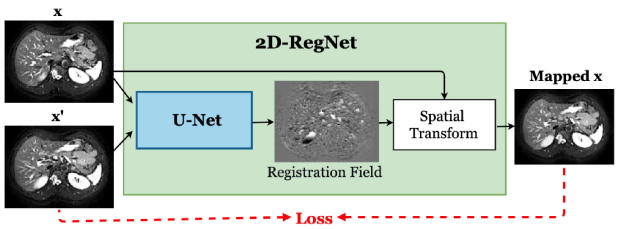

Fig. 2. An overview of the 2D-RegNet network [4]. It learns the spatial transformation that aligns consecutive slices.

registration of consecutive 2D slices along with 3D processing of volumes by a 3D segmentation network can result in a competent performance even when only one annotated slice is available. In the ensuing subsections, we first introduce the components of the proposed model, and then the training process, including the two principal steps mentioned above, will be aptly elaborated upon.

C. PseudoReg

To gradually propagate the labels in the source slice to other slices, we develop a novel auxiliary model, called PseudoReg, to produce the segmentation mask of an unlabeled slice by adapting the segmentation mask of its neighboring slice that is closer to the source slice s_n. We start with generating pseudo-labels for neighboring slices of the source slice and gradually generate pseudo-labels for more distant slices. Thus, in the i-th iteration, pseudo labels for the voxels of the i-th closest slices to the source slice are predicted from the pseudo labels that have been predicted in the previous iteration for voxels of the $(i-1)$-th closest slice to s_n (Step 3 of Algorithm 2). Indeed, it finds a spatial transformation between these two consecutive slices and employs it to adapt the segmentation mask of the $(i-1)$-th closest slice to s_n and achieve the segmentation mask of the i-th closest slice accordingly (PseudoReg box in the left bottom of Fig. 1 shows this process for the first iteration). To this end, we introduce the 2D-RegNet, which is based on a 2D deformable registration and benefits from the model proposed in [4]. The 2D-RegNet architecture is based on the registration network introduced in [4], which employs the 2D U-Net [26] to find registration field. Contrary to [4], since we need local spatial variations between slices, 2D-RegNet learns only spatial transformations between 2D images without imposing any smoothness constraints on the output transformations. An overview of the 2D-RegNet model is outlined in Fig. 2 which takes a pair (x, x') of slices and finds the transformation that tries to map x into x'.

Algorithm 1: The Proposed 3D Image Segmentation Algorithm (3D-SegReg) to Tackle the Sparse Annotation by Self-Training.

Require: $X = X^{(1)}, \ldots, X^{(N)}$
Source slices: $s_1, \ldots, s_N$
Ground truths: $Y^{(1)}_{s_1}, \ldots, Y^{(N)}_{s_N}$
Ensure: The trained M_{3DSeg} model
1: **for** each sample n **do**
2: $\mathcal{L}^{(n)} \leftarrow$ slice s_n of $X^{(n)}$ along its labels $Y^{(n)}_{s_n}$
3: **end for**
4: **for** each iteration i **do**
5: $M_{3DSeg} \leftarrow train(M_{3DSeg}, \mathcal{L}^{(1)}, \ldots, \mathcal{L}^{(N)})$
6: **for** each sample n **do**
7: $\mathcal{L}^{(n)} \leftarrow UpdateLabels(i, n, s_n, \mathcal{L}^{(n)})$
8: **end for**
9: **end for**

Algorithm 2: Update Labels (by Generating Pseudo-Labels).

Require: i: iteration of self-training
n: number of the 3D images in the training data
s_n: the source slice for the n-th 3D sample
$\mathcal{L}^{(n)}$: Labeled or pseudo-labeled voxels of the n-th sample along with their segmentation labels (in the first iteration, it includes just the voxels of the slice s_n along with their labels)
$M_{2DReg} \leftarrow$ The trained 2D-RegNet model
$M_{3DSeg} \leftarrow$ The trained 3D-SegNet model until now
Ensure: Updated label set $\mathcal{L}^{(n)}$
1: Compute the predicted mask via 3D-SegNet as: $P_{3D} \leftarrow M_{3DSeg}(X^{(n)})$
2: **for** each available slice $j \in \{s_n - i, s_n + i\}$ **do**
3: Compute the predicted mask of $X^{(n)}_j$ via PseudoReg: $P_{2D}(j) \leftarrow M_{2DReg}(X^{(n)}_j, X^{(n)}_{j+\text{sign}(s_n-j)})$
4: $P(j) = \frac{P_{3D}(j)+P_{2D}(j)}{2}$
5: Voxels in the j-th slice with a confidence higher than τ according to $P(j,.,.)$ are added to the $\mathcal{L}^{(n)}$ set with their probabilistic predicted label.
6: **end for**

2D-RegNet is trained on all pairs of consecutive slices of 3D volumes by employing an unsupervised loss that assesses the intensity variations between the warped version of the first input image and the second input image (via the spatial transformation). After training of 2D-RegNet, given a pair (x, x') of slices, it can compute the spatial transformation that maps x to its neighboring slice x'. Thus, the segmentation mask of x' can be estimated by applying this spatial transformation to the segmentation mask of x if it is available.

