Sli2Vol+: Segmenting 3D Medical Images Based on an Object Estimation Guided Correspondence Flow Network*

Delin An*
University of Notre Dame
dan3@nd.edu

Pengfei Gu*
The University of Texas Rio Grande Valley
pengfei.gu01@utrgv.edu

Milan Sonka
University of Iowa
milan-sonka@uiowa.edu

Chaoli Wang
University of Notre Dame
chaoli.wang@nd.edu

Danny Z. Chen
University of Notre Dame
dchen@nd.edu

Abstract

*Deep learning (DL) methods have shown remarkable successes in medical image segmentation, often using large amounts of annotated data for model training. However, acquiring a large number of diverse labeled 3D medical image datasets is highly difficult and expensive. Recently, mask propagation DL methods were developed to reduce the annotation burden on 3D medical images. For example, Sli2Vol [59] proposed a self-supervised framework (SSF) to learn correspondences by matching neighboring slices via slice reconstruction in the training stage; the learned correspondences were then used to propagate a labeled slice to other slices in the test stage. But, these methods are still prone to error accumulation due to the inter-slice propagation of reconstruction errors. Also, they do not handle discontinuities well, which can occur between consecutive slices in 3D images, as they emphasize exploiting object continuity. To address these challenges, in this work, we propose a new SSF, called **Sli2Vol+**, for segmenting any anatomical structures in 3D medical images using only a single annotated slice per training and testing volume. Specifically, in the training stage, we first propagate an annotated 2D slice of a training volume to the other slices, generating pseudo-labels (PLs). Then, we develop a novel Object Estimation Guided Correspondence Flow Network to learn reliable correspondences between consecutive slices and corresponding PLs in a self-supervised manner. In the test stage, such correspondences are utilized to propagate a single annotated slice to the other slices of a test volume. We demonstrate the effectiveness of our method on various medical image segmentation tasks with different datasets, showing better generalizability across different organs, modalities, and modals. Code is available at https://github.com/adlsn/Sli2VolPlus.*

1. Introduction

Image segmentation is a critical task in medical image analysis, providing anatomical structure information essential for disease diagnosis and treatment planning [6, 35]. Known deep learning (DL) methods have achieved state-of-the-art (SOTA) performance in many medical image segmentation tasks, including convolutional neural network (CNN)-based methods [13, 20, 33, 39, 69], Transformer-based methods [5, 21], and hybrid approaches [12, 14, 43, 66]. However, these DL methods require abundant labeled training data to attain satisfactory performance. Labeling large amounts of medical image data, especially for 3D images, is highly difficult and expensive as this process requires domain-specific expertise, and pixel/voxel-wise annotations can be very labor-intensive and time-consuming.

Four main methods have been proposed to address the data-annotation burden. The first type explores the potential of non-annotated data through semi-supervised [38, 63, 68] and self-supervised [48, 61, 62, 70] methods to reduce the demand for labeled data. The second type leverages the segment anything model (SAM) [25] for medical image segmentation (e.g., [11, 15, 17, 30–32, 42, 67]). However, the applicability of these methods to medical image segmentation remains limited due to the significant differences between natural images and medical images. The third type is weakly-supervised methods, such as image-level [7, 18, 26], patch-level [10, 56, 57], bounding box [24, 36, 47], scribbles-level [51], and even point-level [37, 60] labeling, using rough annotations as supervision. But these methods typically yield sub-optimal performance [64] because accurate

*⋆ indicates equal contribution. This research was supported in part by NSF grants: IIS-1955395, IIS-2101696, OAC-2104158, and IIS-2401144, NIH grants: 1R01HL177814-01, R01EB004640, 2R56EB004640-16, and R56EB004640.

delineation data are not available for model training. The fourth type annotates only "worthy" samples that help improve the final segmentation accuracy, e.g., active learning methods [9, 44, 52, 58]. However, active learning methods require medical experts to provide annotations interactively, often leading to a "human-machine disharmony" problem. To address the annotation bottleneck, some studies [64, 65] managed to avoid human-machine iterations by selecting representative samples to annotate in one shot, but still needed to annotate a considerable amount of samples.

To reduce the annotation burden, two main types of mask propagation DL methods have been developed. The first type is slice reconstruction, which propagates an annotated 2D slice through the entire 3D volume by matching correspondences between consecutive slices [54, 59]. The second type is slice registration, which propagates an annotated 2D slice throughout the 3D volume by establishing spatial transformations between consecutive slices [3,4,28,34]. For example, in Sli2Vol [59], a self-supervised method was proposed to propagate a provided labeled slice for segmentation. Specifically, it first learns correspondences from adjacent slices by solving a slice reconstruction task in the training stage. Then, in the test stage, the learned correspondences are leveraged to propagate a labeled slice to the other slices for segmentation. Despite their success, these methods still suffer from several drawbacks: (i) They are prone to error drift (i.e., error accumulation) due to the inter-slice propagation of reconstruction/registration errors. (ii) They do not handle discontinuity well (e.g., unseen objects emerging or seen objects ending), which can occur between consecutive slices in 3D images, as they focus on exploiting object continuity. Observe that, intuitively, by incorporating segmentation/pseudo-labels (PLs) to provide certain supervision for slice reconstruction in the training stage, the correspondences can be better learned and more reliable, as object discontinuity can be compensated by PLs.

In this work, we propose a new self-supervised mask propagation framework, called **Sli2Vol+**, for segmenting any anatomical structures in 3D medical images by labeling only a single slice per training and testing volume. Specifically, in the training stage, we first generate PLs for the training volumes by propagating a provided labeled slice to the other slices in each volume using Sli2Vol [59]. Then, we introduce a new Object Estimation Guided Correspondence Flow Network (OEG-CFN) to learn reliable correspondences for subsequent propagation. In the test stage, the learned correspondences are utilized to propagate a labeled slice to the other slices of a test volume. Unlike Sli2Vol [59], we design OEG-CFN to learn correspondences between both consecutive slices and corresponding PLs in a self-supervised manner. Intuitively, the included PLs guide the model to focus on the estimated objects during the correspondence learning process, effectively addressing the error drift and discontinuity issues. Extensive experiments on nine public datasets (both CT and MRI) covering ten different structures of interest (SOIs) show the effectiveness of our new method and the improved generalizability across different SOIs, modalities, and modals.

Our main contributions are three-fold: (i) We propose a new self-supervised mask propagation framework for segmenting any anatomical structures in 3D medical images using only a single annotated slice in each training and testing volume. (ii) We develop a new Object Estimation Guided Correspondence Flow Network (OEG-CFN) to learn reliable correspondences between consecutive slices and corresponding PLs in a self-supervised manner, effectively addressing the error drift and discontinuity issues. (iii) Our method achieves significant improvements on nine public datasets (both CT and MRI) over known methods, and demonstrates better generalizability across different SOIs, modalities, and modals.

2. Related Work

To alleviate the burden of manual annotation and enhance generalizability in 3D medical image segmentation, mask propagation DL methods have been proposed, including slice reconstruction-based methods [54, 59] and slice registration-based methods [3, 4, 28, 34].

Different annotation scenarios were considered using the known mask propagation DL methods. One scenario involves labeling one slice per training volume (e.g., [4]) without any labels on test volumes. However, these methods may not work well with new objects in test volumes. Another scenario involves labeling one slice per test volume (e.g., [3, 59]) without using any labels of training volumes. In particular, Sli2Vol [59] uses only a single annotated slice per test volume. But these methods do not leverage training labels and may not perform well. We combine the above two scenarios, labeling one slice per training volume and one slice per test volume. Compared to [3, 59], our **Sli2Vol+** requires an additional amount of very sparse annotations (i.e., a single annotated slice per training volume).

