Medical Image Analysis 83 (2023) 102599

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Volumetric memory network for interactive medical image segmentation

Tianfei Zhou [a,*], Liulei Li [b], Gustav Bredell [a], Jianwu Li [b], Jan Unkelbach [c], Ender Konukoglu [a]

[a] Computer Vision Laboratory, ETH Zurich, Switzerland
[b] School of Computer Science and Technology, Beijing Institute of Technology, China
[c] Department of Radiation Oncology, University Hospital of Zurich, Zurich, Switzerland

ARTICLE INFO

Keywords:
Interactive image segmentation
Memory-augmented network
Attention
fully convolutional network
Deep learning

ABSTRACT

Despite recent progress of automatic medical image segmentation techniques, fully automatic results usually fail to meet clinically acceptable accuracy, thus typically require further refinement. To this end, we propose a novel Volumetric Memory Network, dubbed as VMN, to enable segmentation of 3D medical images in an interactive manner. Provided by user hints on an arbitrary slice, a 2D interaction network is firstly employed to produce an initial 2D segmentation for the chosen slice. Then, the VMN propagates the initial segmentation mask bidirectionally to all slices of the entire volume. Subsequent refinement based on additional user guidance on other slices can be incorporated in the same manner. To facilitate smooth human-in-the-loop segmentation, a quality assessment module is introduced to suggest the next slice for interaction based on the segmentation quality of each slice produced in the previous round. Our VMN demonstrates two distinctive features: **First**, the memory-augmented network design offers our model the ability to quickly encode past segmentation information, which will be retrieved later for the segmentation of other slices; **Second**, the quality assessment module enables the model to directly estimate the quality of each segmentation prediction, which allows for an active learning paradigm where users preferentially label the lowest-quality slice for multi-round refinement. The proposed network leads to a robust interactive segmentation engine, which can generalize well to various types of user annotations (*e.g.*, `scribble`, `bounding box`, `extreme clicking`). Extensive experiments have been conducted on three public medical image segmentation datasets (*i.e.*, MSD, $KiTS_{19}$, CVC-ClinicDB), and the results clearly confirm the superiority of our approach in comparison with state-of-the-art segmentation models. The code is made publicly available at https://github.com/0liliulei/Mem3D.

1. Introduction

Accurate segmentation of organs or lesions from medical imaging data (*e.g.*, CT, MRI) holds the promise of significant improvement of clinical treatment, by allowing the extraction of accurate models for visualization, quantification or simulation (Pham et al., 2000). The traditional naive manual delineation is extremely inefficient for 3D medical images and its performance highly depends on the physician's experience. Benefiting from the recent advancement of deep neural networks (DNNs), deep learning based automated segmentation systems, including convolutional neural networks (CNNs)-based (Ronneberger et al., 2015; Zhou et al., 2018; Milletari et al., 2016; Hesamian et al., 2019; Zhou et al., 2022; Baumgartner et al., 2019) as well as more recent Transformer-based (Hatamizadeh et al., 2022; Cao et al., 2021; Chang et al., 2021; Shamshad et al., 2022), have achieved vast attention and remarkably advanced the segmentation performance. However, automatic segmentation methods have not demonstrated sufficiently accurate and robust results for clinical purposes due to the inherent challenges of medical images, such as low tissue contrast, highly variable and irregular shapes of segmentation targets, diverse imaging and segmentation protocols, and variations across patients. Consequently, interactive segmentation (Olabarriaga and Smeulders, 2001; Zhao and Xie, 2013; Zhou et al., 2017; Bredell et al., 2018; Wang et al., 2018b,a; Zhou et al., 2021) garners research interests of the medical image analysis community, and recently became the choice in many real-life medical applications.

In interactive segmentation, the user is factored in to play a crucial role in guiding the segmentation process and in correcting errors as they occur (often in an iteratively-refined manner). Research on this topic dates back decades, with early efforts focusing on boundary tracing techniques for natural image segmentation (Kass et al., 1988; Mortensen and Barrett, 1995). For medical imaging segmentation, pioneering approaches treat the task as an restricted optimization problem, which can be solved by max-flow (Boykov and Jolly, 2001), geodesic energy minimization (Criminisi et al., 2008) or random walks (Grady

* Corresponding author.
E-mail address: tianfei.zhou@vision.ee.ethz.ch (T. Zhou).

https://doi.org/10.1016/j.media.2022.102599
Received 15 March 2022; Received in revised form 23 June 2022; Accepted 24 August 2022
Available online 6 September 2022

et al., 2005; Grady, 2006). Yet, these methods require a large amount of input from users to segment targets with low contrast and ambiguous boundaries. With the advent of deep learning, there has been a dramatically increasing interest in deep interactive segmentation. Recent methods demonstrate higher segmentation accuracy with fewer user interactions than classical approaches. Despite this, current methods have two major limitations. **(1)** Many approaches (Kitrungrotsakul et al., 2020; Sun et al., 2018; Sakinis et al., 2019) only focus on 2D medical images, not allowing the integration of prior volumetric knowledge regarding the 3D medical data. Slice-by-slice interactive segmentation also imposes extremely high annotation cost to users. Though some 3D networks (Çiçek et al., 2016; Rajchl et al., 2016; Liao et al., 2020; Wang et al., 2018b,a) are able to exploit high-order, volumetric features to segment voxels at a time, they require significantly more parameters and computations in comparison with the 2D counterparts. This necessitates compromises in the 3D network design to fit into a given memory or computation budget. **(2)** These methods are in practice not flexible for human-in-the-loop segmentation, since they requires the user to manually inspect mis-segmented slices in order to give additional interventions for refinement.

To address these issues, we propose a volumetric memory network (VMN) to solve volumetric medical image segmentation as a memory-based reasoning problem. Fundamental to our model is an external memory component, which allows the model to store historical target information in segmented slices in the memory and later retrieve useful representations from the memory as guidance to segment the incoming slice. In this way, our model makes full use of context within 3D data, and at the same time, avoids computationally expensive 3D operations. During segmentation, we dynamically update the memory to maintain shape or appearance variations of the target, facilitating easy model updating without expensive parameter optimization. This solves **limitation 1**. In addition to predicting the segmentation based on the user's inputs, VMN is equipped with a quality assessment component to estimate a confidence for each segmentation result, allowing automatic identification of mis-segmented slices for user interactions. The quality assessment component is lightweight, bringing negligible burden to VMN in model size or inference speed, meanwhile, it is demonstrated to be a reliable selection criterion to support efficient interactive segmentation (see Table 7). In this manner, our VMN tackles the **limitation 2**.

Based on VMN, we propose a novel interactive segmentation engine, running in a round-based workflow. In each round, the engine processes the input image within three steps:

1. *Initialization:* the physician provides guidance on an arbitrary slice, according to which a 2D interaction segmentation network is employed to produce an initial 2D segmentation of the specified target.
2. *Segmentation:* VMN propagates the initial mask sequentially and bidirectionally to the entire volume, and at the same time, predicts a segmentation quality for each slice.
3. *Correction:* the slice with the lowest quality score is fetched, and the physician provides extra corrections on it for next-round refinement (loop to step 1), or stop segmentation if the result is already satisfactory.

The key contributions of this paper are as follows:

1. We propose a novel memory-augmented network named VMN for interactive segmentation of volumetric medical data. It solves the task by sequential label propagation, while taking into consideration the rich 3D structures, and avoiding expensive 3D operations.
2. We equip the memory network with a quality assessment component to assess the quality of each segmentation. It facilitates automatic suggestion of appropriate slices for iterative correction by involving human intervention in the loop. This self-assessment strategy greatly promotes the practical utility of our VMN so that it enables smooth and efficient interaction segmentation.
3. Our approach outperforms previous methods by a significant margin on three public datasets, while being able to handle various forms of interactions (*e.g.*, `scribble`, `bounding box`, `extreme clicking`).

The present work builds upon our conference paper (Zhou et al., 2021) and extends it in some significant aspects. First, we elaborate on more detailed explanations of our VMN in Section 3. Second, we provide a more inclusive review of relevant works in Section 2. Third, we incorporate considerable new experimental results in Section 4, including a more comparative study with recent approaches, new results on CVC-ClinicDB, more ablative experiments, and visualization results.