D. 3D Segmentation Model

We utilize a 3D segmentation network, called 3D-SegNet, to produce the 3D image segmentation masks. 3D U-Net architecture [25] is employed as our 3D segmentation network. However, in this work we do not focus on segmentation network architecture, since our method can be applied on any 3D segmentation network. In each iteration, 3D-SegNet is trained (Step 5 of Algorithm 1) by utilizing both the labeled data and the unlabeled data with their probabilistic pseudo labels in its loss function. To train 3D-SegNet with sparse annotations, we introduce a supervised loss function wherein labeled voxels, and pseudo-labeled voxels (with their probabilistic target) contribute. To overcome the class imbalance problem and alleviate the propagation error due to the pseudo labels, we propose the weighted cross-entropy loss as:

$$L = -\sum_{n=1}^{N} \sum_{(v,p)\in\mathcal{L}^{(n)}} \sum_{c=1}^{C} \pi_c r_v (p_c \log(\hat{p}_c) + (1-p_c)\log(1-\hat{p}_c)), \quad (1)$$

where π_c illustrates the class balancing weight of class c that is set to the inverse of the number of samples in class c. $\mathcal{L}^{(n)}$ shows labeled or pseudo-labeled voxels of the n-th 3D sample. $v = (j, k, l)$ indicates the index of a labeled voxel in $\mathcal{L}^{(n)}$ where j denotes the slice number and k, l show the location of the voxel in the j-th slice. p denotes the target probability vector for this voxel. The probability vector p for source slices (annotated by experts) are one-hot. For pseudo-labeled voxel v, p denotes the predicted probability vector obtained from the combination of 3D image segmentation and PseudoReg outputs on the voxel v as pseudo-label. $\hat{p}$ shows the output of 3D-SegNet on the voxel v (i.e. voxel (j, k, l) of the predicted label volume $M_{3DSeg}(X^{(n)})$). Using probability maps instead of label maps as the target for pseudo-labeled voxels enables us to consider the confidence of the pseudo labels assigned to voxels in the segmentation loss. Besides, to alleviate the propagation error due to pseudo labels, we consider a weight for voxels of each slice in (1). This weight plays a forgetting mechanism in which the forgetting rate for the voxel v is defined as $r_v = \gamma^{|j-s_n|}$, where s_n denotes the source slice of the n-th 3D sample and γ is a hyper-parameter. In fact, this coefficient is decreased exponentially with the distance of the slice in which v is located from the source slice s_n. If $\gamma = 1$, all the pseudo labels are equally contributed to the loss function, and for $\gamma < 1$, a lower weight is assigned to slices that are more distant from the source slice.

E. Training

Algorithm 1 shows the training process of our method. In the i-th self-training iteration, first, the 3D-SegNet using the last set of labeled voxels is trained and then for each sample $X^{(n)}$, we prepare more pseudo-labeled voxels by Algorithm 2 and update the label set of the n-th sample $\mathcal{L}^{(n)}$. In fact, the segmentation masks of $s_n - i$ or $s_n + i$ slices of $X^{(n)}$ situated in distance i from the source slice s_n are found and the voxels with more confident labels are added to $\mathcal{L}^{(n)}$. Indeed, the results of the 3D segmentation network obtained in the previous iteration and those of PseudoReg network are combined to achieve more accurate pseudo labels (Step 4 of Algorithm 2). To alleviate the propagating errors occurring in the pseudo labels, we select the reliable predictions to update the 3D segmentation network more accurately. Indeed, according to Step 5 of Algorithm 2, only those voxels for which we predict a label with a probability higher than a threshold are added to the $\mathcal{L}^{(n)}$ set (along with their predicted probabilistic label). Eventually, the 3D-SegNet network using both real and pseudo labels will be updated by minimizing its cost function discussed in (1). More details of the proposed segmentation loss have been described in Section II-D. Eventually, after training 3D-SegNet by the proposed method on both labeled and pseudo-labeled voxels during iterations, it is ready to segment an unlabeled 3D image in the testing phase.

III. Experimental Results

In this section, our framework is assessed on the challenging task of specific organ segmentation. In the ensuing, we give empirical supports and conduct several experiments on the 3D medical image segmentation task to evaluate our method.