Incorporating information between nearby slices in volumetric data can be achieved in various ways. For example, CSA-Net [19] used pixel-level cross-slice attention to enhance the segmentation of a central slice, while CSAM [27] employed slice-level attention across feature maps at multiple scales. Both these methods followed a standard 2.5D approach by stacking neighboring slices. However, compared to existing mask propagation approaches, these methods tend to be less memory efficient.

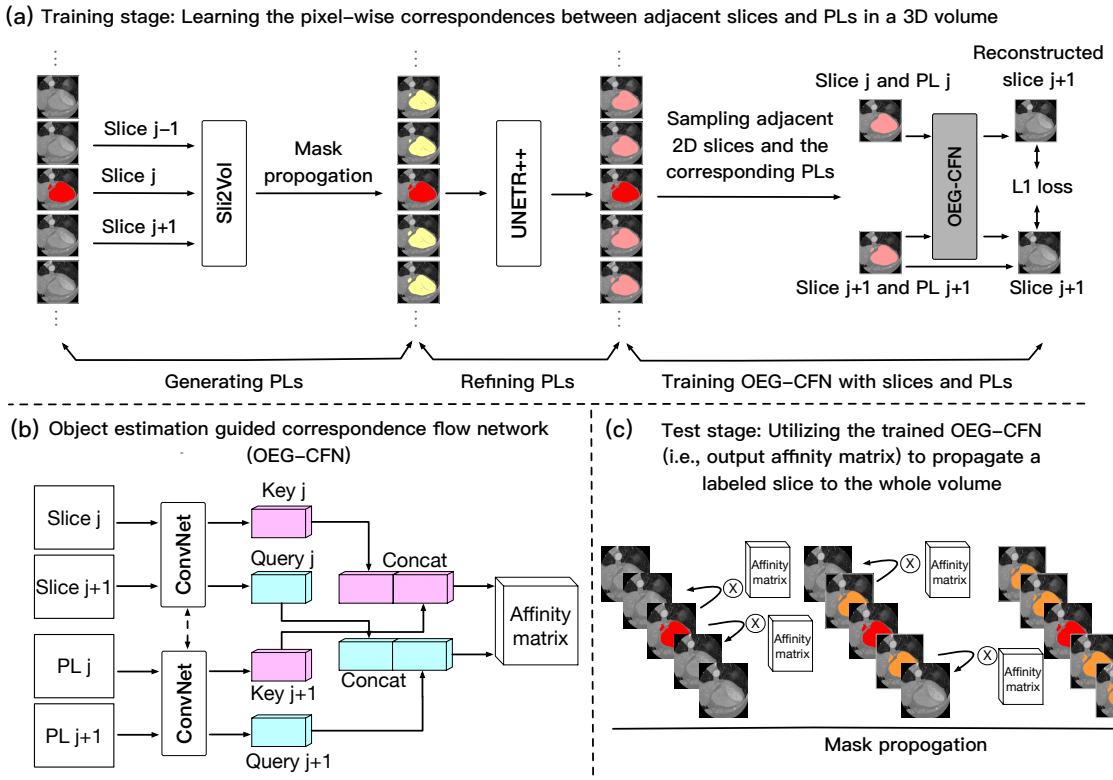

Figure 1. The pipeline of our proposed framework. (a) Training stage: Adjacent 2D slices and their corresponding generated pseudo-labels (PLs) are sampled from a 3D volume to train the Object Estimation Guided Correspondence Flow Network (OEG-CFN). (b) The architecture of OEG-CFN. (c) Test stage: The trained OEG-CFN is used to propagate a labeled slice to the other slices of the entire volume (five slices of a test volume are shown in the example). Red annotations represent ground truth segmentations, yellow and pink annotations represent PLs, and orange annotations represent the final segmentations. $\bigotimes$ denotes matrix multiplication.

3. Method

3.1. Problem Formulation and Method Overview

Problem Formulation. Let $\mathbb{X}_{train} = \{X_1, X_2, \ldots, X_N\}$ be a given set of N 3D training images, where each volume X_i contains D 2D slices, $X_i = (S_1^i, S_2^i, \ldots, S_D^i)$, and only a single annotated slice is provided for each X_i. Given a 3D testing image $X_{test} = (S'_1, S'_2, \ldots, S'_{D'})$, our goal is to segment the SOIs in the test volume with a given 2D segmentation mask of a single slice S'_j in X_{test}.

Overview of the Method. As shown in Fig. 1, our **Sli2Vol+** pipeline consists of the following main steps. In the training stage, we first generate and refine PLs for all the training volumes. Then, we learn the pixel-wise correspondences between adjacent slices and the corresponding PLs in a self-supervised manner using the proposed Object Estimation Guided Correspondence Flow Network (OEG-CFN). In the test stage, the trained OEG-CFN is used to propagate an annotated slice to the other slices in the test volume, generating segmentation of the entire volume.

3.2. Correspondence Learning with Object Estimation Guided Correspondence Flow Network

Review of Sli2Vol. Sli2Vol [59] is an interesting and closely related mask propagation DL method, which proposed a self-supervised approach for learning dense correspondences by matching neighboring slices via slice reconstruction. The learned correspondences (i.e., a set of affinity matrices) are then used for mask propagation by weighting and copying pixels between consecutive slices in the test stage. Specifically, in the training stage, pairs of adjacent slices sampled from a training volume, $\{\mathbf{S}_j, \mathbf{S}_{j+1}\}$, are fed to a CNN network, ConvNet (parameterized by $\psi(\cdot;\theta)$), to learn features of the key $\mathbf{k}_j$ and query $\mathbf{q}_{j+1}$, as:

$$[\mathbf{k}_j,\ \mathbf{q}_{j+1}] = [\psi(g(\mathbf{S}_j);\ \theta),\ \psi(g(\mathbf{S}_{j+1});\ \theta)], \quad (1)$$

where $g(\cdot)$ denotes an edge profile generator for the information bottleneck. Then, an affinity matrix $\mathbf{A}_{j\to j+1}$ is computed from $\mathbf{k}_j$ and $\mathbf{q}_{j+1}$ to represent the feature similarity between slices $\mathbf{S}_j$ and $\mathbf{S}_{j+1}$, as:

$$\mathbf{A}_{j\to j+1}(u,v) = \frac{\exp\langle \mathbf{q}_{j+1}(u,:), \mathbf{k}_j(v,:)\rangle}{\sum_{p\in\Omega}\exp\langle \mathbf{q}_{j+1}(u,:), \mathbf{k}_j(p,:)\rangle}, \quad (2)$$

Method	Dice	
	Decath-spleen	Decath-liver
TransUNet [8]	0.950 ± 0.013	0.944 ± 0.020
CoTr [55]	0.954 ± 0.018	0.942 ± 0.014
UNETR [14]	0.964 ± 0.016	0.961 ± 0.015
UNETR++ [43]	$\mathbf{0.971} \pm 0.012$	$\mathbf{0.964} \pm 0.010$

Table 1. Segmentation results of different models on the Decath-spleen and liver datasets [45]. The best results are marked in **bold**.

where $\langle \cdot, \cdot \rangle$ denotes the dot product of two vectors, and Ω is a window surrounding pixels of v for computing local attention. In the test stage, the affinity matrices thus computed are leveraged to propagate the given mask of a single slice to the other slices for the segmentation of a test volume.

