Slice and Conquer: A Planar-to-3D Framework for Efficient Interactive Segmentation of Volumetric Images

Wonwoo Cho*[1,2] Dongmin Choi*[1,2] Hyesu Lim*[2] Jinho Choi[2]
Saemee Choi[1,2] Hyun-seok Min[3] Sungbin Lim[4] Jaegul Choo[1,2]
[1]Letsur Inc. [2]KAIST [3]Tomocube Inc. [4]Korea University
{wcho, dmchoi, hyesulim, jinhochoi, saemee99, jchoo}@kaist.ac.kr
hsmin@tomocube.com, sungbin@korea.ac.kr

Abstract

Interactive segmentation methods have been investigated to address the potential need for additional refinement in automatic segmentation via human-in-the-loop techniques. For accurate segmentation of 3D images, we propose ***Slice-and-Conquer****, a novel planar-to-3D pipeline formulating volumetric mask construction into two stages: 1) 2D interactive segmentation and 2) guided 3D segmentation. Specifically, the first stage enables users to focus on a single 2D slice and provides the corresponding 2D prediction results as strong shape priors. Taking the planar guidance, an accurate 3D mask can be constructed with minimal interactions. To support a flexible iterative refinement, our system recommends a next slice to annotate at the end of the second stage. Since volumetric segmentation can be completed by consecutively annotating a few recommended 2D slices, our method significantly reduces the cognitive burden of exploring volumetric space for users. Through extensive experiments on various datasets of 3D biomedical images, we demonstrate the effectiveness of the proposed pipeline.*

1. Introduction

Image segmentation, which aims to extract sub-regions of interest, is essential in various applications [19]. For instance, localizing instances (*e.g.*, lesions or organs) and accurately quantifying their volumes in medical images can be a fundamental stage of clinical interventions. Although machine learning (ML)-based algorithms have shown remarkable segmentation performance, which can even be comparable to manual segmentation of domain experts [12], such automatic methods still need further improvement to ensure high accuracy and robustness. To overcome the problem, one can use *interactive segmentation* methods, which allow segmentation models to take user-provided hints. Using interactive tools (*e.g.*, points, bounding boxes, or scribbles), users can obtain satisfying segmentation results.

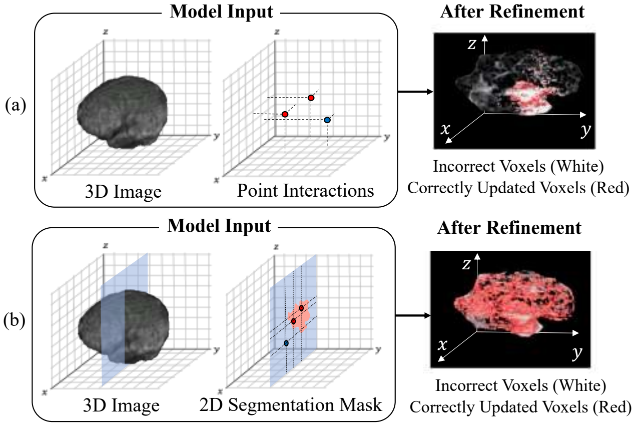


Figure 1. (a) 3D propagation of point interactions. (b) Gradually propagated into a 2D slice first and then to the 3D image, the same number of points can refine larger regions. The incorrect and refined regions are highlighted in white and red respectively.

In the natural image domain, such interactive techniques have been widely studied. After Xu *et al.* [29] introduced deep interactive object selection (DIOS), the first deep interactive segmentation method built upon convolutional neural networks (CNNs), a number of techniques such as iterative training [16] and test-time optimization (backpropagation) [8, 22] were developed to enhance segmentation accuracy. Recently, Sofiiuk *et al.* [23] presented robust interactive segmentation results by using the mask prediction of the previous iteration as an additional cue at each iteration and leveraging diverse and large-scale datasets.

In comparison with 2D applications, interactive segmentation techniques for volumetric images have received rela-

[*] Equal contribution

tively less attention despite their necessity. 3D biomedical imaging can be a representative application requiring accurate volumetric segmentation, where target images usually have low contrast and ambiguous boundaries. Thus, it is required to effectively propagate user-provided hints designed for 2D interfaces in volumetric space. To achieve the goal, previous studies attempted to encode user interactions into a complex distance map (*e.g.,* Geodesic encoding) [15, 25]. However, especially in 3D applications, such complicated distance encoding requires high computational costs at each iteration. Another potential issue of such volumetric propagation methods is that the task imposes a high cognitive burden on users since it is required to review mis-segmented regions in 3D space, *e.g.,* manually inspecting several 2D slices, causing inflexible human-in-the-loop segmentation.

Another source of difficulty in making interactive segmentation via 3D CNNs challenging is data deficiency. As conventional 3D segmentation models require a large number of parameters to handle the entire volume, it is essential to use large-scale and diverse data to effectively optimize the parameters. Without large-scale datasets or effective transfer learning techniques, networks may not sufficiently learn the volumetric structure of the target foreground. As an alternative, Zhou *et al.* [34] proposed a two-staged *slice propagation* method, which first annotates axial slices via 2D models and then propagates them to the other slices back and forth. Despite its need for multiple segmentation inferences in the second round, thus slowing interactive responses, this method might still underperform as it spreads 2D mask information without considering the full 3D volume.

To alleviate the aforementioned problems, we propose a novel two-stage pipeline for effective planar-to-3D propagation, called *Slice and Conquer (SnC)*. In the first stage, SnC forces users to annotate a single slice extracted from each 3D volume. Propagating the user interactions and 3-axis 2D mask predictions to the full volume via a 3D model, the second stage produces an accurate 3D mask (Figure 1 illustrates the effectiveness of mask guidance). To support a flexible iterative refinement, we design a slice recommendation algorithm, which aims to effectively bridge the first and second stages by selecting a challenging slice to be refined in the next iteration. By solving 3D interactive segmentation by 3-axis planar-to-3D propagation, the users are able to complete the task by concentrating only on a few slices.