Datasets: We perform experiments on two different public benchmarks to involve several anatomical structures and different image modalities: the CHAOS [27] and Visceral Anatomy 3 [29]. We follow Task 5 of the CHAOS challenge to segment MRI scans of three abdominal organs including the liver, spleen, and kidneys. It contains 40 scans which are divided into 20 sets for training and 20 sets for testing. The Visceral Anatomy 3 benchmark presents the whole-body segmentation task from contrast-enhanced CT scans. Similar to [28], to address the task of segmenting 10 abdominal organs (i.e. liver, spleen, left kidney, right kidney, left lung, right lung, left psoas, right psoas, aorta, and trachea), we use 65 scans (silver corpus) with noisy labels as training and 20 scans (gold corpus) with exact labels as testing sets, just as in [28]. Although in both benchmarks all volumetric training samples are fully labeled, we use only the segmentation mask of the source slices as the weak supervision during the training phase.

Evaluation Measures: Since ground truths of testing data of CHAOS benchmark are not accessible, the assessment of the results on the testing data is performed by the challenge website. Accordingly, by following the assessment criteria of CHAOS challenge, we report results on 20 testing data based on three different metrics: DICE, relative absolute volume difference (RAVD), and average symmetric surface distance (ASSD) to analyze results in terms of overlapping, volumetric, and spatial

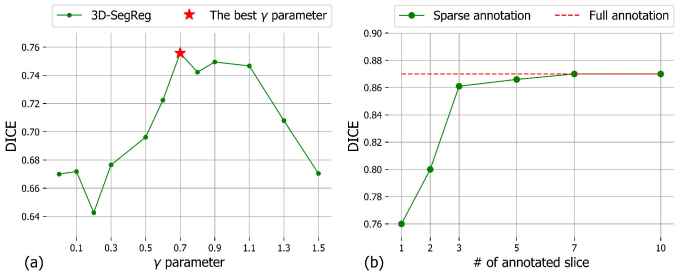

Fig. 3. The impact of the γ hyper-parameter (a), and increasing the number of annotated slices (b) on the liver segmentation task of CHAOS challenge.

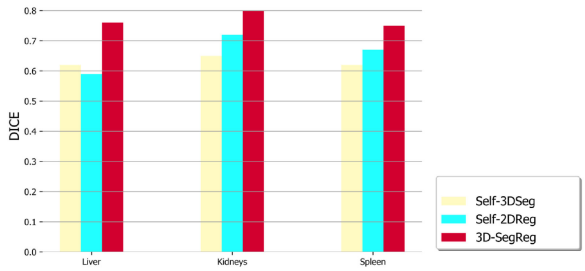

Fig. 4. DICE results of the proposed method, 3D-SegReg, compared to two ablated versions, i.e. the Self-3DSeg and Self-2DReg, on three organs: liver, kidneys, and spleen of the CHAOS challenge.

variations. Besides, we evaluate the segmentation results on gold corpus of Visceral benchmark just with dice coefficient (DICE) as considered in this challenge. Therefore, evaluation metrics can be defined as follows:

$$DICE(A,B) = 2\frac{|A \cap B|}{|A| + |B|}, \tag{2}$$

$$RAVD(A,B) = \frac{abs(|A| - |B|)}{|A|} \times 100, \tag{3}$$

$$ASSD = \frac{1}{|A|+|B|}\left(\sum_{a\in A}\min_{b\in B}\|a-b\| + \sum_{b\in B}\min_{a\in A}\|b-a\|\right), \tag{4}$$

where A and B are the set of voxels in the ground truth and the segmentation result, respectively. Also, abs, $|.|$, and $||.||$ denote the absolute value, cardinality, and $L2$ norm, respectively.

Hyper-parameters: For training network on CHAOS dataset, we use Adam optimizer with the learning rate of 0.001 for 50 epochs in each iteration with a batch size of 1 due to the limitation of GPU memory. The number of iterations in Algorithm 1 is set such that the pseudo labels are propagated and reach the most distant slice from the source slice (i.e. the number of iterations is set to $max(s_n - 1, m_n - s_n)$ where m_n denotes the number of slices in the n-th sample). The γ hyper-parameter is also set to 0.7 via a grid-search technique on the validation set of CHAOS benchmark. To investigate the impact of the γ hyper-parameter, we have reported the DICE metric for different values in the left side of Fig. 3 on the liver segmentation task of CHAOS challenge. It can be inferred $\gamma = 0.7$ achieves the highest performance. However, for $\gamma > 0.7$ and $\gamma < 0.7$, the performance reduces due to accumulating errors of pseudo labels and preventing the label propagation, respectively. Furthermore, in order to demonstrate the impact of the number of annotated slices per volume on our proposed framework, we delineate the right side of Fig. 3 based on the liver segmentation task of CHAOS challenge. According to Fig. 3(b), the performance of our proposed method is quickly improved by increasing the number of annotated slices. However, in order to consider the least supervision, the number of annotated slices of each volume is set to one in the following subsections. Note that for Visceral dataset we use the same set of parameters as used for CHAOS dataset rather than searching for their optimal values.