Sli2Vol [59] focuses on exploiting object continuity, i.e., it is capable of propagating the labels of a slice to its neighboring slices well when there are no drastic changes between these slices. But, this assumption is often not held in 3D medical images, where discontinuities may occur between consecutive slices: Unseen objects may emerge, or seen objects may end. This gives rise to error drift and discontinuity issues, and may incur sub-optimal performance.

Our idea is that by including segmentation/PLs to provide certain supervision for the slice reconstruction, the correspondences can be better learned to handle these issues. Hence, we propose a new self-supervised approach that consists of the following steps: PLs generation, PLs refinement, and correspondence learning with an Object Estimation Guided Correspondence Flow Network (OEG-CFN).

PLs Generation. Given a training volume $X = (S_1, S_2, \ldots, S_D)$ and the annotation Y_j of a single slice S_j, we apply Sli2Vol [59] to propagate the annotation Y_j to the other slices, generating PLs for the whole volume X. However, Sli2Vol [59] neglects global 3D information within the whole volume as it determines the correspondences using only adjacent slices without considering the entire context.

PLs Refinement. To address this issue, we utilize a SOTA 3D model (i.e., UNETR++ [43]) to refine the PLs. Specifically, we use the generated PLs of all the training slices to train UNETR++ [43]. Then, we apply the trained UNETR++ to the entire training volume X to refine the PLs. By considering the 3D information of X, the quality of the PLs is improved. We experimentally choose UNETR++ [43] as our 3D model because it yields the best performance for medical image segmentation (see Table 1).

Object Estimation Guided Correspondence Flow Network (OEG-CFN). To include PLs for providing supervision for slice reconstruction, we propose a new Object Estimation Guided Correspondence Flow Network (OEG-CFN) to learn reliable correspondences. Our idea is to decompose the key and query features into two sets so that those features learned from consecutive slices handle continuity, and the features learned from PLs handle discontinuity. As shown in Fig. 1(b), OEG-CFN has two paths. The top path uses a ConvNet to learn the key and query features from consecutive slices, while the bottom path learns another two sets of key and query features using PLs with a ConvNet. Both the ConvNets share the same architecture and parameters. The key and query features learned from the slices and PLs are then concatenated respectively to form the final key and query features, which are subsequently used to compute the affinity matrix as in Eq. (2).

The PLs thus generated provide information on the estimated objects, helping our OEG-CFN deal with the discontinuity issue effectively and improve label propagation quality by learning correspondences from the PLs.

3.3. Gradient Enhanced Image Generator

The model seeks to learn reliable correspondences by solving the slice reconstruction pre-text task [59]. In order to achieve this, the slice reconstruction pre-text task should not be solved merely in a simple way (e.g., by simply matching the pixel intensities of two neighboring slices). Sli2Vol [59] proposed an edge profile generator (i.e., computing the first-order derivatives of each pixel's intensity value) as an information bottleneck to avoid trivial solutions. This encourages the model to focus more on the edges during slice reconstruction. However, it still has several drawbacks: (1) The edge profile generator that utilizes first-order derivatives is highly sensitive to noise; small variations in pixel intensity can cause significant changes in the derivative values, leading to sub-optimal performance. (2) It can detect regions accurately only with high-intensity changes, resulting in inaccurate representation of the edges.

To address these issues, we propose to generate *gradient-enhanced images*. Specifically, given a slice, we first transform the intensity value of each pixel into a normalized histogram of second-order derivatives, computed in d different directions and at s different scales, and then apply softmax normalization across all the derivative values, as:

$$G(s, d, p) = \text{softmax}\left(\frac{\partial^2}{\partial x_{1,1}^2} I(p), \frac{\partial^2}{\partial x_{1,2}^2} I(p), \ldots, \frac{\partial^2}{\partial x_{(d,s)}^2} I(p)\right), \tag{3}$$

where $I(p)$ denotes the intensity value of a pixel p, and $\partial^2 / \partial x_{i,j}^2$ is the second-order derivative along direction i at scale j. Here, scales refer to the window sizes used to compute the second-order derivative of the center pixel in the window. After that, we concatenate these values with the intensity values to form a gradient-enhanced image. We refer to this as the gradient-enhanced image generator (GEIG).

Using the gradient-enhanced images, the model learns the correspondences. This method is designed to enable the model to learn more reliable correspondences during slice reconstruction because of several benefits of the gradient-enhanced images: (1) They are capable of mitigating noise

sensitivity by considering the rate of change of the gradient rather than the intensity values; (2) they enable more precise edge localization by detecting zero-crossings, which correspond to the actual positions of the edges.

3.4. Inference with Mask Propagation

Our **Sli2Vol+** takes a pair of slices as input and finds the correspondences that map one slice to the other. It is trained on all pairs of consecutive training slices. Once trained, in the test stage, given two consecutive slices x, x', it can compute the correspondences that map x to its neighboring slice x'. Thus, the mask of x' can be estimated by applying the correspondences to the mask of x, allowing the masks of the neighboring slices of a labeled slice to be generated.

Specifically, given a test volume $X_{test} = (S'_1, S'_2, \ldots, S'_{D'})$, two consecutive slices S'_i and S'_{i+1} are sampled from X_{test}. The gradient-enhanced images are then computed using GEIG (Section 3.3), and these images are fed to the trained model to obtain the affinity matrix $\mathbf{A}_{i \to i+1}$. This matrix is then used to propagate a mask $\hat{\mathbf{M}}_i$ of slice S'_i to generate the mask $\hat{\mathbf{M}}_{i+1}$ of slice S'_{i+1}, as:

$$\hat{\mathbf{M}}_{i+1}(u,:) = \sum_{v} \mathbf{A}_{i \to i+1}(u,v)\hat{\mathbf{M}}_i(v,:). \tag{4}$$

Note that, unlike in the training stage, no PLs of the slices are needed for computing the affinity matrix during the test stage. This process of computations is then repeated for another two consecutive slices, say $\mathbf{S'}_{i+1}$ and $\mathbf{S'}_{i+2}$, in either direction until the whole test volume is covered (an example is shown in Fig. 1(c)).

4. Experiments

4.1. Datasets

We evaluate our proposed **Sli2Vol+** framework on nine public datasets (in CT and MRI) with ten different SOIs.

In the CT modality, we train our model on three datasets: C4KC-KiTS [16], CT-LN [41], and CT-Pancreas [40], and test on seven different CT datasets: Sliver07 [50], CHAOS [22], 3Dircadb-01 and 3Dircadb-02 [46], and Decath-Spleen, Decath-Liver, and Decath-Pancreas [45], as well as one MRI dataset, Decath-Heart [45], spanning nine SOIs: Left atrium (LA), Liver (Liv), Spleen (Spl), Pancreas (Pan), Heart (Hea), Gallbladder (Gal), Kidney (Kid), Surrenal gland (Sur), and Lung (Lun).

In the MRI modality, we train our model on the Decath-Brain Tumours dataset [45], with the FLAIR, T1w, T1gd, and T2w modals. Specifically, we train our model on the T1w modal and test on the FLAIR, T1gd, and T2w modals.

We repeat the experiments five times with different random seeds and report the mean Dice scores $\pm$ standard deviation.