For efficient and robust segmentation of volumetric images, we exploit various techniques to improve SnC. First, we consecutively train 2D and 3D models, where the trained parameters of the 2D model are employed as the initial state of the 3D model. Replacing expensive 3D convolution with ACS convolution [31], our approach makes the 3D model benefit from transfer learning. Furthermore, we incorporate a mask attention module into the 3D model for better utilization of mask guidance. To enhance the alignment between the user-completed 2D results and the 3D mask prediction, we additionally design a post-processing technique, which is possible only in planar-to-3D pipelines. Through our extensive experiments using various volumetric biomedical images, we demonstrate the effectiveness of our two-stage pipeline and its sub-components.

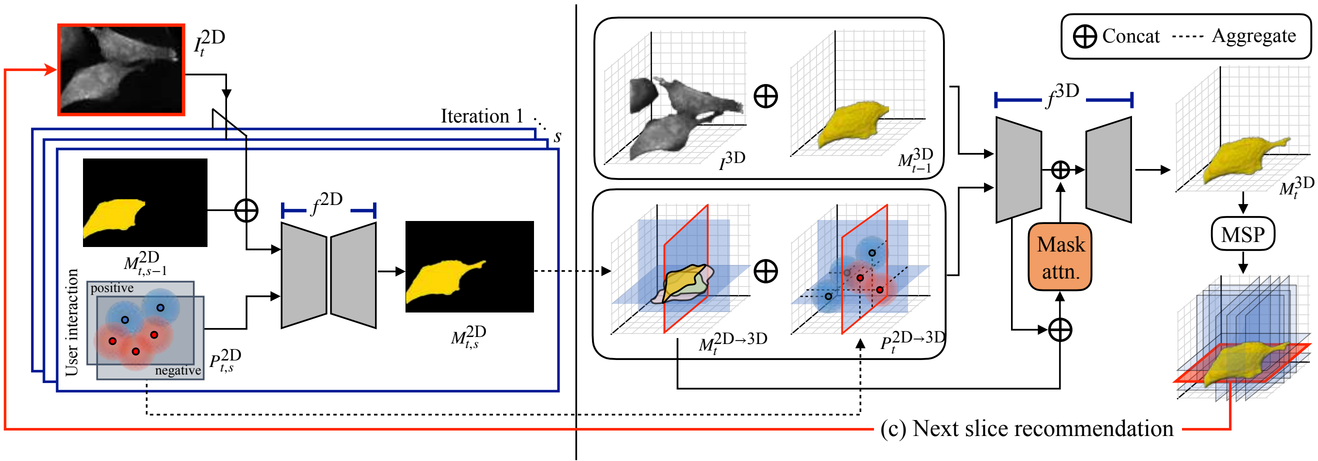
2. Related Work and Our Contribution

2.1. Interactive Segmentation

Deep interactive segmentation. In [29], DIOS presented great potential of convolutional neural networks (CNNs) to advance the field of interactive segmentation. To improve the previous method, Mahadevan *et al.* [16] proposed an iterative training strategy. Afterwards, additional techniques have been widely investigated to enhance the robustness of interactive segmentation [8, 11, 22]. Furthermore, the authors of [23] demonstrated strong interactive segmentation performance by reviving mask predictions generated in the previous iteration, and by leveraging high-quality annotations from a large and diverse collection of datasets.

Volumetric Propagation of Interactions. In interactive segmentation of 2D natural images, models usually utilize click-based interaction and its variants such as scribble and bounding box. However, such interactive methods designed for 2D interfaces may not be suitable for medical applications that often involve 3D images. Moreover, medical images usually have low contrast and ambiguous boundaries. Therefore, it is crucial to effectively propagate user interaction information throughout the entire volume, starting from the interacted slice. To achieve this, Wang *et al.* [25] used Geodesic encoding instead of conventional Gaussian or disk encoding to transform click interactions for effective volumetric propagation. Luo *et al.* [15] later improved the encoding for better generalizability. Not only does such encoding necessitate high computational costs at each interaction, but volumetric models also require a massive number of parameters. Although these problems can lead to reduced performance on smaller datasets, it is challenging to utilize transfer learning appropriate for 3D medical images.

Slice Propagation for 3D Interactive Segmentation. As an alternative, Zhou *et al.* [33, 34] recently proposed a slice propagation approach following the principle of video object segmentation [3, 6], which disseminates information from the interacted slice to the neighboring ones and subsequently to the rest of the slices. In other words, the model uses 2D CNNs to segment each 2D slice of volumetric images, with guidance from the already annotated ones. While the slice propagation method can handle volumetric images in the form of high-resolution 2D slices and benefit from transfer learning techniques, it may neglect 3D information



(a) 2D interactive segmentation

(b) Guided 3D segmentation

Figure 2. An illustration of the inference workflow at the t-th SnC iteration. (a) Users conduct 2D interactive segmentation by clicking points until obtaining a sufficiently accurate segmentation result. (b) Taking the guidance of the accumulated point interactions and 2D mask predictions, f^{3D} conducts 3D image segmentation. The mask attention module is employed to enhance the mask guidance. (c) If the results of guided 3D segmentation are unsatisfactory, SnC recommends a next slice to annotate via maximum softmax probability (MSP) score analysis. Our post-processing steps designed to make a more accurate 3D mask prediction are depicted in Figure 3.

and demand multiple model inferences when each 2D annotation round finishes, which can significantly increase the runtime. Also, it is highly probable that the guidance from annotated slices may weaken as propagation continues.