Ablation Studies: To investigate the performance of blocks (a) and (b) in Fig. 1, we assess our model against the absence of each block. Therefore, we suggest two ablated versions of our main model. The first version, called Self-3DSeg, utilizes self-training but doesn't employ the PseudoReg block in Fig. 1. In fact, Step 4 of Algorithm 2 is replaced with $P(j) = P_{3D}(j)$ in this version. Thus, Self-3DSeg does not use inter-slice registration and just employs the 3D segmentation information for the organ segmentation task. On the other hand, the second version called Self-2DReg covers only the PseudoReg block in Fig. 1; thus, just 2D information is utilized to segment the target organ. A quantitative evaluation of these methods on the abdominal organ segmentation task of CHAOS challenge is represented in Fig. 4. It reveals that the 3D-SegReg improves the result of the Self-3DSeg and achieves 0.14, 0.15, and 0.13 higher DICE on the liver, kidneys, and spleen organs, respectively. Similarly, it obtains 0.17, 0.08, and 0.08 higher DICE respectively on the liver, kidneys, and spleen organs compared to the Self-2DReg. Thus, it can be inferred that after aggregating the information of the Self-2DReg and Self-3DSeg, the proposed 3D-SegReg takes noticeably higher DICE on all organs compared with these stand-alone methods. As a result, it can be confirmed that the incorporation of 2D inter-slice registration information in the self-training process of the proposed method (Step 4 of Algorithm 2) helps to provide more proper pseudo-labels for the training of the 3D segmentation network and thus makes the model more accurate.

Fig. 5 illustrates the result of mask prediction on the liver segmentation task of the CHAOS challenge, where these three methods are applied. Ground truths for the liver organ are indicated in the first row. The second to fourth rows represent the segmentation results of the 3D-SegReg with the green color, Self-3DSeg with the yellow color, and Self-2DReg with the blue color, respectively. The mask with red color is the source slice annotated by experts. It can be observed that incorporating both the Self-3DReg and Self-2DReg models which results in the 3D-SegReg model, can significantly produce more accurate segmentation masks than these stand-alone methods. Although there are some discrepancies between the predicted masks by the 3D-SegReg method and the real segmentation masks outlined in the first rows, the segmentation mask of slices is generally well predicted by 3D-SegReg. It is worth to mention that only one annotated slice in the training stage is available. Since slices of a 3D image gradually change, we can predict the segmentation

TABLE I
COMPARISON OF DIFFERENT METHODS ON CHAOS DATASET (THE BEST RESULTS ARE HIGHLIGHTED IN BOLD). SYMBOLS INDICATE STATISTICAL SIGNIFICANCE WITH $p < 0.01$ ($>$: SIGNIFICANTLY BETTER THAN OURS, $<$: SIGNIFICANTLY WORSE THAN OURS, $\sim$: NO SIGNIFICANT DIFFERENCE, N: STATISTICAL ANALYSIS IS NOT AVAILABLE THAT IS USED FOR RESULTS OF [19] SINCE ITS INDIVIDUAL RESULTS ARE NOT AVAILABLE FOR THE TEST)

	Liver			Kidneys			Spleen			Average		
	DICE	RAVD	ASSD	DICE	RAVD	ASSD	DICE	RAVD	ASSD	DICE	RAVD	ASSD
3D U-Net (Fully automated) [25]	$0.69^{<}$	25.30	34.91	$0.60^{<}$	39.81	59.49	$0.69^{\sim}$	42.56	21.37	0.66	35.89	38.59
3D atlas random augmentation [10]	$0.58^{<}$	23.30	16.20	$0.32^{<}$	38.89	21.90	$0.26^{<}$	32.26	22.70	0.39	31.48	20.26
3D CNN scribble-level [16]	$0.72^{\sim}$	25.82	14.75	$0.67^{<}$	27.87	15.66	$0.61^{\sim}$	25.54	25.54	0.67	26.41	17.64
3D CNN box-level [19]	0.71^{N}	24.1	10.01	0.75^{N}	17.67	8.9	0.69^{N}	27.8	21.12	0.72	20.29	13.07
3D SegReg (ours)	**0.76**	**22.60**	**10.38**	**0.82**	**16.68**	**4.95**	**0.79**	**20.93**	**19.33**	**0.79**	**20.07**	**11.55**
3D U-Net (Fully supervised) [25]	$0.87^{>}$	10.48	4.45	$0.84^{\sim}$	35.07	16.12	$0.82^{\sim}$	31.86	14.41	0.84	25.80	11.66