4.2. Implementation Details

Our experiments are conducted using the PyTorch library. Model training is performed on an NVIDIA A40 Graphics Card with 48GB GPU memory, utilizing the AdamW optimizer [29] with a weight decay of 0.005. The learning rate is set to 0.0001, and the number of training epochs is 4 for the experiments. The batch size for each case is set to the maximum size allowed by the GPU.

For UNETR++ [43], the learning rate is 0.01, with 1000 training epochs. The optimizer is SGD with a weight decay of 0.00003 and a momentum of 0.99.

4.3. Baseline Comparison

Following Sli2Vol [59], we compare our **Sli2Vol+** with three types of baselines. (1) We compare two approaches trained on fully annotated 3D data, demonstrating the performance of SOTA fully supervised models with or without domain shifts. **Fully Supervised-Same Domain (FS-SD)** refers to the scenario where the training and testing data come from the same dataset. Results of SOTA methods [1, 20, 23, 49] and UNETR++ [43] trained by us are reported. **Fully Supervised-Different Domain (FS-DD)** aims to evaluate the generalizability of fully supervised approaches when training and testing data come from different domains. (2) We consider the scenario when only a single annotated slice is provided in each test volume to train a UNETR++ [43] model, referred to as **Fully Supervised-Single Slice (FS-SS)**. For example, in the CHAOS [22] dataset, the UNETR++ model is trained on 20 annotated slices (a single slice from each volume) and tested on the same set of 20 volumes. This approach utilizes the same amount of manual annotations as **Sli2Vol+**, to investigate whether a model trained on single slice annotations is sufficient to generalize to the whole volume. (3) We compare with several SOTA mask propagation methods, including **VoxelMorph (VM) [2], Sli2Vol [59], and Vol2Flow [3]**.

Following Sli2Vol [59], we randomly pick one of the ± 3 slices around the slice with the largest ground truth (GT) annotation as an annotated slice.

5. Results and Discussions

5.1. Quantitative Results

We compare our proposed **Sli2Vol+**, which needs a single annotated slice in each training and testing volume, with two methods trained with fully annotated data and four semi-supervised methods which require only a single annotated slice per test volume.

Table 2 presents a quantitative comparison of various methods on datasets in both CT and MRI. From these results, we observe the following. (1) Compared to the SOTA performance achieved by the Fully Supervised-Same Domain (FS-SD) (**row (a)**), a significant performance drop

Modality	*Abdominal and Chest CT*														
Training Dataset (for rows (e) to (i))	*C4KC-KiTS, CT-LN, and CT-Pancreas*														
Testing Dataset	*Decath-Hea (MRI)*	*Sliver07*	*CHAOS*	*Decath-Liv*	*Decath-Spl*	*Decath-Pan*	*3D-IRCADb-01 and 3D-IRCADb-02*								
ROI	*LA*	*Liv*	*Liv*	*Liv*	*Spl*	*Pan*	*Hea*	*Gal*	*Kid*	*Sur*	*Liv*	*Lun*	*Pan*	*Spl*	*Mean*
# of Volumes	*20*	*20*	*20*	*131*	*41*	*281*	*3*	*8*	*17*	*11*	*22*	*12*	*4*	*7*	
Trained with Fully Annotated Data															
(a) FS-SD	92.7 [20] (**94.4**)	94.8 [1] (**96.2**)	**97.8** [23] (96.4)	95.4 [20] (**95.9**)	96.0 [20] (**97.1**)	79.3 [20] (**79.5**)	- (**97.7**)	- (**72.4**)	- (**97.1**)	- (**69.7**)	**96.5** [49] (**96.5**)	- (**96.9**)	- (**72.4**)	- (**94.2**)	- (**89.7**)
(b) FS-DD	-	74.8 ±13.2	76.5 ±8.8	56.0 ±23.6	-	-	-	-	-	-	-	-	-	-	-
Semi-supervised															
(c) FS-SS	47.7 ±6.8	85.3 ±5.1	79.1 ±6.4	84.6 ±2.7	75.6 ±9.6	50.3 ±11.5	23.3 ±6.9	44.6 ±16.1	58.4 ±14.3	27.7 ±16.3	81.2 ±8.2	80.8 ±9.9	18.2 ±6.7	59.2 ±5.4	58.3
(d) VM [2]	39.5 ±7.6	57.2 ±9.8	66.5 ±10.5	38.5 ±12.5	61.5 ±19.5	21.4 ±6.7	20.3 ±6.5	20.2 ±12.2	70.1 ±18.6	41.1 ±15.3	60.5 ±9.7	38.7 ±21.2	28.3 ±11.0	54.1 ±12.4	41.1
(e) Sli2Vol [59]	51.6 ±8.2	87.9 ±6.3	89.4 ±3.1	84.0 ±8.8	88.7 ±7.7	51.4 ±10.6	79.4 ±7.3	43.2 ±24.3	92.2 ±5.8	45.0 ±15.6	87.0 ±3.6	80.8 ±6.2	54.4 ±6.4	91.6 ±4.0	73.3
(f) Vol2Flow [3]	51.1 ±9.6	92.1 ±4.8	84.4 ±4.1	85.4 ±6.5	88.7 ±10.2	57.3 ±7.3	80.3 ±6.9	57.2 ±16.4	88.4 ±6.2	42.3 ±14.4	88.6 ±2.7	83.5 ±2.2	62.4 ±9.5	87.4 ±7.5	74.9
Sli2Vol+	***Ablation Study***														
(g) PLs + OEG-CFN	52.3 ±8.4	88.2 ±3.9	89.8 ±2.3	85.3 ±6.0	90.2 ±4.7	50.6 ±12.2	88.1 ±5.4	64.2 ±8.4	92.7 ±4.6	49.6 ±17.4	86.2 ±3.2	93.2 ±5.4	52.4 ±7.3	91.1 ±3.6	76.7
(h) Refined PLs + OEG-CFN	52.4 ±7.3	88.4 ±3.3	90.6 ±2.5	86.7 ±7.9	90.7 ±3.5	50.7 ±10.3	88.7 ±4.9	65.1 ±8.2	93.2 ±4.4	50.1 ±17.1	86.8 ±3.3	94.5 ±2.0	52.2 ±7.9	91.8 ±2.7	77.3
(i) Refined PLs + OEG-CFN + GEIG (Ours)	54.9 ±7.4	88.7 ±5.1	91.2 ±3.0	88.4 ±5.6	92.6 ±3.3	52.0 ±10.1	90.6 ±2.9	68.7 ±7.0	94.2 ±2.1	52.0 ±17.2	87.4 ±2.8	96.3 ±1.9	54.4 ±6.1	92.8 ±2.8	78.9

Table 2. Results (mean Dice scores ± standard deviation) of different methods on various datasets. **Row (a)** gives results of SOTA methods [1, 20, 23, 49] and UNETR++ [43] trained by us (values in the brackets). The results in **row (a)** and **row (b)** are only partially available in the literature and are reported to demonstrate the approximate upper-bound and limitations of the fully supervised methods, which are not meant to be compared directly to our proposed approach. The best fully supervised results are marked in **bold**. The best results of the semi-supervised methods are underlined. The same for Table 3.