2.2. Our Contribution

To address the problems associated with prior methods, this paper presents a novel two-stage pipeline for volumetric interactive segmentation, where its in-depth details are described in the next section. Our approach is carefully designed by taking into account the following factors:

- As the two-stage iterative pipeline consists of 1) 2D interactive segmentation followed by 2) guided 3D segmentation, users can construct an accurate 3D mask by consecutively annotating only a few 2D slices.
- The 3-axis 2D mask predictions from the first stage can serve as a strong shape prior to the second stage, thus eliminating the necessity of complex 3D distance map encoding (*e.g.,* Geodesic) for volumetric propagation.
- Our two-stage SnC pipeline enables weight parameter transfer from the 2D model to the 3D model and post-processing of 3D results in Stage 2 based on the results of Stage 1 so that the 3D model can provide more elaborated results through a single forward pass.

3. Proposed Method

Overview. Our proposed *Slice and Conquer (SnC)* aims to complete a segmentation mask from scratch or refine an existing mask for a 3D image. In SnC, a single iteration of 3D interactive segmentation (outer-loop) consists of *2D interactive segmentation (inner-loop)* and *guided 3D segmentation*, where our *slice recommendation* algorithm enables SnC to start a new iteration of 3D refinement. Throughout the paper, we consider a conventional click-based interactive scheme [23, 29], which can be easily extended to other interaction types, *e.g.*, bounding box and scribble.

By counting the number of outer- and inner-loops with t and s ($t = 0$ or $s = 0$ cases can be used for empty or pre-existing masks), respectively, we present more detailed process making a 3D segmentation prediction M_t^{3D} for a 3D image I^{3D} at the t-th SnC iteration:

1) **2D interactive segmentation** (Sec. 3.1): Users concentrate on a single slice I_t^{2D}, which is extracted from I^{3D}. By following a conventional 2D interactive segmentation scheme such as [23], f^{2D} generates the corresponding 2D segmentation mask M_t^{2D}.

2) **Guided 3D segmentation** (Sec. 3.2): Using the accumulated click interactions and $\{M_i^{2D}\}_{i=1}^{t}$ as a shape prior of the target foreground region, our 3D segmentation model f^{3D} produces a segmentation mask M_t^{3D}.

3) **Slice recommendation** (Sec. 3.3): At the first iteration ($t = 1$), users select the first slice I_1^{2D} to annotate and acquire M_1^{3D}. Afterwards, this process can be iterated until a satisfactory 3D mask is constructed. To assist the users with refining the result, our system automatically selects the next 2D slice to annotate which is highly likely to improve 3D segmentation accuracy.

For 2D interactive segmentation, f^{2D} can be optimized by a loss function $\mathcal{L}_{\text{seg}}^{\text{2D}}$, where we use the normalized focal loss (NFL) [23] as $\mathcal{L}_{\text{seg}}^{\text{2D}}$. Similarly, we train f^{3D} by $\mathcal{L}_{\text{seg}}^{\text{3D}}$, a 3D version of $\mathcal{L}_{\text{seg}}^{\text{2D}}$. To effectively reflect the mask guidance $\{M_i^{\text{2D}}\}_{i=1}^{t}$ in f_{3D}, we design a *mask attention module* for f_{3D}, which can be trained based on another NFL-based loss $\mathcal{L}_{\text{attn}}$. To sum up, the loss functions for f^{2D} and f^{3D} are

$$\mathcal{L}_{\text{2D}} = \mathcal{L}_{\text{seg}}^{\text{2D}} \quad \text{and} \quad \mathcal{L}_{\text{3D}} = \mathcal{L}_{\text{seg}}^{\text{3D}} + \lambda \mathcal{L}_{\text{attn}}, \tag{1}$$

respectively, where each term in Eq. (1) is NFL and λ is a hyperparameter. We illustrate the overall workflow of SnC in Figure 2 and provide in-depth details about each step and the corresponding components in the following subsections. Also, the SnC pipeline is summarized in Algorithm 1.

3.1. 2D Interactive Segmentation

We consider that users iteratively provide positive and negative clicks to refine mis-segmented regions. This click method can be flexibly adapted to extreme clicking [17].

User-friendly environment. At the first stage, our SnC forces users to focus on completing a 2D mask M_t^{2D} for a single slice I_t^{2D}, which significantly reduces users' cognitive burden in 3D interactive segmentation. In what follows, we provide a detailed description of the first stage:

- At the t-th SnC iteration, a user starts to refine $M_{t,0}^{\text{2D}}$, which is extracted from M_{t-1}^{3D}. When $t = 1$, $M_{1,0}^{\text{2D}}$ can be a zero matrix or a 2D slice of existing M_0^{3D}.
- Given positive or negative clicks, which aim to indicate false-negative or false-positive segmented regions, respectively, $M_{t,s-1}^{\text{2D}}$ can be refined to $M_{t,s}^{\text{2D}}$ by f^{2D}.
- At the s-th 2D refinement, the user accept $M_{t,s}^{\text{2D}}$ as M_t^{2D} if satisfactory (*e.g.*, sufficiently high accuracy).

In this process, we use disk click encoding and Conv1S module of [23], *i.e.*, f^{2D} takes $(I_t^{\text{2D}} \oplus M_{t,s-1}^{\text{2D}})$ and $P_{t,s}^{\text{2D}}$ as its input at the s-th refinement, where $P_{t,s}^{\text{2D}}$ is the disk-encoded click map. f^{2D} allows users to concentrate on annotating a single 2D slice (without considering the other slices of I^{3D}), thus serving as an intermediate module for *user-friendly interaction* between the user-side and f^{3D}.