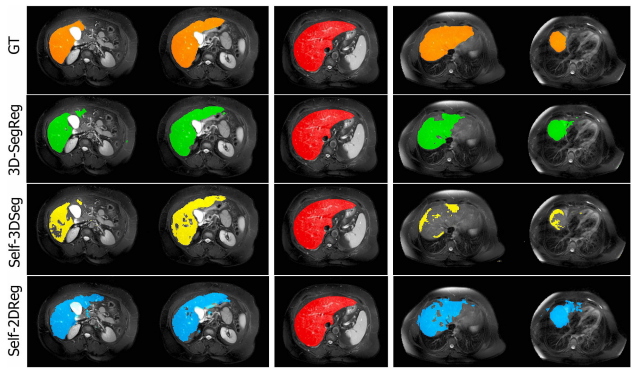

Fig. 5. Visual examples of liver segmentation task results of the CHAOS challenge. The first row exhibits ground truth (GT). The second to fourth rows are segmentation results of the 3D-SegReg with the green, Self-3DSeg with the yellow, and Self-2DReg with the blue color, respectively. The source slice annotated by experts is shown with the red color.

mask of all slices by following gradual variations (as accomplished in the proposed method). However, the performance of 3D-SegReg can be improved when more annotated slices are used in the training stage as illustrated in Fig. 3(b).

Compared Methods and Results: To investigate the performance of our method, we compare 3D-SegReg against existing approaches which attempt to leverage different types of supervisions, including: 1) 3D U-Net [25] in its fully-automated setup, that works exactly in the setting of the proposed method (which employs the annotation of only one slice of each 3D training image). 2) 3D Atlas Data Augmentation [10], which trains its network applying only one fully annotated 3D image. 3) 3D CNN box-level [19], which leverages bounding-box supervision in which the box around the target object is delineated in all slices of 3D training images. 4) 3D CNN scribble-level [16], which tackles scribble labels where the location of the target object in all slices of a 3D image is specified by a small ratio of the object's voxels. Furthermore, 3D U-Net in its fully-supervised setup, which assumes the network is trained on a completely annotated training set, is considered to produce upper bounds for results. To achieve a fair comparison, we design our 3D segmentation network architecture similar to 3D U-Net and find all parameters and the randomly selected source slices (utilized in our method and fully-automated 3D U-Net) the same. To report results of [10], we opt the best 3D sample of the training set with its full 3D segmentation mask as the source segmented volume. To provide scribble annotations for 3D CNN scribble-level method, we consider the random point scribble along with individual bounds strategy, both introduced in [16] on all 2D slices as it outperforms other strategies. Also, according to [19], the margin of the bounding box is set 5 pixels in all 2D slices for this method.

Table I and Table II present a comparison between these methods on CHAOS and Visceral datasets, respectively, on different abdominal organs. The average of these specific organs is also reported in the last column. In addition, the statistical significance tests are made on DICE values of test samples using Wilcoxon signed rank test [30] with a significance threshold of $p < 0.01$. It can be inferred from Table I that our method outperforms other related methods (according to DICE, RAVD, and ASSD measures) on all organs of CHAOS dataset with a large margin. Moreover, the proposed method generally shows close results to those of the fully supervised 3D U-Net while utilizing only a very small portion of annotated slices. The best RAVD scores have been obtained for 3D-SegReg (even compared to fully supervised 3D U-Net). It indicates that our method can preserve the volume of organs better. Besides, according to the average ASSD, our method can handle boundaries and spatial variations successfully. In fact, due to employing inter-slice registration during training, our method learns to tackle local spatial variations more accurately. On the kidneys, it even outperforms fully supervised 3D U-Net according to ASSD. Fig. 6 depicts the segmentation boundaries obtained by our method (with the blue color) and the fully-supervised U-Net (with the yellow color) on the kidneys. The arrows indicate voxels that are incorrectly predicted and have a long distance to the nearest point of ground truth. As it is clear from Fig. 6, there is much diffusion in the generated segmentation masks by fully-supervised U-Net.