In the remainder of this paper, we first provide a thorough review of existing interactive segmentation and memory-augmented networks in medical imaging in Section 2. In Section 3, we introduce our VMN network for volumetric image segmentation, and verify it through extensive experiments in Section 4. The paper is concluded in Section 5.

2. Related work

In this section, we discuss recent advances in two relevant fields, *i.e.*, *interactive medical image segmentation* and *memory-aware neural networks*.

2.1. Interactive medical image segmentation

Segmenting targets interactively is a long standing research topic, which shows superiority in producing higher-quality segmentation than fully-automatic methods. It is often a key step in many medical applications, where image segmentation is particularly difficult due to restrictions imposed by image acquisition, pathology and biological variation (Olabarriaga and Smeulders, 2001). User interactions can be supplied in several typical ways such as scribbles (Boykov and Funka-Lea, 2006; Grady et al., 2005; Rother et al., 2004; Wang et al., 2016), bounding boxes (Castrejon et al., 2017), extreme points (Maninis et al., 2018; Agustsson et al., 2019) or point clicks (Sakinis et al., 2019; Koohbanani et al., 2020; Zhang et al., 2021b). Most conventional approaches (Boykov and Funka-Lea, 2006; Grady et al., 2005; Rother et al., 2004; Wang et al., 2016) formulate the task as energy minimization on a regular pixel grid, with unary potential capturing low-level appearance properties and pairwise or higher-order potentials encouraging regular segmentation outputs. In recent years, deep learning based techniques (Rajchl et al., 2016; Çiçek et al., 2016; Sakinis et al., 2019; Liao et al., 2020; Kitrungrotsakul et al., 2020; Wang et al., 2018b,a; Koohbanani et al., 2020; Zhang et al., 2021b) have received considerable attention and significantly boosted segmentation performance. The pioneering DeepCut method (Rajchl et al., 2016) directly replaces the Gaussian mixture model in GrabCut by a CNN for MRI segmentation, while most subsequent approaches (Çiçek et al., 2016; Sakinis et al., 2019; Wang et al., 2018a,b; Luo et al., 2021; Wang et al., 2020) solve the task by 2D or 3D fully convolutional networks (FCNs), with user hints serving as network inputs. To reduce the cost of initial annotations, some methods (Wang et al., 2018b; Bredell et al., 2018; Zhou et al., 2021; Liao et al., 2020) take automatically-segmented masks as network inputs and refine them via neural networks.

Despite the progress, previous methods employ computationally expensive 3D CNNs for interactive volumetric image segmentation, which causes low reaction speed in practice, especially for scenarios requiring multi-round iterative interactions. In addition, most methods require users to manually check the segmentation slice-by-slice and identify inaccurate segments to provide corrections, which is laborious and highly inefficient. To alleviate this, in addition to predicting segmentation based on user's inputs, some studies assess uncertainty of the segmentation as the guidance of follow-up corrections. Conventional

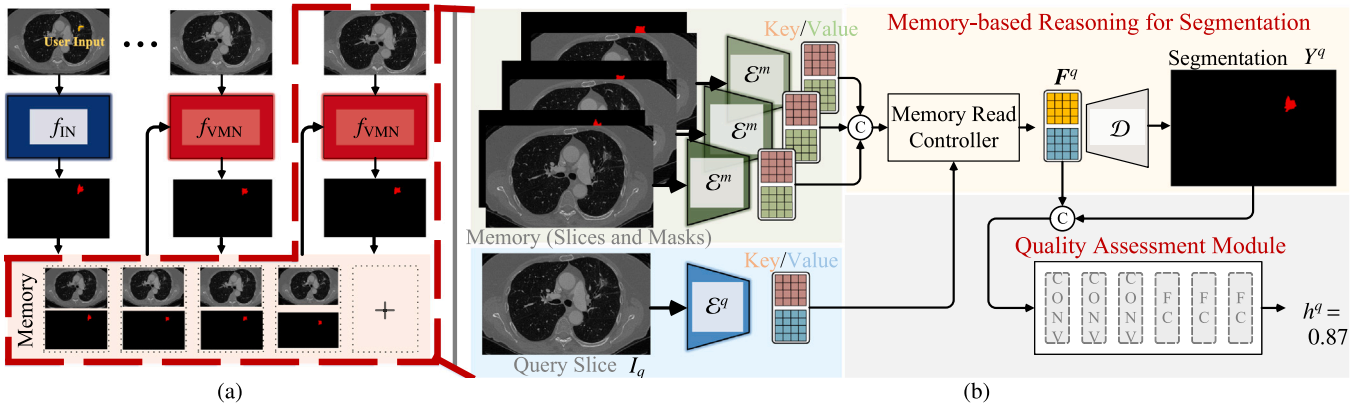

Fig. 1. Illustration of the proposed 3D interactive segmentation engine. (a) Simplified schematization of our engine that solves the task with an 2D interaction network (f_{IN}) and a volumetric memory network (f_{VMN}). (b) Detailed network architecture of f_{VMN}. © denotes concatenation.

algorithms (Baxter et al., 2016; Top et al., 2011) estimate uncertainty of the whole volume by multi-step, iterative optimization, which is very time-consuming. More recently, (Yoo and Kweon, 2019) introduces a loss prediction module for uncertainty estimation within deep neural networks. The method is task-agnostic, however, it relies on elaborate designs of the loss function to account for changes of loss scale during training. Our quality assessment module is superior to Yoo and Kweon (2019) in that it is specifically designed for segmentation; considering that IoU is a normalized metric, we can directly apply a simple MSE loss for optimization.

2.2. Memory-aware neural networks

Memory networks augment neural networks with an external memory component, allowing for the network to explicitly access the past experiences (Sukhbaatar et al., 2015; Kumar et al., 2016; Santoro et al., 2016). They have been shown effective in various tasks such as few-shot learning (Santoro et al., 2016), video tracking (Yang and Chan, 2018), and also been explored to solve reasoning problems in visual dialog (Sukhbaatar et al., 2015; Kumar et al., 2016). In this work, we, for the first time, explore memory-augmented neural networks for solving the problem of 3D medical image segmentation. In contrast to 3D CNN networks that perceive volumetric patterns through 3D convolutions, our network stores volumetric cues within the memory, and for each query slice, produces a memory summarization representation by taking into account the similarity between the query slice and the stored memory. As a result, the memory network has the ability to retrieve volumetric cues for each slice, and thus enables accurate 3D segmentation in a cheap manner. Concurrent to our conference paper (Zhou et al., 2021; Cheng et al., 2021) introduces a memory-augmented neural network for interactive video object segmentation. However, it requires manual selection of frames for user correction and only verifies scribble-guided segmentation.

3. Our approach

3.1. Overview

Let $V \in \mathbb{R}^{h\times w\times c}$ be a volumetric image to be segmented, which has a spatial size of $h \times w$ and c slices. Our approach aims to obtain a 3D binary mask $Y \in \{0,1\}^{h\times w\times c}$ for a specified target by utilizing user guidance. As shown in Fig. 1(a), the physician is asked to provide an initial input on an arbitrary slice $I_i \in \mathbb{R}^{h\times w}$, where I_i denotes the ith slice of V. Then, an interaction network (f_{IN}, Section 3.2) is employed to obtain a coarse 2D segmentation $Y_i \in [0,1]^{h\times w}$ for I_i. Subsequently, Y_i is propagated to all other slices by VMN (f_{VMN}, Section 3.3) to obtain Y. Our approach also takes into account iterative refinement allowing the segmentation performance to be progressively improved with multi-round inference. To aid the refinement, the memory network has a module that estimates the segmentation performance on each slice and suggests the user to place guidance on the slice with the worst segmentation quality.

3.2. Interaction network

The interaction network takes the user annotation at an interactive slice I_i to segment the specified target (or refine the previous result). At the tth round, its input consists of three images: the original gray-scale image I_i, the segmentation mask from the previous round Y_i^{t-1}, and a cue map $M_i \in \{0,1\}^{h\times w}$ that encodes user guidance. Note that in the first round (*i.e.*, $t=0$), the segmentation mask Y_i^{-1} is initialized as a neutral mask with 0.5 for all pixels. These inputs are concatenated along the channel dimension to form an input tensor $\boldsymbol{X}_i^t \in \mathbb{R}^{h\times w\times 3}$. The interaction network f_{IN} conducts the segmentation for I_i as follows:

$$Y_i^t = f_{IN}(\boldsymbol{X}_i^t) \quad \in \mathbb{R}^{h\times w}. \tag{1}$$

To further enhance performance and avoid mistakes in case of small targets or low-contrast tissues, we propose to crop the image according to the rough bounding-box estimation of user input, and apply f_{IN} only to the ROI. We extend the bounding box by 10% along sides to preserve more context. Each ROI region is resized into a fixed size for network input. After segmentation, the mask made within the ROI is inversely warped and pasted back to the original location.