Modality	*MRI*			
Training Dataset (for rows (e) to (i))	*Brain Tumor (T1w modal)*			
Testing Dataset	*FLAIR modal*	*T1gd modal*	*T2w modal*	*Mean*
ROI	*Tumor*	*Tumor*	*Tumor*	
# of Volumes	*266*	*266*	*266*	
Trained with Fully Annotated Data				
(a) FS-SD	**93.6**	**92.1**	**93.3**	**93.0**
(b) FS-DD	72.6 ±11.7	72.3 ±12.5	69.2 ±14.8	71.4
Semi-supervised				
(c) FS-SS	44.6 ±7.9	45.8 ±7.2	43.3 ±8.6	44.6
(d) VM [2]	35.9 ±8.4	36.5 ±12.4	34.4 ±10.7	35.6
(e) Sli2Vol [59]	49.5 ±7.8	50.9 ±6.6	51.3 ±7.5	50.6
(f) Vol2Flow [3]	48.8 ±9.7	51.0 ±10.1	52.1 ±8.8	50.6
Sli2Vol+	***Ablation Study***			
(g) PLs + OEG-CFN	53.9 ±7.4	51.6 ±7.1	51.9 ±6.2	52.5
(h) Refined PLs + OEG-CFN	54.2 ±7.7	52.3 ±7.2	52.2 ±6.7	52.9
(i) Refined PLs + OEG-CFN + GEIG (Ours)	54.9 ±7.6	53.6 ±6.9	53.7 ±6.4	54.1

Table 3. Results of various methods on the MRI datasets with different modals. **Row (a)** gives results of UNETR++ [43] trained by us. The results in **row (a)** and **row (b)** are reported just to demonstrate the approximate upper-bound and limitations of the fully supervised methods, which are not meant to be compared directly to our proposed approach.

(over 20 in Dice score) can be seen for cross-domain evaluation (same SOIs, different datasets) (**row (b)**), by comparing **row (b)** with the values inside the brackets in **row (a)**. Yet, our method shows a drop of less than 7 Dice score, suggesting its potential for addressing domain shift problems. (2) With the same amount of annotations (i.e., only a single annotated slice per test volume), our method (**row (i)**) significantly outperforms Fully Supervised-Single Slice (FS-SS) (**row (c)**) on all the datasets ($p < 0.05$, t-

Method	Dice
SAM [25]	0.602 ± 0.024
MedSAM [30]	0.774 ± 0.019
ScribblePrompt [53]	**0.955** ± 0.007

Table 4. Segmentation results of different segmentation foundation models or interactive segmentation tools on the Decath-Brain Tumours dataset [45].

Modality	*MRI*			
Training Dataset	*Brain Tumor (T1w modal)*			
Testing Dataset	*FLAIR modal*	*T1gd modal*	*T2w modal*	*Mean*
ROI	*Tumor*	*Tumor*	*Tumor*	
# of Volumes	*266*	*266*	*266*	
(a) Sli2Vol+ (Ours)	**54.9** ±7.6	**53.6** ±6.9	53.7 ±6.4	**54.1**
(b) Sli2Vol+ (using ScribblePrompt [53])	54.6 ±7.2	52.8 ±7.3	**53.9** ±7.1	53.8

Table 5. Results of various methods on the MRI datasets with different modals. In row (b), ScribblePrompt [53] is utilized to generate the single slice annotations of the training volumes.

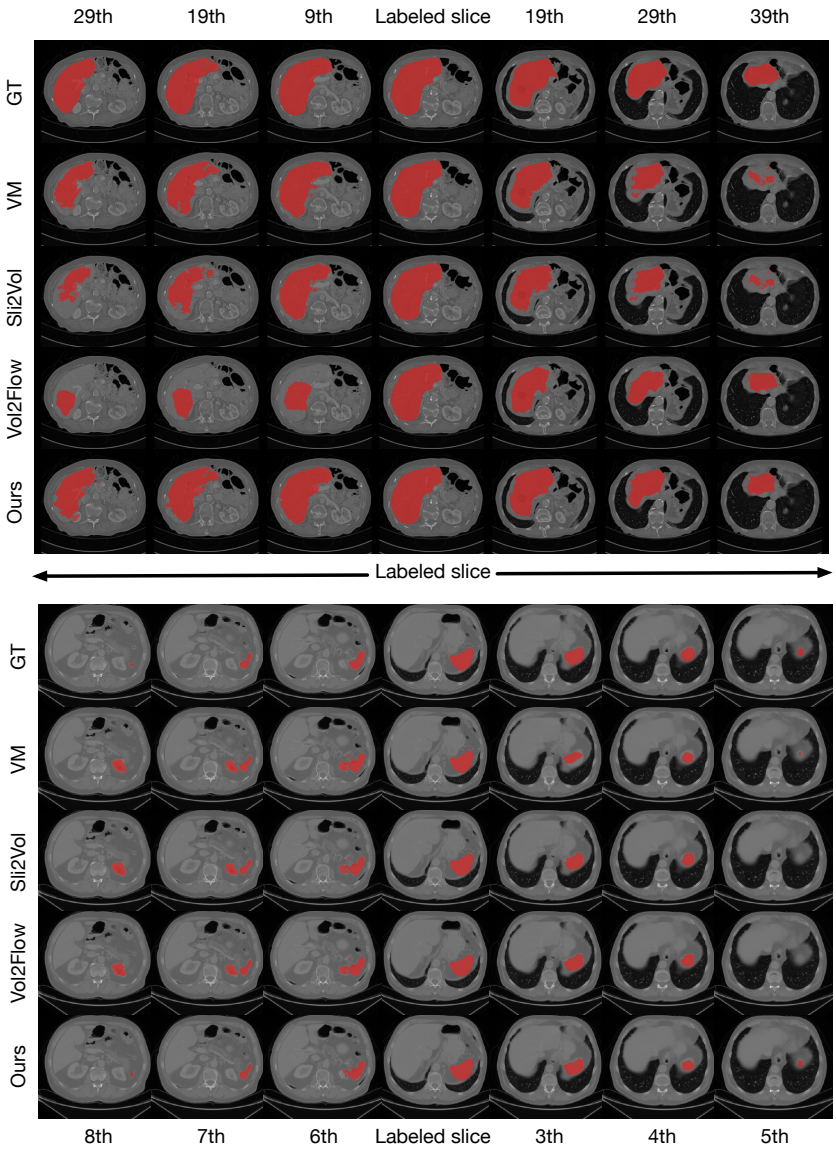

Figure 2. Examples of segmentation results by different methods on the Decath-Liver (top) and Decath-Spleen (bottom) datasets [45]. In the examples, a labeled slice is propagated in two directions, and the "i-th" represents the position of a slice to which the labeled slice is propagated in that direction.

test), with an average Dice score margin of over 20. (3) For propagation-based methods (**rows (d)-(f)**), our method outperforms VM [2], Sli2Vol [59], and Vol2Flow [3] on all the datasets. This suggests that our method incurs less severe error drifts thanks to its inclusion of PLs. Specifically, our method improves Sli2Vol [59] by an average of 5.6 Dice score, demonstrating that our assumption (i.e., including PLs to provide supervision for slice reconstruction leads to better and more reliable correspondences) is valid. (4) Our method outperforms FS-SS, VM [2], Sli2Vol [59], and Vol2Flow [3] for cross-modality evaluation (i.e., different SOIs, different modalities), showing its better generalizability in cross-modality tasks.