The segmentation results based on DICE score on 10 abdominal organs of Visceral dataset are reported in Table II. Note that hyper-parameters for this dataset are set to the same values as the selected ones for CHAOS dataset, rather than selecting their optimal values, which could lead to better results. In general, we can observe that the proposed 3D-SegReg exceeds other methods and achieves closer performance to fully-supervised U-Net on average. When comparing the performance of 3D-SegReg with two compared weakly-supervised methods, 3D-CNN-scribble-level and 3D-CNN-box-level methods, it can be seen that despite 3D-SegReg uses supervision on only one slice (while the compared methods use supervision on all slices), it

TABLE II

COMPARISON OF DIFFERENT METHODS ON VISCERAL DATASET (THE BEST RESULTS ARE HIGHLIGHTED IN BOLD). SYMBOLS INDICATE STATISTICAL SIGNIFICANCE WITH $p < 0.01$ ($>$: SIGNIFICANTLY BETTER THAN OURS, $<$: SIGNIFICANTLY WORSE THAN OURS, $\sim$: NO SIGNIFICANT DIFFERENCE, N: STATISTICAL ANALYSIS IS NOT AVAILABLE THAT IS USED FOR RESULTS OF [19] SINCE ITS INDIVIDUAL RESULTS ARE NOT AVAILABLE FOR THE TEST)

	Liver	Spleen	Aorta	Trachea	L.Lung	R.Lung	L.Kidney	R.Kidney	L.Psoas	R.Psoas	**Avg.**
3D U-Net (Fully automated) [25]	$0.78^{\sim}$	$0.65^{<}$	$0.62^{\sim}$	$0.84^{\sim}$	$0.91^{\sim}$	$0.91^{\sim}$	$0.77^{\sim}$	$0.84^{<}$	$0.71^{<}$	$\mathbf{0.75}^{\sim}$	0.77
3D atlas random augmentation [10]	$0.60^{<}$	$0.38^{<}$	$0.29^{<}$	$0.15^{<}$	$0.63^{<}$	$0.78^{<}$	$0.40^{<}$	$0.38^{<}$	$0.41^{<}$	$0.48^{<}$	0.45
3D CNN scribble-level [16]	$\mathbf{0.83}^{\sim}$	$0.67^{\sim}$	$\mathbf{0.75}^{>}$	$0.83^{\sim}$	$0.91^{<}$	$0.91^{<}$	$0.65^{<}$	$0.76^{<}$	$0.73^{\sim}$	$0.69^{\sim}$	0.77
3D CNN box-level [19]	0.80^{N}	0.68^{N}	0.66^{N}	$\mathbf{0.84}^{N}$	0.86^{N}	0.84^{N}	0.74^{N}	0.80^{N}	0.74^{N}	0.71^{N}	0.76
3D SegReg (ours)	0.80	**0.75**	0.62	0.82	**0.92**	**0.92**	**0.78**	**0.88**	**0.78**	0.74	**0.80**
3D U-Net (Fully supervised) [25]	$0.84^{>}$	$0.84^{>}$	$0.81^{>}$	$0.82^{\sim}$	$0.93^{\sim}$	$0.92^{\sim}$	$0.83^{\sim}$	$0.84^{\sim}$	$0.80^{>}$	$0.79^{>}$	0.84

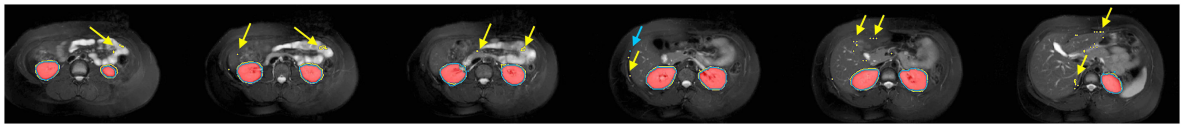

Fig. 6. Visual comparison of the segmentation mask borders generated by our method (with the blue contour) and fully-supervised U-Net (with the yellow contour) on the kidney organ of CHAOS dataset. The yellow arrows indicate diffusion caused by fully-supervised U-Net and the blue caused by our method. The grand truths are shown with the red color.