3.3. Volumetric memory network

Given the initial 2D segmentation Y_i^t, our VMN learns from the interactive slice I_i and segments the desired target in other slices. It stores previously segmented slices in an external memory $\mathcal{M}$, and takes into consideration of the stored 3D image and corresponding segmentation masks to improve the segmentation of each 2D slice. The network architecture is shown in Fig. 1(b). In the following paragraphs, the superscript 't' is omitted for conciseness unless necessary.

3.3.1. Key and value embedding

Given a query slice I_q, the network mines useful information from memory $\mathcal{M}$ for segmentation. Here, each memory cell $\mathcal{M}_j \in \mathcal{M}$ is comprised of a slice image I_{n_j} and its segmentation mask Y_{n_j}, where n_j indicates the index of the slice in the original volume. As illustrated in Fig. 1(b), we first encode the query I_q as well as each memory cell $\mathcal{M}_j = \{I_{n_j}, Y_{n_j}\}$ into pairs of *key* and *value* using dedicated encoders (*i.e.*, query $\mathcal{E}^q$ and memory encoder $\mathcal{E}^m$):

$$\boldsymbol{K}^q, \boldsymbol{V}^q = \mathcal{E}^q(I_q), \tag{2}$$

$$\boldsymbol{K}^m_{n_j}, \boldsymbol{V}^m_{n_j} = \mathcal{E}^m(I_{n_j}, Y_{n_j}). \quad (3)$$

Here, $\boldsymbol{K}^q \in \mathbb{R}^{H\times W\times C/8}$ and $\boldsymbol{V}^q \in \mathbb{R}^{H\times W\times C/2}$ indicate key and value embedding of the query I_q, respectively, whereas $\boldsymbol{K}^m_{n_j}$ and $\boldsymbol{V}^m_{n_j}$ correspond to the key and value of the memory cell $\mathcal{M}_j$. H, W and C denote the height, width and channel dimension of the feature map from the backbone network, respectively. Note that for each memory cell, we apply Eq. (3) to obtain key–value embedding pairs. Subsequently, the key and value maps from different memory slices are stacked together to build a pair of 4D key and value features (*i.e.*, $\boldsymbol{K}^m \in \mathbb{R}^{N\times H\times W\times C/8}$ and $\boldsymbol{V}^m \in \mathbb{R}^{N\times H\times W\times C/2}$), where $N = |\mathcal{M}|$ denotes memory size.

3.3.2. Memory reading

The memory read controller retrieves useful information from the memory that is relevant to the current query. To achieve this, we first compute the similarities between all the pixels of the query key map (*i.e.*, $\boldsymbol{K}^q$) and the memory key map ($\boldsymbol{K}^m$). Following the key–value retrieval mechanism in Kumar et al. (2016), Sukhbaatar et al. (2015), the similarity matching is established in a non-local manner by comparing every 3D location $i \in \mathbb{R}^3$ in $\boldsymbol{K}^m$ in with each spatial location $j \in \mathbb{R}^2$ in $\boldsymbol{K}^q$ as follows:

$$s(i,j) = \frac{\boldsymbol{K}^m(i)\cdot\boldsymbol{K}^q(j)}{\|\boldsymbol{K}^m(i)\|\|\boldsymbol{K}^q(j)\|} \quad \in [-1,1], \quad (4)$$

where $\boldsymbol{K}^m(i) \in \mathbb{R}^{C/8}$ and $\boldsymbol{K}^q(j) \in \mathbb{R}^{C/8}$ denote the features at the ith and jth position of $\boldsymbol{K}^m$ and $\boldsymbol{K}^q$, respectively. Next, we compute the read weight w_k by softmax normalization:

$$w(i,j) = \frac{\exp(s(i,j))}{\sum_o \exp(s(o,j))} \quad \in [0,1]. \quad (5)$$

Here, $w(i,j)$ measures the matching probability between i and j. Then, the value of the memory is retrieved by a weighted summation with the soft weights:

$$\boldsymbol{H}^q(j) = \sum_i w(i,j)\boldsymbol{V}^m(i) \quad \in \mathbb{R}^{C/2}, \quad (6)$$

where $\boldsymbol{V}^m(i) \in \mathbb{R}^{C/2}$ denotes the feature of the ith 3D position in $\boldsymbol{V}^m$ and $\boldsymbol{H}^q(j)$ indicates the summarized representation of location j. For all $H \times W$ locations in $\boldsymbol{K}^q$, we independently apply Eq. (6) and obtain the feature map $\boldsymbol{H}^q \in \mathbb{R}^{H\times W\times C/2}$. To achieve a more comprehensive representation, the feature map is concatenated with query value $\boldsymbol{V}^q$ to compute a final query representation $\boldsymbol{F}^q = \mathtt{cat}(\boldsymbol{H}^q, \boldsymbol{V}^q) \in \mathbb{R}^{H\times W\times C}$.

3.3.3. Segmentation readout

Given $\boldsymbol{F}^q$, our VMN leverages a decoder network D to predict the final segmentation probability map for the query slice:

$$Y^q = D(\boldsymbol{F}^q) \quad \in [0,1]^{h\times w}. \quad (7)$$

3.3.4. Quality assessment

While VMN provides a compelling way to produce 3D segmentation, it does not efficiently support human-in-the-loop scenarios. To solve this, we equip the memory network with a lightweight quality assessment head, which computes a quality score for each segmentation mask. In particular, we consider *mean intersection-over-union (mIoU)* as the basic index for quality measurement. For each query I_q, we take its feature representation $\boldsymbol{F}^q$ and the corresponding segmentation mask Y^q together to regress a mIoU score h^q:

$$h^q = \mathcal{Q}(\boldsymbol{F}^q, Y^q) \quad \in [0,1], \quad (8)$$

where $\boldsymbol{Y}^q$ is firstly resized to a size of $H \times W$ and then concatenated with $\boldsymbol{F}^q$ for regression. The slice with the lowest score is curated for next-round interaction. The quality-aware module $\mathcal{Q}(\cdot)$ has two appealing properties: (1) it provides the interactive engine a mechanism to automatically suggest the lowest-quality slice for user refinement; (2) the quality regression loss provides auxiliary supervision signal to guide the learning of VMN.

3.4. Multi-round segmentation

Our VMN performs multi-round interaction to progressively improve the segmentation performance. In particular, in the first round, a user hint as well as the corresponding slice image are provided to the 2D interaction network f_{IN} to predict an initial segmentation mask. Next, VMN (*i.e.*, f_{VMN}) will transfer the segmentation mask bidirectionally to all other slices to produce segmentation predictions and corresponding segmentation quality scores. A new slice with the lowest score will be chosen and the corresponding mask will be provided to users for corrections, which will trigger next-round segmentation.

3.5. Detailed network architecture

3.5.1. Loss function

Our VMN is end-to-end trainable. The training loss is a combination of the segmentation loss and quality regression loss with the same weights:

$$\mathcal{L} = \sum_q \mathcal{L}_{\text{CE}}(Y^q, \hat{Y}^q) + \mathcal{L}_2(h^q, \hat{h}^q), \quad (9)$$

where $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_2$ represents the cross entropy loss and ℓ_2 loss, respectively. $\hat{Y}^q$ is the ground-truth mask of I_q, while $\hat{h}^q$ is equal to the pixel-level IoU between the predicted mask Y^q and its matched ground truth mask $\hat{Y}^q$.