Dense-hint generator. The 2D model f^{2D} not only provides a user-friendly environment but also serves as a *dense-hint generator*. As our SnC approach aims at constructing an accurate segmentation result of I^{3D} by *conquering only a few 2D slices*, f^{3D} of the next stage may not take sufficient hints about the shape of target foreground if we only utilize click interactions as guidance. Thus, we also exploit M_t^{2D} to provide additional guidance in 3D segmentation. To further enhance the effectiveness of mask guidance, we train f^{2D} to handle 2D slices from three orthogonal planes (axial (XY), sagittal (ZX), and coronal (YZ)). By investigating the three-axes slices of I^{3D}, users can additionally provide complementary shape information to f^{3D} that cannot be obtained when using only single-axis slices [18]. Also, it is noteworthy that the strategy of supporting 3D interactive segmentation in multiple-axes 2D slices is suitable for conventional biomedical image visualization interfaces such as [9].

Algorithm 1: Slice and Conquer Pipeline

```
input  : 3D image $I^{\text{3D}}$, 2D and 3D models $\{f^{\text{2D}}$,
         $f^{\text{3D}}\}$, and existing (or empty) mask $M_0^{\text{3D}}$
output: 3D segmentation mask $M^{\text{3D}}$
$t \leftarrow 1$
repeat
    Extract selected or recommended $I_t^{\text{2D}}$ from $I^{\text{3D}}$
    Initialize $P_{t,0}^{\text{2D}}$ and $M_{t,0}^{\text{2D}}$
    $s \leftarrow 1$
    repeat
        Update $P_{t,s}^{\text{2D}}$ by adding a new click
        $M_{t,s}^{\text{2D}} \leftarrow f^{\text{2D}}(I_t^{\text{2D}} \oplus M_{t,s-1}^{\text{2D}}, P_{t,s}^{\text{2D}})$
        $P_t^{\text{2D}} \leftarrow P_{t,s}^{\text{2D}}$ and $M_t^{\text{2D}} \leftarrow M_{t,s}^{\text{2D}}$
        $s \leftarrow s + 1$
    until $M_{t,s}^{2D}$ is satisfactory;
    Aggregate $\{P_i^{\text{2D}}\}_{i=1}^{t}$ and transform to $P_t^{\text{2D}\rightarrow\text{3D}}$
    Aggregate $\{M_i^{\text{2D}}\}_{i=1}^{t}$ and transform to $M_t^{\text{2D}\rightarrow\text{3D}}$
    $M_t^{\text{3D}} \leftarrow f^{\text{3D}}(I^{\text{3D}} \oplus M_{t-1}^{\text{3D}}, P_t^{\text{2D}\rightarrow\text{3D}} \oplus M_t^{\text{2D}\rightarrow\text{3D}})$
    Recommend a next 2D slice via Eq. (2)
    Update $M_t^{\text{3D}}$ via the post-processing of Eq. (3)
    $t \leftarrow t + 1$
until $M_t^{3D}$ is satisfactory;
$M^{\text{3D}} \leftarrow M_t^{\text{3D}}$
return $M^{\text{3D}}$
```

3.2. Guided 3D Segmentation

Transfer learning. To alleviate data deficiency issues in biomedical applications, there has been a line of research adapting the weights of pre-trained 2D feature extractors for 3D medical image analysis. For instance, previous studies [30, 31] showed that transferring ImageNet-pretrained 2D weights to 3D models can enhance 3D segmentation accuracy despite the domain difference.

Inspired by the techniques, we design our SnC pipeline to flexibly incorporate the weights of f^{2D} into f^{3D}. Specifically, we use the Mean-ACS convolution [31] technique, which interprets a 3D convolutional filter as a combination of three 2D convolutional filters of axial, sagittal, and coronal planes, thus supporting *more efficient computation* than 3D convolution. Since we train f^{2D} by exploiting 2D slices of three orthogonal planes (extracted from 3D images), its convolutional filters can be seamlessly transferred to the axial, sagittal, and coronal components of the corresponding 3D ACS convolutional filters of f^{3D}.

Remark: To the best of our knowledge, our SnC frame-

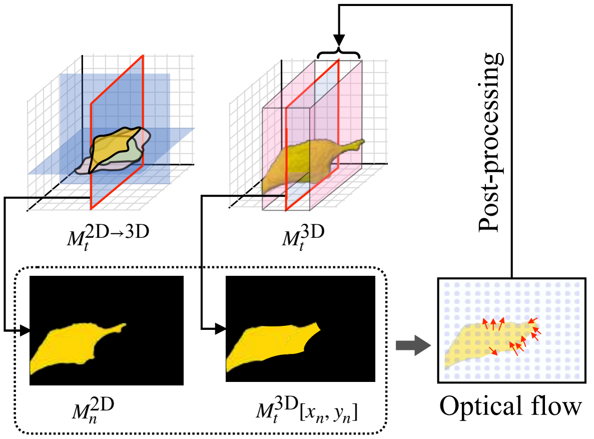


Figure 3. An illustration of our post-processing method, which aims to enhance the alignment between a user-annotated 2D mask M_n^{2D} and the corresponding 2D mask $M_t^{3D}[x_n, y_n]$ of M_t^{3D}.

work introduces the first approach adopting transfer learning in volumetric interactive segmentation models. By using our carefully designed SnC pipeline (3D segmentation following 2D interactive segmentation), we expect that f^{3D} can receive the prior knowledge of f^{2D}, which learns inter-slice contexts of target volumetric images.

Mask attention module. Making a slight modification to the input feeding method of f^{2D}, f^{3D} takes $(I^{3D} \oplus M_{t-1}^{3D})$ and $(P_t^{2D\to3D} \oplus M_t^{2D\to3D})$ at the t-th SnC iteration, where $M_t^{2D\to3D}$ and $P_t^{2D\to3D}$ are 3D volumes containing 2D masks and point interactions accumulated since the first iteration, respectively (I^{3D}, $M_t^{2D\to3D}$, and $P_t^{2D\to3D}$ have the same dimension). It is intuitive that the masks $M_t^{2D\to3D}$ can serve as a stronger shape prior than $P_t^{2D\to3D}$ to construct M_t^{3D}.