Table 3 gives a quantitative comparison of various methods on the MRI datasets with different modals. We use the T1w modal data for training and test the data with the FLAIR, T1gd, and T2w modals. From these results, we observe the following. (1) Our method significantly outperforms the FS-SS (**row (c)**) ($p < 0.05$, t-test), with an average Dice score margin of over 9, using only a single annotated slice per volume. This demonstrates the effectiveness of our method in reducing the annotation burden. (2) Our method outperforms the SOTA mask propagation methods (i.e., VM [2], Sli2Vol [59], and Vol2Flow [3]) in

cross-modal evaluation (i.e., same SOIs, different modals), showing its better generalizability across different modals.

5.2. Qualitative Results

Fig. 2 presents some examples of segmentation results by VM [2], Sli2Vol [59], Vol2Flow [3], and our method on the Decath-Liver (top) and Decath-Spleen (bottom) datasets [45], with GT given as reference.

From the visual segmentation results in Fig. 2, we can observe the following. (1) When propagating to slices that are far away from the labeled slice, the segmentation results produced by our method are significantly better than those produced by the known mask propagation methods (e.g., see the segmentation results of the 29th and 39th slices). This validates that our method better handles the error accumulation issue during propagation. (2) When propagating to slices where seen objects end, our method can still propagate well. For example, the spleen almost ends from the 7th slice to the 8th slice, yet our method generates accurate segmentation results. However, the other mask propagation methods generate false positives. This demonstrates that our method effectively deals with the discontinuity issue.

5.3. Annotation Cost Analysis

Although our method requires an additional single slice annotation per training volume compared to Sli2Vol [59], which only needs a single slice annotation per test volume, the annotation efforts for our method are still relatively low (e.g., involving only a few hundred slices even for large training datasets, such as C4KC-KiTS [16] with 300 training volumes). Considering the significant improvements it provides, the additional annotation efforts are worthwhile.

We would like to note that the additional annotation effort of **Sli2Vol+** can be further reduced by utilizing segmentation foundation models (e.g., SAM [25] and MedSAM [30]) or interactive biomedical image segmentation tools (e.g., ScribblePrompt [53]) to automatically generate annotations for training volumes. We evaluate SAM [25], MedSAM [30], and ScribblePrompt [53] on the Decath-Brain Tumours dataset [45]. Their quantitative segmentation results are reported in Table 4, and some visual results are shown in Fig. 3. Since ScribblePrompt [53] yielded the best performance in these experiments, we selected it to generate the annotations for the training volumes.

Table 5 reports the results of our method using segmentations generated by ScribblePrompt [53] (not the GT annotations). It achieves comparable performance.

5.4. Ablation Study

To examine the effects of different key components in our method, we conduct an ablation study on both the CT and MRI datasets, as shown in Tables 2 and 3. We observe the following. (1) When using our OEG-CFN to learn the correspondences between consecutive slices and PLs, the Dice scores are improved by an average of 3.4 and 1.9 (comparing row (e) and row (g)) on the CT and MRI datasets, respectively. This demonstrates the effectiveness of including PLs to learn reliable correspondences. (2) The performance is further improved slightly when refining the quality of the generated PLs. (3) When applying our proposed gradient-enhanced image generator to enhance the images, the performance is further improved by an average of 1.6 and 1.2 (comparing row (h) and row (i)) on the CT and MRI datasets, respectively. This verifies the effectiveness of the gradient-enhanced image generator.

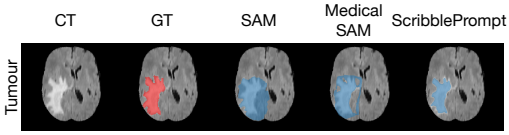

Figure 3. Examples of segmentation results by various segmentation foundation models or interactive segmentation tools on the Decath-Brain Tumours dataset [45].

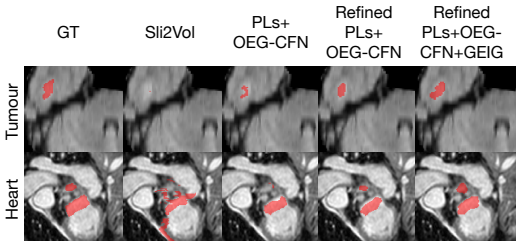

Figure 4. Examples of segmentation results generated by our method when combining different key components on the Decath-Brain Tumours and Heart datasets [45].

Fig. 4 presents some visual results. From these results, one can see the following. (1) When using our OEG-CFN to learn the correspondences from the slices and PLs, the quality of segmentation results is significantly improved. (2) When refining the PLs, the segmentation quality is further enhanced. (3) When applying the gradient-enhanced image generator, the model is able to detect more precise edges.

6. Conclusions

In this paper, we presented a new self-supervised mask propagation framework, **Sli2Vol+**, for segmenting any anatomical structures in 3D medical images using only a single annotated slice per training and testing volume. Our **Sli2Vol+** can learn reliable correspondences between consecutive slices and pseudo-labels by utilizing information on estimated objects provided by PLs, effectively addressing the error drift and discontinuity issues. Experiments on nine public datasets spanning ten different SOIs demonstrated the effectiveness of our new **Sli2Vol+** framework.

References

[1] Mubashir Ahmad, Danni Ai, Guiwang Xie, Syed Furqan Qadri, Hong Song, Yong Huang, Yongtian Wang, and Jian Yang. Deep belief network modeling for automatic liver segmentation. *IEEE Access*, 7:20585–20595, 2019. 5, 6

[2] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019. 5, 6, 7, 8

[3] Adeleh Bitarafan, Mohammad Farid Azampour, Kian Bakhtari, Mahdieh Soleymani Baghshah, Matthias Keicher, and Nassir Navab. Vol2Flow: Segment 3D volumes using a sequence of registration flows. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 609–618, 2022. 2, 5, 6, 7, 8

[4] Adeleh Bitarafan, Mahdi Nikdan, and Mahdieh Soleymani Baghshah. 3D image segmentation with sparse annotation by self-training and internal registration. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2665–2672, 2021. 2

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-UNet: UNet-like pure Transformer for medical image segmentation. In *European Conference on Computer Vision Workshops*, pages 205–218, 2023. 1

[6] M Emre Celebi, Hassan A Kingravi, Bakhtiyar Uddin, Hitoshi Iyatomi, Y Alp Aslandogan, William V Stoecker, and Randy H Moss. A methodological approach to the classification of dermoscopy images. *Computerized Medical Imaging and Graphics*, 31(6):362–373, 2007. 1

[7] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8988–8997, 2020. 1

[8] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 4

[9] Chengliang Dai, Shuo Wang, Yuanhan Mo, Kaichen Zhou, Elsa D. Angelini, Yike Guo, and Wenjia Bai. Suggestive annotation of brain tumour images with gradient-guided sampling. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, volume 12264, pages 156–165, 2020. 2

[10] Vien Ngoc Dang, Francesco Galati, Rosa Cortese, Giuseppe Di Giacomo, Viola Marconetto, Prateek Mathur, Karim Lekadir, Marco Lorenzi, Ferran Prados, and Maria A Zuluaga. Vessel-CAPTCHA: An efficient learning framework for vessel annotation and segmentation. *Medical Image Analysis*, 75:102263, 2022. 1

[11] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson, et al. Segment Anything Model (SAM) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023. 1

[12] Pengfei Gu, Yejia Zhang, Chaoli Wang, and Danny Z Chen. ConvFormer: Combining CNN and Transformer for medical image segmentation. In *IEEE International Symposium on Biomedical Imaging*, pages 1–5, 2023. 1

[13] Pengfei Gu, Hao Zheng, Yizhe Zhang, Chaoli Wang, and Danny Z Chen. kCBAC-Net: Deeply supervised complete bipartite networks with asymmetric convolutions for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347, 2021. 1