3.5.2. Interaction network

The interaction network f_{IN} is implemented with a cascaded structure as (Chen et al., 2018) to produce the segmentation in a coarse-to-fine manner. First, the inputs are fed into a FPN-like network (Lin et al., 2017) which progressively fuses the high-level semantic information (from the deeper layers) with low-level details (from the earlier layers) via lateral connections to produce informative representations. A pyramid scene parsing module is appended at the deepest layer to gather global contextual information. Second, we apply multi-scale fusion to aggregate the information across different levels in the FPN network as (Chen et al., 2018). Finally, a 1×1 convolutional layer is used to produce the initial segmentation mask. We note that our approach is not limited to this specific interaction network, and other architectures like U-Net (Ronneberger et al., 2015) can also be used instead. The network is trained using a standard cross-entropy loss.

3.5.3. Volumetric memory network

We utilize ResNet-50 (He et al., 2016) as the backbone network for both $\mathcal{E}^q$ (Eq. (2)) and $\mathcal{E}^m$ (Eq. (3)). The `res4` feature map of ResNet-50 is taken for computing the key and value embedding. Note that $\mathcal{E}^q$ and $\mathcal{E}^m$ has the same structure except for the inputs. The input to $\mathcal{E}^q$ is only a slice image, while the input to $\mathcal{E}^m$ consists of a slice image and the corresponding segmentation mask. For D, we first apply Atrous Spatial Pyramid Pooling module after the memory read operation to enlarge the receptive field. We use three parallel dilated convolution layers with dilation rates 2, 4 and 8. Then, the learned feature is decoded with a residual refinement module proposed in Qin et al. (2019). The quality-aware module, $\mathcal{Q}$, consists of three 3×3 convolutional layers and three fully connected layers.

4. Experiment

4.1. Experimental setup

4.1.1. Data

To evaluate the effectiveness of our method, we conduct extensive experiments on three public datasets:

Table 1
Quantitative segmentation results on MSD (Simpson et al., 2019) `test` in terms of DSC (Mean ± Standard Deviation, %). Automatic (non-interactive) approaches are shown in gray. ∗ denotes current best-performed model in the leaderboard of MSD challenge (https://decathlon-10.grand-challenge.org/evaluation/challenge/leaderboard/). † indicates statistically significant results (p-value < 0.05) in comparison with MIDeepSeg. See Section 4.4.1 for details.

Approach	Interaction	Lung cancer	Colon cancer
SNAS (Kim et al., 2019)	No	68.6	N/A
V-NAS (Zhu et al., 2019)	No	55.3	N/A
UMCT (Xia et al., 2020)	No	N/A	56.0
C2FNAS (Yu et al., 2020)	No	70.4	58.9
3D nnU-Net (Isensee et al., 2018)	No	66.9	56.0
*Swin UNETR	No	77.0	59.0
Interactive 3D nnU-Net (Isensee et al., 2018)	`scribble`	73.9 ± 16.8	68.1 ± 34.7
	`bounding box`	74.7 ± 16.3	68.5 ± 33.2
	`extreme clicking`	75.1 ± 15.5	69.8 ± 31.0
UGIR (Wang et al., 2020)	`scribble`	76.0 ± 13.8	71.9 ± 19.6
	`bounding box`	76.5 ± 13.4	72.4 ± 19.5
	`extreme clicking`	76.9 ± 12.8	72.5 ± 19.6
DeepIGeoS (Wang et al., 2018b)	`scribble`	76.6 ± 13.5	72.3 ± 19.5
	`bounding box`	77.2 ± 13.3	73.0 ± 19.2
	`extreme clicking`	77.5 ± 12.6	73.2 ± 19.1
MIDeepSeg (Luo et al., 2021)	`scribble`	78.9 ± 10.1	74.8 ± 12.5
	`bounding box`	79.3 ± 9.9	75.6 ± 12.3
	`extreme clicking`	79.9 ± 9.8	76.0 ± 11.8
VMN (ours)	`scribble`	†80.9 ± 9.2	†79.7 ± 11.7
	`bounding box`	†81.5 ± 9.1	†79.3 ± 11.4
	`extreme clicking`	†**82.0** ± 8.8	†**80.4** ± 11.2

Table 2
Quantitative segmentation results on $KiTS_{19}$ (Heller et al., 2019) `test` in terms of DSC (Mean ± Standard Deviation, %). Automatic (non-interactive) approaches are shown in gray. †: current best-performed model in the leaderboard of $KiTS_{19}$ challenge (https://kits21.kits-challenge.org/results). † indicates statistically significant results (p-value < 0.05) in comparison with MIDeepSeg. See Section 4.4.2 for details.

Approach	Interaction	Kidney organ	Kidney tumor
Mu et al. (Mu et al., 2019)	No	97.2	78.9
MSS U-Net (Zhao et al., 2020)	No	96.9	80.5
Zhang et al. (Zhang et al., 2019)	No	97.4	83.1
Hou et al. (Hou et al., 2019)	No	96.7	84.5
3D nnU-Net (Isensee et al., 2018)	No	96.9	85.7
†3D U-Net (Isensee and Maier-Hein, 2019)	No	97.4	85.1
Interactive 3D nnU-Net (Isensee et al., 2018)	`scribble`	94.5 ± 4.0	86.3 ± 15.9
	`bounding box`	95.3 ± 3.8	86.8 ± 15.8
	`extreme clicking`	95.6 ± 3.1	87.6 ± 14.4
UGIR (Wang et al., 2020)	`scribble`	96.0 ± 3.8	87.1 ± 16.3
	`bounding box`	96.3 ± 3.0	87.8 ± 15.6
	`extreme clicking`	96.7 ± 2.7	88.1 ± 13.5
DeepIGeoS (Wang et al., 2018b)	`scribble`	95.7 ± 3.4	87.6 ± 14.3
	`bounding box`	96.4 ± 2.8	88.5 ± 13.0
	`extreme clicking`	96.7 ± 2.4	88.9 ± 11.4
MIDeepSeg (Luo et al., 2021)	`scribble`	96.3 ± 2.9	87.9 ± 9.2
	`bounding box`	96.6 ± 2.8	88.1 ± 9.0
	`extreme clicking`	**97.1** ± 2.3	88.5 ± 8.8
VMN (ours)	`scribble`	†96.9 ± 1.9	88.2 ± 7.5
	`bounding box`	†97.0 ± 2.1	†88.4 ± 7.5
	`extreme clicking`	97.0 ± 1.7	†**89.1** ± 7.4

- MSD (Simpson et al., 2019) is a large-scale dataset with a total of 2633 3D volumetric images. They are grouped into ten subsets according to the anatomy of interest (*e.g.*, liver, lung, hippocampus, colon). In our experiments, we study the most challenging two subsets: lung (64/32 for `train`/`test`) and colon (126/64 for train/`test`).
- $KiTS_{19}$ (Heller et al., 2019) contains 300 arterial phase abdominal CT scans with annotations of kidney and tumor. We use the released 210 scans in the experiments, which are split into 168 for `train` and 42 for `test`.
- CVC-ClinicDB (Bernal et al., 2015) consists of 29 colonoscopy sequences. We follow (Fan et al., 2020; Jha et al., 2020) to divide them into 23, 3 and 3 for `train`, `val` and `test`, respectively.

4.1.2. Baseline model

For comparison, we build a baseline model, named Interactive 3D nnU-Net, by adapting nnU-Net (Isensee et al., 2018) into the setting of interactive segmentation. In particular, we first leverage the 2D interaction network f_{IN} (Section 3.2) to produce a segmentation for the interactive slice. Then, the segmentation mask is concatenated with the volume to form the input of 3D nnU-Net. The quality-aware iterative refinement is also applied. In addition, we compare our approach against three interactive segmentation methods, *i.e.*, DeepIGeoS (Wang et al., 2018b), UGIR (Wang et al., 2020), and MIDeepSeg (Luo et al., 2021). We also report the performance of several famous automated (*i.e.*, non-interactive) alternatives for reference.

Table 3
Quantitative results on CVC-ClinicDB (Bernal et al., 2015) `test` in terms of mIoU (%) and DSC (%). Automatic (non-interactive) approaches are shown in gray. † indicates statistically significant results (p-value < 0.05) in comparison with Interactive 3D nnU-Net. See Section 4.4.3 for details.