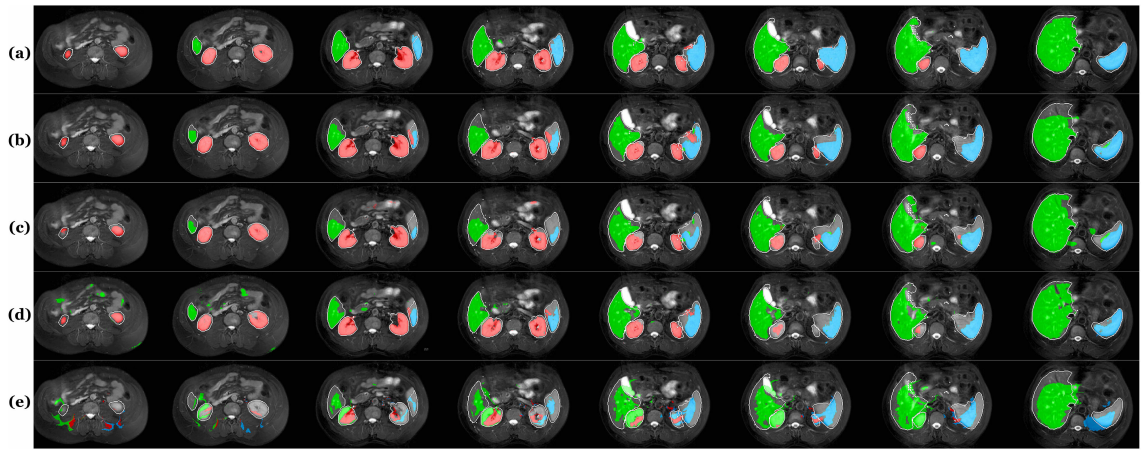

Fig. 7. Visual examples of a test sample of CHAOS benchmark on different organs. The liver is indicated in green, the spleen in blue, and kidneys (left and right) in red. White contours represent the mask-border generated by fully-supervised U-Net. The first to fifth rows represent masks predicted by (a) our 3D-SegReg method, (b) 3D CNN box-level, (c) 3D CNN scribble-level, (d) fully-automated 3D U-Net, and (e) 3D atlas random augmentation, respectively.

generally outperforms them. However, on the most deformed organ, aorta, it underperforms 3D-CNN-scribble-level and 3D-CNN-box-level methods since our model can not handle special organs like aorta that change from one-part to multi-part across slices.

A qualitative evaluation of the compared methods is shown in Fig. 7, which illustrates the segmentation results of different abdominal organs on a test 3D image of CHAOS benchmark. Generally, it can be observed that the proposed 3D-SegReg remarkably improves the segmentation results on all organs compared to other methods presented in the sparse annotation scenario. The superiority of our work can be attributed to two factors: (1) utilizing pseudo labels in the training process aids our network to remarkably strengthen itself, (2) applied 2D registration network focusing on organ-inherent pattern detection conducts the 3D segmentation network to yield more accurate segmentation. Indeed, our proposed method benefits from the inter-slice registration in 2D-level (gradually registration of a slice with the annotated slice) for 3D segmentation that has not been employed in the related work so far to the best of our knowledge.

IV. DISCUSSION AND CONCLUSION

In this work, we proposed a novel self-training model for 3D image segmentation with sparse annotation wherein the ground truth on only one slice of training volume images was achievable. In this situation, to train a better 3D image segmentation model, we propagated the label of the annotated slice to its neighbors via inter-slice registration, and the unlabeled voxels with their prepared pseudo labels are gradually incorporated in the learning process. The proposed iterative method to produce pseudo labels in each iteration do not use only the results of 3D-SegNet in the previous iteration but also developed the PseudoReg network, which was able to gradually propagate labels between 2D slices to lead the 3D-SegNet network more accurate. Although time

complexity of our iterative framework during training phase is about twice the time of fully-supervised and fully-annotated 3D U-Net, our testing time cost is equal to that of a simple 3D U-Net. However, in comparison with other related methods such as [16] and [19], it has approximately equal training time.

Experimental results exhibited that our method generally outperforms other related methods introduced for sparse annotations. According to Table I and Table II, [10] underperforms other methods. Indeed, [10] relies strongly on the selected 3D atlas, and contrary to our method it can not handle test samples that are partly different in size, shape, or appearance with the atlas. Although, both compared weakly-supervised methods, 3D CNN scribble-level [16], and 3D CNN box-level [19], improve the segmentation performance in sparse annotation problem, they are less practical in real clinical scenarios. Indeed, one of the limitation of these methods is that since they need the supervision (i.e. bounding box and scribble) in all slices of a 3D image, more time-consuming annotation process is needed. However, our framework requires only one annotated slice and therefore has a strong potential to be practical in clinical. Besides, it can be inferred from Table I and Table II that our method can improve the performance of segmentation task on sparse annotations with a large margin against other related methods. Moreover, statistical comparison between 3D-SegReg with other methods shows that in many cases our method achieves significantly better results. Also in many cases, there is no significant difference between our method and fully supervised method.

Limitation: Despite our method's superior performance, its limitation should also be noted. Our experiments show that in order to obtain good results, it is important to see a variety of source slices (along the axial axis) for each organ in different 3D samples altogether (i.e., having various source slices across the organ's visible range is more suitable than annotating a fixed one in all 3D samples). Therefore, in our experiments, for each 3D sample, we have selected the source slice randomly from the ones in which the target organ is visible.

REFERENCES

[1] A. bitarafan, A. Amini, M. Soleymani, and H. Khodajou-Chokami, "A hybrid deep model for automatic arrhythmia classification based on LSTM recurrent networks," in *Proc. IEEE Int. Symp. Med. Meas. Appl.*, Jun. 2020, pp. 1–6.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[3] G. Yang *et al.*, "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1310–1321, Jun. 2018.