Inspired by FCA [11], we enhance the mask guidance by applying an add-on module to f^{3D} as illustrated in Figure 1 (b). As presented in Eq. (1), the mask attention module is also trained to produce 3D segmentation predictions.

3.3. Slice Recommendation and Post-processing

Next slice recommendation. Our SnC framework aims to generate an accurate 3D mask prediction via iterative refinement. In other words, if the quality of the 3D mask prediction M_t^{3D} is not satisfactory at the t-th iteration, it is necessary to initiate another iteration by selecting I_{t+1}^{2D}. However, manually investigating erroneous regions in the volumetric space, such as scrolling through multiple 2D slices back and forth, to select the next 2D slice to annotate can hinder the efficient human-in-the-loop refinement pipeline.

To support flexible iterative refinement steps by reducing user efforts in the selection of the next slice, it is necessary to recommend one that needs additional refinement. However, such processes may require additional computation or memory resources to find challenging slices in volumetric segmentation. For instance, previous studies used a quality assessment module [33, 34] or an uncertainty estimation technique based on multiple prediction results [4, 24]. In order to avoid the usage of an additional module or multiple predictions, we make an assessment policy as in [10].

For each M_t^{3D}, we compute a volumetric confidence map C_t, whose elements contain pixel-level maximum softmax probability (MSP) [5] scores corresponding to M_t^{3D}. Then, we define $C_t^{[\tau]}$, whose element is 1 if the corresponding element of C_t is greater than a threshold τ, or 0 otherwise. By using the confidence map $C_t^{[\tau]}$, our recommendation algorithm selects $C_t^{[\tau]}[x^*, y^*]$ as a candidate of I_{t+1}^{2D}, where

$$x^*, y^* = \arg\min_{x,y} \left\| C_t^{[\tau]}[x, y] \right\|_1 . \tag{2}$$

In Eq. (2), $\|C_t^{[\tau]}[x, y]\|_1$ denotes the L_1-norm of the y-th 2D slice in the x-th axis of $C_t^{[\tau]}$, which implies that we select a 2D slice if the number of under-confident pixels is greater than the others. For a better slice recommendation, we skip the slices adjacent to the annotated slices and apply the edge confidence enhancement [10] and LogitNorm [28] techniques for robust construction of C_t.

3D mask post-processing. At each iteration of SnC, users complete 2D interactive segmentation and then acquire a 3D segmentation prediction. In addition to the 2D model-assisted click annotation, our pipeline is able to allow users to refine 2D segmentation via manual annotation tools (*e.g.*, brush or polygon tools) before feeding the 2D results to f^{3D}. Although it is necessary to maintain the regions where the users put their effort to modify, 3D results of f^{3D} may not reflect the 2D mask completed at each iteration.

To alleviate the problem, we propose a *post-processing technique* which enhances the alignment between M_t^{3D} and the user-annotated 2D masks in $M_t^{2D\to3D}$. Let $M_t^{3D}[x, y]$ be the y-th 2D slice of x-th axis, where $x \in \{0, 1, 2\}$. For each $M_n^{2D} \in \{M_i^{2D}\}_{i=1}^t$ and the corresponding slice of M_t^{3D}, or $M_t^{3D}[x_n, y_n]$, one can define

$$\phi_n(M_t^{3D}[x_n, y_n], \alpha) \approx M_n^{2D}, \tag{3}$$

where ϕ_n is an arbitrary operation that can approximately transform $M_t^{3D}[x_n, y_n]$ into M_t^{3D} and α is a parameter controlling the magnitude of operation. In our method, we implement the operation via optical flow estimation and warping operation, where vector fields required for optical flow registration can be obtained by the TV-L1 solver [27, 32]. Employing ϕ_n, we replace $M_t^{3D}[x_n, y_n]$ and its neighboring slices with $\phi_n(M_t^{3D}[x_n, y_n + k], \max(1 - \gamma|k|, 0))$, where $\gamma \in [0, 1]$ is a hyperparameter. Note that it is not compulsory to utilize the post-processing technique on every mask within the set $\{M_i^{2D}\}_{i=1}^t$. To prevent accuracy degradation, we only use sufficiently accurate 2D masks. For a better understanding, we illustrate the post-processing in Figure 3.

Methods	Models	MSD						KiTS19	
		Lung	Colon	Atrium	Liver	Pancreas	Brain	Kidney	Tumor
Non-Interactive	3D nnU-Net [7]	0.689	0.580	0.929	0.951	0.802	0.680	0.969	0.857
Interactive Segmentation	DIOS [29]	0.746	0.721	0.910	0.952	0.814	0.876	0.959	0.867
	f-BRS [22]	0.753	0.743	0.921	0.948	0.833	0.890	0.965	0.870
	FCA [11]	0.786	0.740	0.927	0.956	0.830	0.899	0.968	0.881
	DeepIGeoS [25]	0.775	0.742	0.918	0.950	0.837	0.875	0.967	0.889
	MIDeepSeg [15]	0.793	0.763	0.934	0.961	0.852	0.901	**0.971**	0.883
	RITM [23]	0.807	0.757	0.930	0.959	0.865	0.898	0.962	0.872
	VMN [34]	0.815	0.798	0.937	0.963	0.868	0.905	0.969	0.889
	SnC (Ours)	**0.849**	**0.817**	**0.945**	**0.971**	**0.882**	**0.907**	0.969	**0.892**

Table 1. Comparison between our SnC method and the other volumetric segmentation methods (including automatic and interactive methods). By following the evaluation protocol of [34], we reported the results of SnC after 6 iterations (annotating 6 slices).