Approach	Interaction	mIoU (%)	DSC (%)
SFA (Fang et al., 2019)	No	N/A	70.0
U-Net++ (Zhou et al., 2018)	No	72.9	79.4
U-Net (Ronneberger et al., 2015)	No	75.5	82.3
ResUNet++ (Jha et al., 2019)	No	79.6	79.6
PraNet (Fan et al., 2020)	No	84.9	89.9
TransFuse-S (Zhang et al., 2021a)	No	86.8	91.8
DoubleUNet (Jha et al., 2020)	No	86.1	92.4
Interactive 3D nnU-Net (Isensee et al., 2018)	`bounding box`	88.1 ± 4.5	93.2 ± 6.5
	`extreme clicking`	88.3 ± 4.1	93.3 ± 6.0
VMN (ours)	`bounding box`	†90.4 ± 3.1	†94.6 ± 4.7
	`extreme clicking`	†**90.7** ± 2.8	†**94.9** ± 4.2

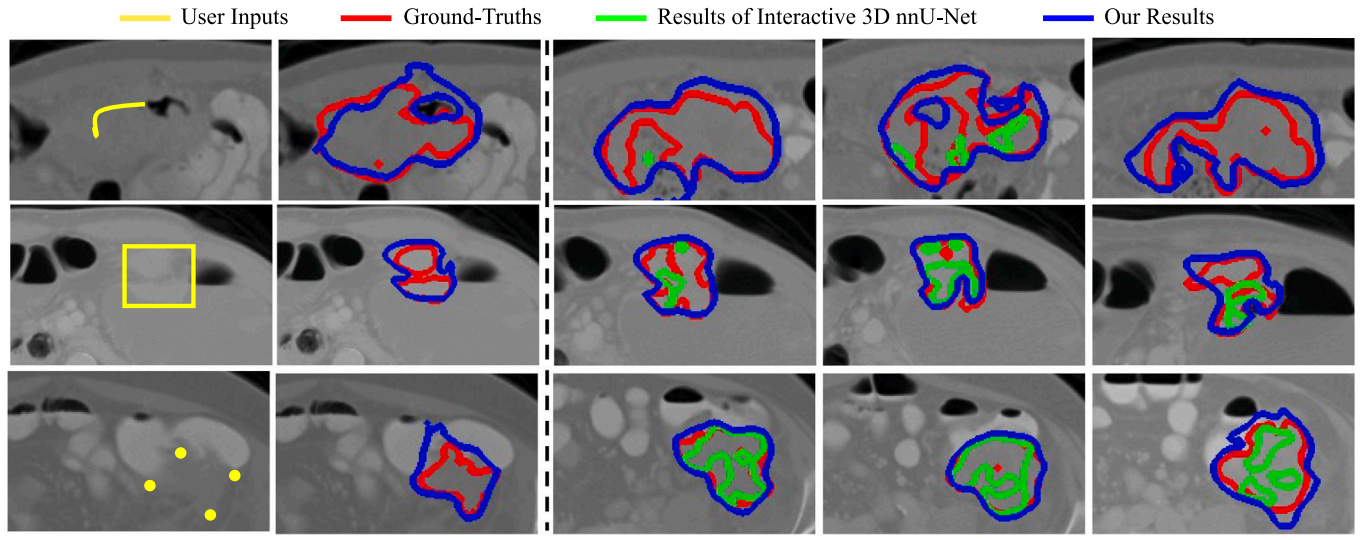

Fig. 2. Qualitative results of our approach vs. Interactive 3D nnU-Net on representative samples in the colon set of MSD (Simpson et al., 2019) `test`. From top to bottom: `scribble`, `bounding box` and `extreme clicking`. From left to right: interactive slices, segmentation results of interactive slices by the interaction network, segmentation results of other three slices by VMN. Note that VMN and Interactive 3D nnU-Net share a same 2D interaction network, thus only one contour is depicted in the second column.

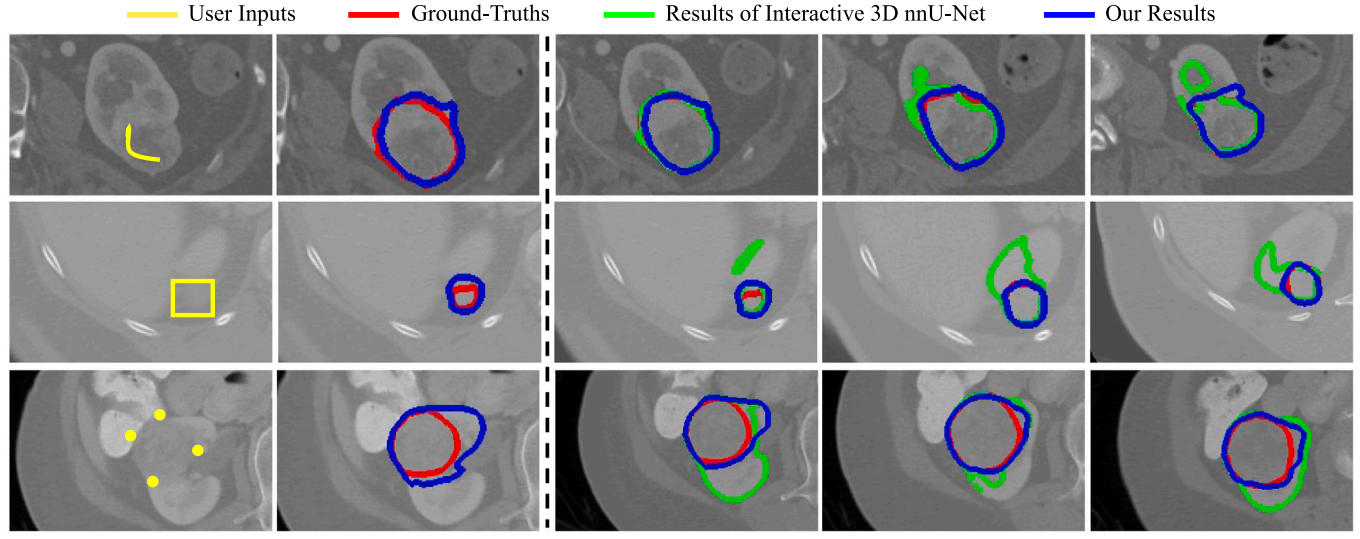

Fig. 3. Qualitative results of our approach vs. Interactive 3D nnU-Net on representative samples of kidney tumor segmentation in $KiTS_{19}$ (Heller et al., 2019) `test`. From top to bottom: `scribble`, `bounding box` and `extreme clicking`. From left to right: interactive slices, segmentation results of interactive slices by the interaction network, segmentation results of other three slices by VMN. Note that VMN and Interactive 3D nnU-Net share a same 2D interaction network, thus only one contour is depicted in the second column.

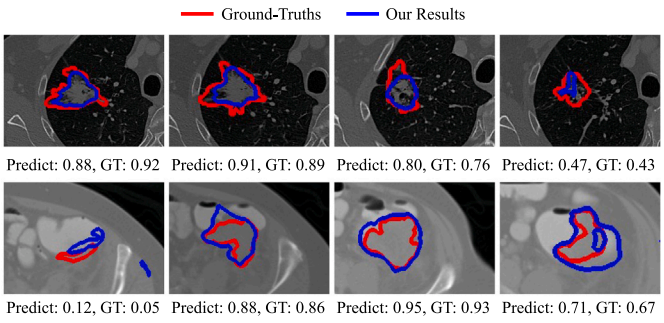

Fig. 4. Qualitative results of quality assessment on the lung (top) and colon (bottom) sets of MSD `test`. For each segmentation, we report the predicted quality score as well as corresponding IOU score computed between the segmentation prediction and ground-truth.

4.1.3. Evaluation protocol

To assess the performance of our approach, we employ the common Dice Similarity coefficient (DSC) as the main metric. It measures the similarity between the predicted volumetric segmentation Y and corresponding ground-truth $\hat{Y}$:

$$DSC = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|}, \tag{10}$$

where $|\cdot|$ represents the operation of cardinality computation, that returns the number of elements in a binary mask.