[4] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "An unsupervised learning model for deformable medical image registration," in *Pro. IEEE Conf. Comput. Vis. Pattern Recognit*. Salt Lake City, UT, USA, 2018, pp. 9252–9260.

[5] J. Hofmanninger, F. Prayer, J. Pan, S. Rohrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem," 2020, *arXiv:2001.11767*.

[6] L. Zhang, V. Gopalakrishnan, L. Lu, R. M. Summers, J. Moss, and J. Yao, "Self-learning to detect and segment cysts in lung CT images without manual annotation," in *Proc. IEEE 15th Int. Symp. Biomed. Imag.* (ISBI 2018). Washington, DC, USA, 2018, pp. 1100–1103.

[7] Y. B. Tang, S. Oh, Y. X. Tang, J. Xiao, and R. M. Summers, "CT-realistic data augmentation using generative adversarial network for robust lymph node segmentation," in *Proc. Medical Imaging: CAD*, vol. 10950, 2019.

[8] S. Tang, K. Yang, Y. Chen, and L. Xiang, "X-ray-induced acoustic computed tomography for 3D breast imaging: A simulation study," *Med. Phys.*, vol. 45, no. 4, pp. 1662–1672, 2018.

[9] N. Tong, S. Gou, S. Yang, D. Ruan, and K. Sheng, "Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks," *Med. Phys.*, vol. 45, no. 10, pp. 4558–4567, 2018.

[10] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8543–8553.

[11] M. Ahmad *et al.*, "Deep belief network modeling for automatic liver segmentation," *IEEE Access*, vol. 7, pp. 20585–20595, 2019.

[12] W. Bai *et al.*, "Recurrent neural networks for aortic image sequence segmentation with sparse annotations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2018, pp. 586–594.

[13] W. Bai *et al.*, "Semi-supervised learning for network-based cardiac MR image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2017, pp. 253–260.

[14] Y. Zhou, Y. Wang, P. Tang, W. Shen, E. K. Fishman, and A. L. Yuille, "Semi-supervised multi-organ segmentation via multi-planar co-training," 2018, *arXiv:1804.02586*.

[15] S. Min, X. Chen, Z. J. Zha, F. Wu, and Y. Zhang, "A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 4578–4585, 2019.

[16] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. B. Ayed, "Constrained-CNN losses for weakly supervised segmentation," *Med. Image Anal.*, vol. 54, pp. 88–99, 2019.

[17] J. Cai *et al.*, "Accurate weakly supervised deep lesion segmentation on CT scans: Self-paced 3D mask generation from RECIST," 2018, *arXiv:1801.08614*.

[18] M. Rajchl *et al.*, "DeepCut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 674–683, Feb. 2017.

[19] G. Yang *et al.*, "Weakly-supervised convolutional neural networks of renal tumor segmentation in abdominal CTA images," *BMC Med. Imag.*, vol. 20, no. 1, pp. 1–12, 2020.

[20] O. Z. Kraus, J. L. Ba, and B. J. Frey, "Classifying and segmenting microscopy images with deep multiple instance learning," *Bioinformatics*, vol. 32, no. 12, pp. i52–i59, 2016.

[21] L. Zhou, Y. Zhao, J. Yang, Q. Yu, and X. Xu, "Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images," *IET Image Proc.*, vol. 12, no. 4, pp. 563–571, 2017.

[22] X. Feng, J. Yang, A. F. Laine, and E. D. Angelini, "Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2017, pp. 568–576.

[23] H. Roth *et al.*, "Weakly supervised segmentation from extreme points," in *Proc. Large-Scale Annotation Biomed. Data Expert Label Synth. Hardware Aware Learn. Med. Imag. Comput. Assist. Interv.*, Springer, 2019, pp. 42–50.

[24] Z. Zhang, J. Li, Z. Zhong, Z. Jiao, and X. Gao, "A sparse annotation strategy based on attention-guided active learning for 3D medical image segmentation," 2019, *arXiv:1906.07367*.

[25] 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. Interv.*, Springer, 2016, pp. 424–432.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, 2015, pp. 234–241.

[27] A. E. Kavur *et al.*, "CHAOS challenge - combined (CT-MR) healthy abdominal organ segmentation," 2020, *arXiv:2001.06535*.

[28] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel'squeeze and excitation' blocks," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 540–549, Feb. 2019.

[29] O. Jimenez-del-Toro *et al.*, "Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks," *IEEE Trans. Med. Imag.*, vol. 35, no. 11, pp. 2459–2475, Nov. 2016.

[30] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.