4. Experiments

4.1. Training Details

For f^{2D}, we used the HRNet-18 backbone [26] as in [23], where the backbone of f^{3D} was constructed by applying the Mean-ACS convolution [31] to f^{2D}. Also, we employed the pre-trained weight of [23] and our trained weight of f^{2D} as the initial states of f^{2D} and f^{3D}, respectively. At training time, we used the click simulation strategies of [13, 29] and the iterative training scheme of [23], where the clicks were transformed into disk encodings of 5-pixel radius.

We employed the AdamWR optimizer [14] for optimization and the cosine learning rate scheduling. With an initial learning rate of 0.0003, f^{2D} and f^{3D} were trained for 90 and 120 epochs, where the restart scheme was applied at 30 and 40 epochs, respectively. At training time, the batch sizes of f^{2D} and f^{3D} were set to 32 and 2, respectively.

In addition, we applied the principle of curriculum learning [1] after restarting the learning rate (after 30 and 40 epochs for f^{2D} and f^{3D}, respectively) to enhance the robustness of segmentation models. For the 2D model f^{2D} employing positive and negative clicks, we decreased the number of simulated user clicks as the epoch increased. In a similar manner, we controlled the number of 2D mask inputs for f^{3D}, where the mask inputs included not only the ground truth masks but also the output predictions of f^{2D}.

4.2. Datasets

In our experiments, we employed the medical segmentation decathlon (MSD) dataset [21] and the dataset released for 2019 kidney tumor segmentation challenge (KiTS19). Using the MSD dataset, we conducted segmentation of lung tumor, colon cancer, left atrium, liver, pancreas, and brain tumor, where the subsets of the MSD dataset consist of 3D scans and the corresponding segmentation masks. In addition, we performed kidney and kidney tumor segmentation with the KiTS19 dataset, which includes arterial phase abdominal CT scans and the corresponding masks.

# clicks	MSD-lung		MSD-colon	
	VMN	Ours	VMN	Ours
1	0.769	0.815	0.715	0.719
2	0.780	0.820	0.720	0.750
3	0.784	0.835	0.752	0.764
4	0.798	0.840	0.771	0.793
5	0.807	0.844	0.782	0.807
6	0.815	0.849	0.798	0.817
7	0.821	0.852	0.805	0.823
8	0.830	0.856	0.811	0.833
9	0.834	0.860	0.820	0.840
10	0.839	0.862	0.832	0.843

Table 2. Segmentation DSC measures of our proposed method and VMN [34] with respect to the number of slices.

4.3. Compared Models

As a non-interactive segmentation baseline, nnU-Net [7] was employed. In addition, we tested 3D versions of interactive segmentation methods for 2D images (DIOS [29], FCA [11], f-BRS [22], and RITM [23]) by converting the 2D convolutions via ACS ones [31]. Volumetric segmentation models developed for medical images, DeepIGeoS [25] and MIDeepSeg [15] were also used as baseline methods. In addition to the volumetric propagation methods, we compared SnC with a slice propagation baseline VMN [34].

4.4. Evaluation Protocol

To measure the segmentation accuracy, we employed the dice similarity coefficient (DSC) which is computed by

$$\mathrm{DSC} = \frac{2\|M_a \cap M_b\|_1}{\|M_a\|_1 + \|M_b\|_1}, \tag{4}$$

where M_a and M_b are a ground truth segmentation mask and the corresponding mask prediction, respectively.

In [34], the authors reported volumetric interactive segmentation results by annotating 6 slices. Following the protocol, we reported DSC by using the same number of 2D

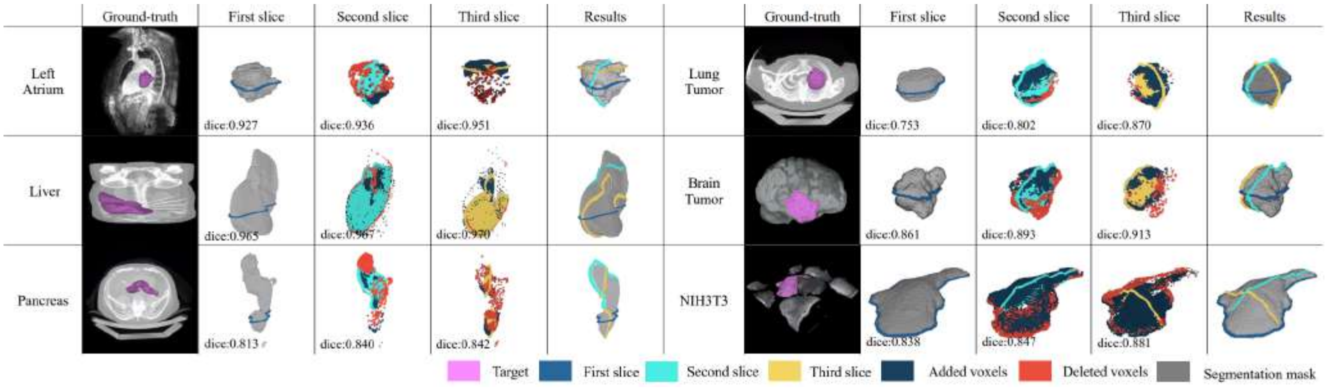


Figure 4. The input image and the ground-truth mask (1st column). Initial prediction obtained from the first slice (2nd column). Added voxels and deleted voxels to the predicted mask after adding the second slice (3rd column) and the third slice (4th column). The segmentation results obtained from three slices (5th column).