4.2. Interaction simulation

Our approach can support various types of user interactions, which facilitates diverse uses in clinical routine. In our experiments, we study three common interactions, *i.e.*, `scribble`, `bounding box` and `extreme clicking`. `Scribble` provides sparse labels to describe the targets and rough extent, `bounding box` outlines the sizes and locations of targets, whereas `extreme clicking` (Maninis et al., 2018) outlines a more compact area of a target by labeling its *leftmost*, *rightmost*, *top*, *bottom* pixels. To simulate scribbles, we manually label the data in MSD and KiTS_{19}, resulting in 3,585 slices. We will make the annotations public available. Bounding boxes and extreme clicks can be easily simulated from ground-truths. To mimic real users' behaviors that may not provide precise annotations, we randomly jitter the position of each extreme click up to 10 pixels. This also applies to corners of bounding boxes. Given user's annotation, we employ geodesic distance transformation in DeepIGeoS (Wang et al., 2018b) to encode the input into a cue map M, which serves as the input of f_{IN} (Section 3.2). Only axial slices are employed for interaction in our experiments. To maximize the performance, we train an independent f_{IN} for each interaction type.

4.3. Implementation details

Training. Our engine is implemented in PyTorch and trained using four Geforce RTX 2080Ti GPUs, each with a 11 GB memory. For f_{IN} (Section 3.2), we follow the setting in Zhang et al. (2020) for training. It is trained for a total of 100 epochs with batch size 10 and learning rate 1e-6. The volumetric memory network f_{VMN} (Section 3.3) is trained using Adam with learning rate 1e-5 and batch size 8 for 120 epochs. To create a training example, we randomly sample 5 ordered slices from a 3D image. During training, the memory is dynamically updated by adding the slice and mask of the previous step to the memory as support for the next slice. All these training settings are determined via 5-fold cross-validation on KiTS_{19} `train`, and subsequently applied to other datasets (*i.e.*, MSD and CVC-ClinicDB).

Testing. During inference, simulated user hints are provided to f_{IN} for an initial segmentation of the interactive slice. Then, for each query slice, we put this interactive slice and the previous slice with corresponding segmentation masks into the memory as the most important reference information. In addition, we save a new memory item every S slices for each segmentation direction independently, where S is empirically set to 5. We do not add all slices and corresponding masks into memory to avoid large storage and computational costs. In this way, our memory network achieves the effect of online learning and adaption without additional training.

Table 4
Quantitative comparison of generalization ability of VMN across different datasets. DSC (%) is used as the metric. See Section 4.4.4.

Train set	Test set	`scribble`	`bounding box`	`extreme clicking`
KiTS_{19}	MSD (Lung)	77.8	78.2	78.6
	MSD (Colon)	77.2	77.5	77.9
MSD	KiTS_{19} (Organ)	89.6	90.3	90.7
	KiTS_{19} (Tumor)	86.1	86.4	86.5

4.4. Main results

4.4.1. Segmentation performance on MSD

Table 1 provides segmentation results of lung cancer and colon cancer on MSD `test`. For interactive models, we report scores at the 6-th round which well balances segmentation accuracy and model efficiency. **First**, we find that segmentation of lung cancer and colon cancer is highly challenging. The best-performed automatic model, *i.e.*, Swin UNETR, only produces DSC scores of 77.0% for lung cancer and 59.0% for colon cancer, respectively. The performance cannot meet the requirements of clinical practice. **Second**, our VMN, working in an interactive manner, significantly improves the segmentation performance against automatic models. With `extreme clicking`, VMN outperforms Swin UNETR by **15%** for lung cancer and **21.4%** for colon cancer. **Third**, we observe a significant improvement of VMN against the four interactive competitors, *i.e.*, Interactive 3D nnU-Net (Isensee et al., 2018), DeepIGeoS (Wang et al., 2018b), UGIR (Wang et al., 2020), and MIDeepSeg (Luo et al., 2021), which is consistent across interaction types (*i.e.*, `scribble`, `bounding box` and `extreme clicking`). In particular, VMN outperforms Interactive 3D nnU-Net (Isensee et al., 2018) by large margins, *i.e.*, more than **7%** for lung cancer and **10%** for colon cancer on average. In comparison with the second best MIDeepSeg (Luo et al., 2021), VMN also establishes promising gains of more than **2%** and **4%** for Lung Cancer and Colon Cancer, respectively. **Last**, for the three interaction types, *i.e.*, `scribble`, `bounding box` and `extreme clicking`, VMN delivers very similar segmentation performance, demonstrating its high robustness to user inputs.

4.4.2. Segmentation performance on $KiTS_{19}$

Table 2 presents performance comparisons on KiTS_{19} `test` for kindey organ and tumor segmentation. As seen, the best-performed automatic model (Isensee and Maier-Hein, 2019) has already demonstrated compelling performance for segmentation of kidney organ, even better than interactive models. However, automatic models still encounter difficulties in kidney tumor segmentation. Moreover, all the five interactive segmentation models deliver more precise segmentation of kidney tumor than (Isensee and Maier-Hein, 2019). Among them, our VMN yields the best overall performance, with the smallest standard deviations across three interactive types.

4.4.3. Segmentation performance on CVC-ClinicDB

We now assess the performance of VMN against eight automatic and one interactive competitors on CVC-ClinicDB `test`. Following the protocol of the dataset, we evaluate the approaches in terms of both mIoU and DSC scores. As summarized in Table 3, our VMN with `extreme clicking` yields the best performance. It outperforms the

Table 5
Ablation study on memory size in terms of DSC (%). See Section 4.5.1 for details.

Memory size	Lung cancer			Colon cancer		
	scribble	bounding box	extreme clicking	scribble	bounding box	extreme clicking
0	58.2	59.3	58.9	54.7	54.7	54.8
1	76.2	75.6	77.0	67.3	67.1	68.0
5	79.6	79.8	80.9	72.9	73.1	73.9
10	80.9	81.4	81.8	75.2	75.3	75.8
15	**81.0**	**81.5**	**82.1**	78.7	79.2	**80.4**
20	80.9	**81.5**	82.0	**79.7**	**79.3**	**80.4**

Table 6
Ablation study on 2D interaction network f_{IN} in terms of DSC (%). See Section 4.5.2 for details.

2D Interaction network	Lung cancer			Colon cancer		
	scribble	bounding box	extreme clicking	scribble	bounding box	extreme clicking
shared	80.5	81.2	81.4	79.3	78.8	79.8
not shared	80.9	81.5	82.0	79.7	79.3	80.4

Table 7
Ablation study of the quality assessment module in terms of DSC (%). See Section 4.5.3 for details.

Variant	Lung cancer	Colon cancer
oracle	81.4	80.4
random	80.1	77.5
quality assessment	81.3	79.7

Table 8
Quantitative comparison of the quality assessment module against the loss prediction module in Yoo and Kweon (2019) on MSD test. DSC (%) is used as the metric. Section 4.5.3 for details.

Uncertainty technique	Lung cancer			Colon cancer		
	scrib.	bound. box	extreme click	scrib.	bound. box	extreme click
(Yoo and Kweon, 2019)	79.8	80.1	80.8	78.4	78.3	79.0
Ours	80.9	81.5	82.0	79.7	79.3	80.4

Table 9
Performance comparison between VMN and automatic 3D nnU-Net in low-data regime, in terms of DSC (%). See Section 4.5.5.

Model	Lung cancer			Colon cancer		
	10%	20%	50%	10%	20%	50%
3D nnU-Net	25.8	41.6	64.0	30.7	45.3	54.5
VMN (Ours)	69.7	74.5	81.1	63.2	71.9	80.0

best automatic model DoubleUNet (Jha et al., 2020) by **4.6%** and **2.5%** in terms of mIoU and DSC, respectively. In addition, VMN performs consistently better than the Interactive 3D nnU-Net. Note that since we only provide scribble annotations for MSD and KiTS_{19}, we do not evaluate scribble-guided segmentation on CVC-ClinicDB.

4.4.4. Cross-dataset validation

Next, we investigate the generalization ability of VMN across different datasets. As presented in Table 4, we train the model on KiTS_{19} (or MSD) and test the performance of the model on MSD (or KiTS_{19}). In comparison with results in Tables 1 and 2, our VMN suffers a minor performance decrease in this cross-dataset study. This suggests the strong generalization capability of our model.

4.4.5. Qualitative results

Figs. 2 and 3 depict visual results of our approach against Interactive 3D nnU-Net on representative examples from MSD and KITS_{19} test, under different types of interactions, *i.e.*, scribble, bounding box and extreme clicking. As we can see, our approach produces more accurate segmentation results than the competitor over all the three forms of interactions.

4.5. Diagnostic experiment

To gain more insights into our model, we investigate the influence of essential components in VMN on MSD test.