[14] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. UNETR: Transformers for 3D medical image segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1748–1758, 2022. 1, 4

[15] Sheng He, Rina Bao, Jingpeng Li, P Ellen Grant, and Yangming Ou. Accuracy of Segment-Anything Model (SAM) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324*, 2023. 1

[16] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, et al. C4kc kits challenge kidney tumor segmentation dataset, 2019. 5, 8

[17] Chuanfei Hu and Xinde Li. When SAM meets medical images: An investigation of Segment Anything Model (SAM) on multi-phase liver tumor segmentation. *arXiv preprint arXiv:2304.08506*, 2023. 1

[18] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. 1

[19] Alex Ling Yu Hung, Haoxin Zheng, Kai Zhao, Xiaoxi Du, Kaifeng Pang, Qi Miao, Steven S. Raman, Demetri Terzopoulos, and Kyunghyun Sung. CSAM: A 2.5D Cross-Slice attention module for anisotropic volumetric medical image segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision WACV, 2024*, pages 5911–5920, 2024. 2

[20] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 1, 5, 6

[21] Davood Karimi, Serge Didenko Vasylechko, and Ali Gholipour. Convolution-free medical image segmentation using Transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 78–88, 2021. 1

[22] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 5

[23] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, et al. CHAOS Challenge - Combined (CT-MR) Healthy Abdominal Organ Segmentation, Jan. 2020. 5, 6

[24] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1665–1674, 2017. 1

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 6, 8

[26] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, volume 9908, pages 695–711, 2016. 1

[27] Amarjeet Kumar, Hongxu Jiang, Muhammad Imran, Cyndi Valdes, Gabriela Leon, Dahyun Kang, Parvathi Nataraj, Yuyin Zhou, Michael D. Weiss, and Wei Shao. A flexible 2.5D medical image segmentation approach with In-Slice and Cross-Slice attention. *arXiv:2405.00130*, 2024. 2

[28] Shumeng Li, Heng Cai, Lei Qi, Qian Yu, Yinghuan Shi, and Yang Gao. PLN: Parasitic-like network for barely supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2022. 2

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[30] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 1, 6, 8

[31] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment Anything Model for medical image analysis: An experimental study. *arXiv preprint arXiv:2304.10517*, 2023. 1

[32] Sovesh Mohapatra, Advait Gosai, and Gottfried Schlaug. SAM vs BET: A comparative study for brain extraction and segmentation of magnetic resonance images using deep learning. *arXiv preprint arXiv:2304.04738*, 2023. 1

[33] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 1

[34] Yousuf Babiker M Osman, Cheng Li, Weijian Huang, Nazik Elsayed, Zhenzhen Xue, Hairong Zheng, and Shanshan Wang. Semi-supervised and self-supervised collaborative learning for prostate 3D MR image segmentation. *arXiv preprint arXiv:2211.08840*, 2022. 2

[35] Danielle F Pace, Adrian V Dalca, Tal Geva, Andrew J Powell, Mehdi H Moghari, and Polina Golland. Interactive whole-heart segmentation in congenital heart disease. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 80–88, 2015. 1

[36] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *IEEE International Conference on Computer Vision*, pages 1742–1750, 2015. 1

[37] Hui Qu, Pengxiang Wu, Qiaoying Huang, Jingru Yi, Zhennan Yan, Kang Li, Gregory M. Riedlinger, Subhajyoti De, Shaoting Zhang, and Dimitris N. Metaxas. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE Transactions on Medical Imaging*, 39(11):3655–3666, 2020. 1

[38] Ilija Radosavovic, Piotr Dollár, Ross B. Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4119–4128, 2018. 1

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pages 234–241, 2015. 1

[40] Holger Roth, Amal Farag, Evrim B. Turkbey, Le Lu, Jiamin Liu, and Ronald M. Summers. Data from pancreas-CT, 2016. 5

[41] Holger Roth, Le Lu, Ari Seff, Kevin M Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, and Ronald M. Summers. A new 2.5 D representation for lymph node detection in CT, 2015. 5

[42] Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H Maier-Hein. SAM.MD: Zero-shot medical image segmentation capabilities of the Segment Anything Model. *arXiv preprint arXiv:2304.05396*, 2023. 1

[43] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. UNETR++: Delving into efficient and accurate 3D medical image segmentation. *arXiv preprint arXiv:2212.04497*, 2022. 1, 4, 5, 6

[44] Xueying Shi, Qi Dou, Cheng Xue, Jing Qin, Hao Chen, and Pheng-Ann Heng. An active learning approach for reducing annotation cost in skin lesion analysis. In *International Workshop on Machine Learning in Medical Imaging*, volume 11861, pages 628–636, 2019. 2

[45] Amber L Simpson, Michela Antonelli, Spyridon Bakas, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 4, 5, 6, 7, 8

[46] L Soler, A Hostettler, V Agnus, A Charnoz, J Fasquel, J Moreau, A Osswald, M Bouhadjar, and J Marescaux. 3D image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. *IRCAD, Strasbourg, France, Tech. Rep*, 2010. 5

[47] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3136–3145, 2019. 1

[48] Xing Tao, Yuexiang Li, Wenhui Zhou, Kai Ma, and Yefeng Zheng. Revisiting Rubik's cube: Self-supervised learning with volume-wise transformation for 3D medical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, volume 12264, pages 238–248, 2020. 1

[49] Song-Toan Tran, Ching-Hwa Cheng, and Don-Gey Liu. A multiple layer U-Net, U^n-Net, for liver and liver tumor segmentation in CT. *IEEE Access*, 2020. 5, 6

[50] Bram Van Ginneken, Tobias Heimann, and Martin Styner. 3D segmentation in the clinic: A grand challenge. In *International Conference on Medical Image Computing and Computer-assisted Intervention Workshop on 3D Segmentation in the Clinic: A Grand Challenge*, volume 1, pages 7–15, 2007. 5

[51] Guotai Wang, Wenqi Li, Maria A. Zuluaga, Rosalind Pratt, Premal A. Patel, Michael Aertsen, Tom Doel, Anna L. David, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Transactions Medical Imaging*, 37(7):1562–1573, 2018. 1

[52] Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, Xin Liu, Jie Chen, Huihui Zhou, Ismail Ben Ayed, and Hairong Zheng. Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*, 12(1):1–13, 2021. 2

[53] Hallee E. Wong, Marianne Rakic, John Guttag, and Adrian V. Dalca. ScribblePrompt: Fast and flexible interactive segmentation for any biomedical image. *arXiv:2312.07381*, 2024. 6, 8

[54] Yixuan Wu, Bo Zheng, Jintai Chen, Danny Z Chen, and Jian Wu. Self-learning and one-shot learning based single-slice annotation for 3D medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 244–254, 2022. 2

[55] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. CoTr: Efficiently bridging CNN and Transformer for 3D medical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 171–180, 2021. 4

[56] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. CAMEL: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10682–10691, 2019. 1

[57] Yan Xu, Jun-Yan Zhu, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical Image Analysis*, 18(3):591–604, 2014. 1

[58] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, volume 10435, pages 399–407, 2017. 2

[59] Pak-Hei Yeung, Ana IL Namburete, and Weidi Xie. Sli2Vol: Annotate a 3D volume from a single slice with self-supervised learning. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 69–79, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[60] Jian Zhang, Yinghuan Shi, Jinquan Sun, Lei Wang, Luping Zhou, Yang Gao, and Dinggang Shen. Interactive medical image segmentation via a point-based interaction. *Artificial Intelligence in Medicine*, 111:101998, 2021. 1