Method	Model	MSD-lung	MSD-colon
Interactive Refinement	InterCNN [2]	0.793	0.689
	DeepIGeoS [25]	0.814	0.703
	MIDeepSeg [15]	0.821	0.721
	RITM [23]	0.827	0.717
	TIS [13]	0.849	0.740
	SnC (Ours)	**0.855**	**0.805**

Table 3. Comparison between our SnC pipeline and the previous approaches, where the models were required to refine the existing 3D nnU-Net mask predictions of the MSD-lung and MSD-colon dataset. By following the evaluation protocol of [13], we reported the results of the SnC method when applying 10 clicks.

slices in SnC. For the first slice to annotate, we randomly selected a 2D slice adjacent to the largest foreground region in its ground truth mask. In addition, we finished the first stage of SnC (2D interactive segmentation) when the 2D mask prediction is sufficiently accurate (*e.g.,* DSC > 0.9) or the number of used clicks is 5. At the t-th SnC iteration, our post-processing method of Eq. (3) was applied to the mask prediction of f^{3D} by using sufficiently accurate ones (*e.g.,* DSC > 0.9) among the t accumulated masks. For the other interactive baselines, we used the same number of interactions provided in SnC for each data sample.

We reported each segmentation result of SnC based on the resolution of each pre-processed image, which was automatically given by the nnU-Net protocol [7].

5. Experimental Results

5.1. Quantitative Results

In Table 1, we compared the segmentation results of our SnC approach to those of the previous methods described in Section 4.3. The results present that our proposed SnC pipeline is able to construct more accurate volumetric segmentation results in comparison with previous automatic (non-interactive) and interactive segmentation methods. Especially in the MSD datasets, which have a significantly fewer number of images than KiTS, our method showed robust performance of volumetric interactive segmentation.

The table also implies that two-stage approaches, which first annotate 2D slices and then propagate them to volumetric space, can achieve a higher segmentation accuracy than single-stage volumetric propagation techniques. To emphasize the robustness of our framework, we compared the segmentation results of SnC to those of VMN [34], the state-of-the-art slice propagation method, with respect to the number of slices, where the results are presented in Table 2.

5.2. Visual Demonstration

For a better understanding of our SnC pipeline, we illustrated interactive segmentation processes in Figure 4 by selecting a data item from each dataset. In addition to MSD, we also visualized results for a NIH3T3 [20] data sample to show the effectiveness of SnC in multiple-instance cases.

In the 1st column, we visualize the input image and the ground-truth mask. The 2nd column shows the 3D segmentation masks in the first iteration of SnC. The 3rd and 4th columns show the updated voxels in the second and third iterations, respectively. 3D segmentation accuracy drastically increased when additional 2D mask guidance is added.

5.3. Additional Results

Refinement of existing masks. In [13], Liu *et al.* demonstrated interactive refinement results of existing masks with respect to the number of click interactions. The interactive refinement results of the previous methods including InterCNN [2], DeepIGeoS [25], MIDeepSeg [15], RITM [23], and TIS [13] were reported. Starting from the most erroneous slice in each existing mask, our SnC method was trained to exploit via positive and negative clicks.

Model	MSD-lung				MSD-colon			
	10%	20%	50%	100%	10%	20%	50%	100%
3D nnU-Net [7]	0.267	0.430	0.655	0.689	0.334	0.482	0.565	0.580
RITM [23]	0.615	0.690	0.744	0.807	0.650	0.704	0.727	0.757
VMN [34]	0.689	0.740	0.802	0.815	0.627	0.714	0.773	0.798
SnC (Ours)	**0.784**	**0.802**	**0.823**	**0.849**	**0.749**	**0.774**	**0.796**	**0.817**

Table 4. In the setting of interactive segmentation from scratch, we compare our SnC method and the other volumetric segmentation methods in low-data regime (using the random sampled subsets of the MSD-lung and the MSD-colon datasets).

	Transfer.	Mask attn.	Post-proc.	MSD-lung
(1)	-	-	-	0.816
(2)	✓	-	-	0.825
(3)	✓	✓	-	0.833
(4)	✓	✓	✓	0.849

Table 5. Ablation study results by adding the sub-components of the SnC pipeline. We used the MSD-lung dataset for ablation.

Table 3 presents a comparison between the refinement results of our proposed approach and the previous methods, where the refinement initiates from the existing 3D nnU-Net mask predictions for each dataset. For SnC, we first selected a 2D slice that has the largest erroneous region. When its 2D refinement is satisfactory (*e.g.,* DSC > 0.9), SnC can start its next iteration. The table implies that SnC can also be effective when refining existing segmentation masks. In other words, one can utilize our pipeline to effectively refine mis-segmented 3D masks given by automatic segmentation models such as nnU-Net [7].

Low-data regime. As we mentioned, ML-based biomedical image analysis algorithms usually suffer from data deficiency problem. To demonstrate the segmentation robustness of SnC in such a low-data regime, we adopted the experiments of [34], which sub-sampled 10%, 20%, and 50% of the MSD training data uniformly at random. By using the experiment protocol of Table 1, we compared our method to 3D nnU-Net (non-interactive) [7], 3D RITM (volumetric propagation) [23], and VMN (slice propagation) [34].

As shown in Table 4, our proposed SnC method maintains sufficiently high segmentation accuracy even when the number of training images is significantly reduced. Such results imply that our methods can take benefits from the post-processing step based on 2D mask predictions as well as the transferred knowledge and 2D mask guidance of f^{2D}.

Runtime analysis. We observed that a single SnC iteration takes less than 2.5 s on a NVIDIA RTX2080Ti GPU, while it was reported that VMN [34] and an interactive version of 3D nnU-Net [7] costs more than 5 s and 50 s, respectively. Since SnC requires only a single forward pass for volumetric segmentation, which does not require multiple network inferences for slice propagation and replaces 3D convolutions with efficient ACS convolutions [31], our proposed method can achieve high accuracy with less runtime, thus supporting more flexible real-time interaction.