4.5.1. Memory size

First, we study the impact of memory size to our model. Table 5 lists the DSC segmentation scores for lung cancer and colon cancer. Here, the size '0' indicates that we use the VMN without an external memory, which means that we only use a ResNet-50 for slice-by-slice segmentation. We see that the baseline yields poor results across all the three interaction types. With memory size '1', we see sharp performance improvements across all the settings. For instance, for colon cancer with extreme clicking, the DSC score improves by 13.2%, *i.e.*, from 54.8% to 68.0%. Moreover, it can be seen that the model performance progressively improves when further increasing the memory size, and the gain becomes marginal around the values of '15' and '20'. Hence, we use a default value of '20' in all our experiments.

4.5.2. Interaction network

By default, we train different 2D interaction networks for different types of interactions (*i.e.*, scribble, bounding box and extreme clicking). In this manner, each interaction network can better account for unique features of the corresponding interaction type, and can be expected to yield superior performance. However, this strategy is practically inflexible in model training or deployment. To address this, we further design a 'universal', shared interaction network which is trained by a combination of all training samples with different interaction types. As reported in Table 6, the shared interaction network only encounters minor performance degradation (*i.e.*, 0.4%~0.6%) against the non-shared ones. These results clearly demonstrate the high flexibility of our engine, which facilitates users to use different types of interaction tools in different rounds to better correct segmentation errors.

4.5.3. Quality assessment module

The quality assessment module endows our VMN to automatically find informative slices for further corrections. To prove its effectiveness, we design two baseline models: 'oracle' selects the worst segmented slice by comparing the masks with corresponding ground-truths, while 'random' selects each slice randomly at each round. As presented in Table 7, our quality assessment module performs consistently better than 'random' across the two sets on MSD test, and is comparable to 'oracle'. Fig. 4 presents some visual results of the module on two examples of MSD test. We can see that our predicted quality scores are very close with the true IoU produced by 'oracle'. These results are remarkable since our module is automatic and lightweight, thereby showing

Table 10
Quantitative comparison of different encoding techniques, in terms of DSC (%). See Section 4.5.6 for details.

Encoding technique	Lung cancer			Colon cancer		
	scribble	bounding box	extreme clicking	scribble	bounding box	extreme clicking
Gaussian	80.5	81.6	81.8	79.5	79.5	80.1
Geodesic	80.9	81.5	82.0	79.7	79.3	80.4
Exp. Geodesic	80.8	81.7	81.9	79.7	79.4	80.2

Table 11
Investigation of model robustness with respect to jittering degree when simulating bounding box and extreme clicking annotations. DSC (%) is used as the metric. See Section 4.5.7 for details.

Jitter degree	Lung cancer		Colon cancer	
	bounding box	extreme clicking	bounding box	extreme clicking
0	81.8	82.4	79.5	80.9
5	81.7	82.2	79.3	80.7
10	81.5	82.0	79.3	80.4
20	81.2	81.5	78.7	79.8
30	81.1	81.3	78.6	79.5
40	79.5	79.6	76.9	77.3

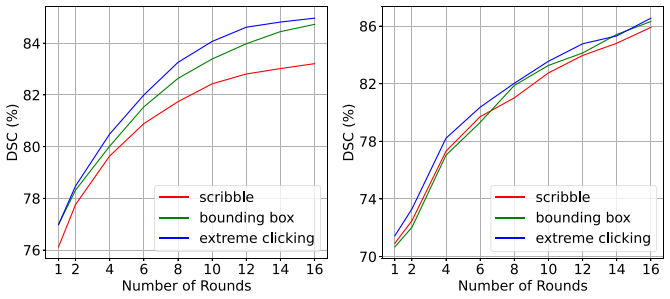

Fig. 5. The effect of multi-round segmentation (Section 3.4) on MSD `test`. Left: Lung Cancer; Right: Colon Cancer. See Section 4.5.4 for details.

great potential to facilitate effortless human-in-the-loop segmentation in practice.

Furthermore, we compare our quality assessment module with the loss prediction module in Yoo and Kweon (2019). As shown in Table 8, our module consistently outperforms the module in Yoo and Kweon (2019) across all the metrics on MSD `test`, demonstrating its superiority.

4.5.4. Multi-round interactive segmentation

Furthermore, we investigate the multi-round segmentation mechanism presented in Section 3.4. Fig. 5 shows DSC results with growing number of interactions (from 1 to 16) on lung and colon subsets of MSD `test`. Clearly, as the number of interaction rounds increase, segmentation accuracy becomes better and better, confirming the effectiveness of multi-round refinement. To gain a good trade-off between accuracy and efficiency, we run our VMN by six rounds by default in all experiments.

4.5.5. Comparison with automatic methods in low-data regime

We further analyze the influence of the number of training data to our interactive model on MSD, and compare it with automatic 3D nnU-Net. We design three sets of experiments for each target (Lung Cancer or Colon Cancer), in which 10%, 20%, 50% of training data are randomly sampled for model training, respectively. As shown in Table 9, automatic 3D nnU-Net is more "data-hungry", encountering serious degradation with the reduction of training data. Our interactive approach, even with 10% of training data, notably outperforms 3D nnU-Net using 50% of training data.

4.5.6. Robustness to interaction encoding techniques

By default, our method utilizes geodesic distance transform introduced in DeepIGeoS (Wang et al., 2018b) for interaction encoding. We compare it with other two techniques, *i.e.*, Gaussian distance transform and exponentialized geodesic distance transform (Luo et al., 2021). For all transformations, we directly use the implementations in Luo et al. (2021). Though exponentialized geodesic distance transform is demonstrated to be more effective than the other two techniques in Luo et al. (2021), Table 10 shows that our model is robust to all of them. This is because our 2D interaction network, as a further encoding procedure, reduces the differences of initial cue maps generated by these techniques.

4.5.7. Robustness to user variance of interactions

We next examine the robustness of our VMN to annotation quality. More precisely, we evaluate the impacts of jittering degree (see Section 4.2) in the forms of bounding box and extreme clicking. The results are reported in Table 11. As seen, our model demonstrates high robustness to annotation disturbations when the degree is smaller than 30, and only degrades with severe annotation noises, *i.e.*, a jittering degree of 40. By default, a degree of 10 is used in all our experiments.

4.5.8. Runtime analysis

Our VMN has no expensive operations like 3D convolutional layers, thus is highly efficient. For a 3D volume with size 512×512×100, our VMN needs 5.13 s on average for one-round segmentation on a NVIDIA RTX2080Ti GPU, whereas it costs more than 50 s for Interactive 3D nnU-Net. Hence our engine enables a significant increase in inference speed.

5. Conclusion

This work presents a novel interactive segmentation engine for 3D medical data. It consists of two essential networks, *i.e.*, a 2D interactive segmentation network that accepts users' hints in a specified slice and gives an initial segmentation prediction, as well as a volumetric memory network (VMN) to propagate the initial mask into other slices. The VMN exploits an external memory to store relevant information, which are retrieved to support the segmentation of each incoming slice. VMN avoids computationally expensive operations like 3D convolutions, thus is more efficient than 3D networks; it takes into account volumetric structural prior, thus is able to deliver more accurate segmentation than 2D counterparts. Moreover, the VMN is equipped with a quality assessment module that endows the model to automatically select informative slices for user feedback, which we believe is an important added value of the engine, and will greatly benefit the usage of the engine in clinical practice. Extensive experiments on three public datasets demonstrate that our engine is capable of producing superior results with a reasonable number of user interactions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research is supported by the Varian Research, Switzerland Grant.

References

Agustsson, E., Uijlings, J.R., Ferrari, V., 2019. Interactive full image segmentation by considering all regions jointly. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 11622–11631.

Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E., 2019. Phiseg: Capturing uncertainty in medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. Springer, pp. 119–127.

Baxter, J.S., Rajchl, M., Peters, T.M., Chen, E.C., 2016. Optimization-based interactive segmentation interface for multiregion problems. J. Med. Imaging 3 (2), 024003.

Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F., 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. 43, 99–111.

Boykov, Y., Funka-Lea, G., 2006. Graph cuts and efficient ND image segmentation. Int. J. Comput. Vis. 70 (2), 109–131.

Boykov, Y.Y., Jolly, M.-P., 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 105–112.