[61] Yejia Zhang, Pengfei Gu, Nishchal Sapkota, Hao Zheng, Peixian Liang, and Danny Z Chen. A point in the right direction: Vector prediction for spatially-aware self-supervised volumetric representation learning. In *IEEE 20th International Symposium on Biomedical Imaging*, pages 1–5, 2023. 1

[62] Yejia Zhang, Nishchal Sapkota, Pengfei Gu, Yaopeng Peng, Hao Zheng, and Danny Z Chen. Keep your friends close & enemies farther: Debiasing contrastive learning with spatial priors in 3D radiology images. In *IEEE International Conference on Bioinformatics and Biomedicine*, pages 1824–1829, 2022. 1

[63] Yizhe Zhang, Lin Yang, Jianxu Chen, Maridel Fredericksen, David P. Hughes, and Danny Z. Chen. Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 10435, pages 408–416, 2017. 1

[64] Hao Zheng, Lin Yang, Jianxu Chen, Jun Han, Yizhe Zhang, Peixian Liang, Zhuo Zhao, Chaoli Wang, and Danny Z. Chen. Biomedical image segmentation via representative annotation. In *The 33rd AAAI Conference on Artificial Intelligence*, pages 5901–5908, 2019. 1, 2

[65] Hao Zheng, Yizhe Zhang, Lin Yang, Chaoli Wang, and Danny Z. Chen. An annotation sparsification strategy for 3D medical image segmentation via representative selection and self-training. In *The 34th AAAI Conference on Artificial Intelligence*, pages 6925–6932, 2020. 2

[66] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnFormer: Interleaved Transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021. 1

[67] Tao Zhou, Yizhe Zhang, Yi Zhou, Ye Wu, and Chen Gong. Can SAM segment polyps? *arXiv preprint arXiv:2304.07583*, 2023. 1

[68] Yuyin Zhou, Yan Wang, Peng Tang, Song Bai, Wei Shen, Elliot K. Fishman, and Alan L. Yuille. Semi-supervised multi-organ segmentation via deep multi-planar co-training. In *IEEE Winter Conference on Applications of Computer Vision*, pages 121–140, 2019. 1

[69] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, volume 11045, pages 3–11, 2018. 1

[70] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B. Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3D medical image analysis. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, volume 11767, pages 384–393, 2019. 1

Appendix

Fig. 1 gives some examples of segmentation results by VM [1], Sli2Vol [7], Vol2Flow [2], and our method on on the Decath-Liver, Decath-Spleen, Decath-Heart, and Decath-Brain Tumours datasets [5], with GT given as reference. From the visual segmentation results, we can observe that the segmentation results produced by our method are significantly better than those produced by the known mask propagation methods on all the four datasets. In particular, our method generates accurate segmentation results, while the other mask propagation methods generate false negatives on the Decath-Liver and Decath-Brain Tumours datasets. This demonstrates the effectiveness of our method.

Fig. 2 presents some visual results from various segmentation foundation models (i.e., SAM [3] and MedSAM [4]) or interactive segmentation tools (i.e., ScribblePrompt [6]) on the Decath-Liver, Decath-Spleen, and Decath-Heart datasets [5]. From the visual segmentation results, we can observe that the segmentation results produced by ScribblePrompt [6] are better than those produced by the known segmentation foundation models on all the three datasets. Fig. 3 shows some examples of segmentation results generated by our method when combining different key components on the Decath-Liver and Decath-Spleen datasets [5]. From the visual segmentation results, we can observe that the segmentation results are improved when different key components are combined. This demonstrates the effectiveness of the proposed components.

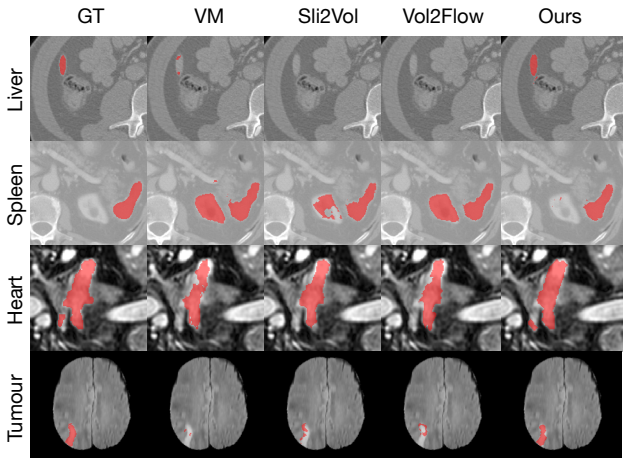

Figure 1. Examples of segmentation results by different methods on the Decath-Liver, Decath-Spleen, Decath-Heart, and Decath-Brain Tumours datasets [5].

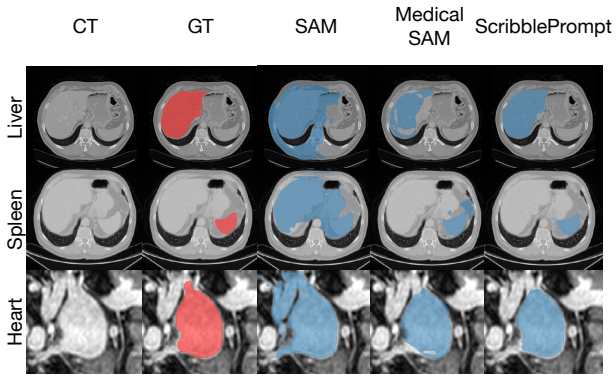

Figure 2. Examples of segmentation results by various segmentation foundation models or interactive segmentation tools on the Decath-Liver, Decath-Spleen, and Decath-Heart datasets [5].

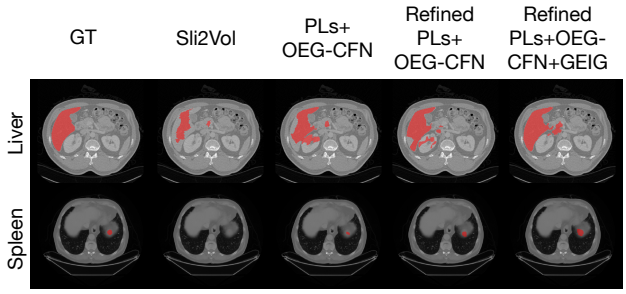

Figure 3. Examples of segmentation results generated by our method when combining different key components on the Decath-Liver and Decath-Spleen datasets [5].

References

[1] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019. 1

[2] Adeleh Bitarafan, Mohammad Farid Azampour, Kian Bakhtari, Mahdieh Soleymani Baghshah, Matthias Keicher, and Nassir Navab. Vol2Flow: Segment 3D volumes using a sequence of registration flows. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 609–618, 2022. 1

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[4] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 1

[5] Amber L Simpson, Michela Antonelli, Spyridon Bakas, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 1

[6] Hallee E. Wong, Marianne Rakic, John Guttag, and Adrian V. Dalca. ScribblePrompt: Fast and flexible interactive segmentation for any biomedical image. *arXiv:2312.07381*, 2024. 1

[7] Pak-Hei Yeung, Ana IL Namburete, and Weidi Xie. Sli2Vol: Annotate a 3D volume from a single slice with self-supervised learning. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 69–79, 2021. 1