5.4. Ablation Study

To analyze the effect of the sub-components in our SnC pipeline, we conducted an ablation study by using the MSD-lung dataset. Table 5 shows that the transfer learning (transfer.) technique, which employs the trained weight of f^{2D} as the initial state to train f^{3D}, and the mask attention module (mask attn.) can enhance the segmentation performance. In addition, the table shows that our post-processing method (post-proc.) that enhances the alignment between 2D mask predictions of the first stage and 3D segmentation results, can be effective in improving 3D segmentation accuracy.

6. Conclusion

This paper proposes a novel two-stage approach for interactive 3D image segmentation called Slice-and-Conquer (SnC). The 2D interactive segmentation stage not only enables users to concentrate on a single slice at each iteration but also provides dense hints for 3D image segmentation. By empowering the sparse user-provided point interactions with the 2D mask predictions, the 3D segmentation module constructs a 3D mask based on the guidance. To further enhance the flexibility in iterative refinement, SnC automatically recommends 2D slices to annotate, *i.e.,* the users are only required to consecutively annotate a few recommended slices for 3D image segmentation. Our extensive experimental results show that the proposed approach outperforms the previous methods in efficiency and effectiveness.

Acknowledgements. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)), National Research Foundation of Korea (NRF) funded by the Korea government (MSIT) (2021R1C1C1009256), the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare, Korea (HI21C0977), and a grant from Korea University (K2304791).

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 6

[2] Gustav Bredell, Christine Tanner, and Ender Konukoglu. Iterative interaction training for segmentation editing networks. In *International workshop on machine learning in medical imaging*, pages 363–370. Springer, 2018. 7

[3] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5559–5568, 2021. 2

[4] Karol Gotkowski, Camila Gonzalez, Isabel Jasmin Kaltenborn, Ricarda Fischbach, Andreas Bucher, and Anirban Mukhopadhyay. i3deep: Efficient 3d interactive segmentation with the nnu-net. In *Medical Imaging with Deep Learning*, 2021. 5

[5] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 5

[6] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Guided interactive video object segmentation using reliability-based attention maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7322–7330, 2021. 2

[7] 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. 6, 7, 8

[8] Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5297–5306, 2019. 1, 2

[9] Ron Kikinis, Steve D Pieper, and Kirby G Vosburgh. 3d slicer: a platform for subject-specific image analysis, visualization, and clinical support. In *Intraoperative imaging and image-guided therapy*, pages 277–289. Springer, 2014. 4

[10] Alexandros Kouris, Stylianos I Venieris, Stefanos Laskaridis, and Nicholas Lane. Multi-exit semantic segmentation networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 330–349. Springer, 2022. 5

[11] Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, and Shao-Ping Lu. Interactive image segmentation with first click attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13339–13348, 2020. 2, 5, 6

[12] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017. 1

[13] Wentao Liu, Chaofan Ma, Yuhuan Yang, Weidi Xie, and Ya Zhang. Transforming the interactive segmentation for medical imaging. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV*, pages 704–713. Springer, 2022. 6, 7

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[15] Xiangde Luo, Guotai Wang, Tao Song, Jingyang Zhang, Michael Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Medical image analysis*, 72:102102, 2021. 2, 6, 7

[16] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively trained interactive segmentation. *arXiv preprint arXiv:1805.04398*, 2018. 1, 2

[17] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 616–625, 2018. 4

[18] John R Mayo. Ct evaluation of diffuse infiltrative lung disease: dose considerations and optimal technique. *Journal of thoracic imaging*, 24(4):252–259, 2009. 4

[19] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 1

[20] YongKeun Park, Christian Depeursinge, and Gabriel Popescu. Quantitative phase imaging in biomedicine. *Nature photonics*, 12(10):578–589, 2018. 7

[21] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 6

[22] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. F-brs: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8623–8632, 2020. 1, 2, 6

[23] Konstantin Sofiiuk, Ilia A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. *arXiv preprint arXiv:2102.06583*, 2021. 1, 2, 3, 4, 6, 7, 8

[24] Guotai Wang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Uncertainty-guided efficient interactive refinement of fetal brain segmentation from stacks of mri slices. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 279–288. Springer, 2020. 5

[25] Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David,

Jan Deprest, Sébastien Ourselin, et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1559–1572, 2018. 2, 6, 7

[26] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 6

[27] Andreas Wedel, Thomas Pock, Christopher Zach, Horst Bischof, and Daniel Cremers. An improved algorithm for tv-l 1 optical flow. In *Statistical and geometrical approaches to visual motion analysis*, pages 23–45. Springer, 2009. 5

[28] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, pages 23631–23644. PMLR, 2022. 5

[29] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 373–381, 2016. 1, 2, 3, 6

[30] Jiancheng Yang, Yi He, Kaiming Kuang, Zudi Lin, Hanspeter Pfister, and Bingbing Ni. Asymmetric 3d context fusion for universal lesion detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 571–580. Springer, 2021. 4

[31] Jiancheng Yang, Xiaoyang Huang, Yi He, Jingwei Xu, Canqian Yang, Guozheng Xu, and Bingbing Ni. Reinventing 2d convolutions for 3d images. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3009–3018, 2021. 2, 4, 6, 8

[32] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 5

[33] Tianfei Zhou, Liulei Li, Gustav Bredell, Jianwu Li, and Ender Konukoglu. Quality-aware memory network for interactive volumetric image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 560–570. Springer, 2021. 2, 5

[34] Tianfei Zhou, Liulei Li, Gustav Bredell, Jianwu Li, Jan Unkelbach, and Ender Konukoglu. Volumetric memory network for interactive medical image segmentation. *Medical Image Analysis*, 83:102599, 2023. 2, 5, 6, 7, 8