Bredell, G., Tanner, C., Konukoglu, E., 2018. Iterative interaction training for segmentation editing networks. In: International Workshop on Machine Learning in Medical Imaging. pp. 363–370.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2021. Swin-unet: Unet-like pure transformer for medical image segmentation. ArXiv preprint arXiv:2105.05537.

Castrejon, L., Kundu, K., Urtasun, R., Fidler, S., 2017. Annotating object instances with a polygon-rnn. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 5230–5238.

Chang, Y., Menghan, H., Guangtao, Z., Xiao-Ping, Z., 2021. Transclaw u-net: Claw u-net with transformers for medical image segmentation. ArXiv preprint arXiv:2107.05188.

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J., 2018. Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 7103–7112.

Cheng, H.K., Tai, Y.-W., Tang, C.-K., 2021. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 5559–5568.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer Assisted Intervention. pp. 424–432.

Criminisi, A., Sharp, T., Blake, A., 2008. Geos: Geodesic image segmentation. In: Proceedings of the IEEE/CVF European Conference on Computer Vision. pp. 99–112.

Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L., 2020. Pranet: Parallel reverse attention network for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 263–273.

Fang, Y., Chen, C., Yuan, Y., Tong, K.-y., 2019. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 302–310.

Grady, L., 2006. Random walks for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 28 (11), 1768–1783.

Grady, L., Schiwietz, T., Aharon, S., Westermann, R., 2005. Random walks for interactive organ segmentation in two and three dimensions: Implementation and validation. In: Medical Image Computing and Computer Assisted Intervention. pp. 773–780.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022. Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 770–778.

Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al., 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. ArXiv preprint arXiv:1904.00445.

Hesamian, M.H., Jia, W., He, X., Kennedy, P., 2019. Deep learning techniques for medical image segmentation: achievements and challenges. J. Digit. Imaging 32 (4), 582–596.

Hou, X., Xie, C., Li, F., Nan, Y., 2019. Cascaded semantic segmentation for kidney and tumor. In: 2019 Kidney Tumor Segmentation Challenge: KiTS19.

Isensee, F., Maier-Hein, K.H., 2019. An attempt at beating the 3D U-Net. ArXiv preprint arXiv:1908.02182.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2018. Nnu-net: Self-adapting framework for u-net-based medical image segmentation. ArXiv preprint arXiv:1809.10486.

Jha, D., Riegler, M.A., Johansen, D., Halvorsen, P., Johansen, H.D., 2020. Doubleu-net: A deep convolutional neural network for medical image segmentation. ArXiv preprint arXiv:2006.04868.

Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D., 2019. Resunet++: An advanced architecture for medical image segmentation. In: IEEE International Symposium on Multimedia. ISM, IEEE, pp. 225–2255.

Kass, M., Witkin, A., Terzopoulos, D., 1988. Snakes: Active contour models. Int. J. Comput. Vis. 1 (4), 321–331.

Kim, S., Kim, I., Lim, S., Baek, W., Kim, C., Cho, H., Yoon, B., Kim, T., 2019. Scalable neural architecture search for 3d medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. Springer.

Kitrungrotsakul, T., Yutaro, I., Lin, L., Tong, R., Li, J., Chen, Y.-W., 2020. Interactive deep refinement network for medical image segmentation. ArXiv preprint arXiv:2006.15320.

Koohbanani, N.A., Jahanifar, M., Tajadin, N.Z., Rajpoot, N., 2020. Nuclick: a deep learning framework for interactive segmentation of microscopic images. Med. Image Anal. 65, 101771.

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R., 2016. Ask me anything: Dynamic memory networks for natural language processing. In: Proc. ACM Int. Conf. Mach. Learn.. pp. 1378–1387.

Liao, X., Li, W., Xu, Q., Wang, X., Jin, B., Zhang, X., Wang, Y., Zhang, Y., 2020. Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 9394–9402.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 2117–2125.

Luo, X., Wang, G., Song, T., Zhang, J., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2021. MiDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning. Med. Image Anal. 72, 102102.

Maninis, K.-K., Caelles, S., Pont-Tuset, J., Van Gool, L., 2018. Deep extreme cut: From extreme points to object segmentation. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 616–625.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. pp. 565–571.

Mortensen, E.N., Barrett, W.A., 1995. Intelligent scissors for image composition. In: Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques. pp. 191–198.

Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y., 2019. Segmentation of Kidney Tumor by Multi-Resolution VB-Nets. University of Minnesota Libraries Publishing.

Olabarriaga, S.D., Smeulders, A.W., 2001. Interaction in the segmentation of medical images: A survey. Med. Image Anal. 5 (2), 127–142.

Pham, D.L., Xu, C., Prince, J.L., 2000. Current methods in medical image segmentation. Annu. Rev. Biomed. Eng. 2 (1), 315–337.

Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M., 2019. Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 7479–7489.

Rajchl, M., Lee, M.C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., et al., 2016. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans. Med. Imaging 36 (2), 674–683.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 234–241.

Rother, C., Kolmogorov, V., Blake, A., 2004. "Grabcut" interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23 (3), 309–314.

Sakinis, T., Milletari, F., Roth, H., Korfiatis, P., Kostandy, P., Philbrick, K., Akkus, Z., Xu, Z., Xu, D., Erickson, B.J., 2019. Interactive segmentation of medical images through fully convolutional neural networks. ArXiv preprint arXiv:1903.08205.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., Lillicrap, T., 2016. Meta-learning with memory-augmented neural networks. In: Proc. ACM Int. Conf. Mach. Learn. pp. 1842–1850.

Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H., 2022. Transformers in medical imaging: A survey. ArXiv preprint arXiv:2201.09873.

Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al., 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. ArXiv preprint arXiv:1902.09063.

Sukhbaatar, S., Weston, J., Fergus, R., et al., 2015. End-to-end memory networks. Proc. Advances Neural Inf. Process. Syst. 28.

Sun, J., Shi, Y., Gao, Y., Wang, L., Zhou, L., Yang, W., Shen, D., 2018. Interactive medical image segmentation via point-based interaction and sequential patch learning. ArXiv preprint arXiv:1804.10481.

Top, A., Hamarneh, G., Abugharbieh, R., 2011. Active learning for interactive 3D image segmentation. In: Medical Image Computing and Computer Assisted Intervention. Springer, pp. 603–610.

Wang, G., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2020. Uncertainty-guided efficient interactive refinement of fetal brain segmentation from stacks of MRI slices. In: Medical Image Computing and Computer Assisted Intervention. Springer, pp. 279–288.

Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al., 2018a. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Trans. Med. Imaging 37 (7), 1562–1573.

Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al., 2018b. DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 41 (7), 1559–1572.

Wang, G., Zuluaga, M.A., Pratt, R., Aertsen, M., Doel, T., Klusmann, M., David, A.L., Deprest, J., Vercauteren, T., Ourselin, S., 2016. Slic-Seg: A minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal MRI in multiple views. Med. Image Anal. 34, 137–147.

Xia, Y., Liu, F., Yang, D., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H., 2020. 3D semi-supervised learning with uncertainty-aware multi-view co-training. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3646–3655.

Yang, T., Chan, A.B., 2018. Learning dynamic memory networks for object tracking. In: Proceedings of the IEEE/CVF European Conference on Computer Vision. pp. 152–167.

Yoo, D., Kweon, I.S., 2019. Learning loss for active learning. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 93–102.

Yu, Q., Yang, D., Roth, H., Bai, Y., Zhang, Y., Yuille, A.L., Xu, D., 2020. C2FNAS: Coarse-to-fine neural architecture search for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 4126–4135.

Zhang, S., Liew, J.H., Wei, Y., Wei, S., Zhao, Y., 2020. Interactive object segmentation with inside-outside guidance. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 12234–12244.

Zhang, Y., Liu, H., Hu, Q., 2021a. Transfuse: Fusing transformers and cnns for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 14–24.

Zhang, J., Shi, Y., Sun, J., Wang, L., Zhou, L., Gao, Y., Shen, D., 2021b. Interactive medical image segmentation via a point-based interaction. Artif. Intell. Med. 111, 101998.

Zhang, Y., Wang, Y., Hou, F., Yang, J., Xiong, G., Tian, J., Zhong, C., 2019. Cascaded volumetric convolutional network for kidney tumor segmentation from ct volumes. ArXiv preprint arXiv:1910.02235.

Zhao, W., Jiang, D., Queralta, J.P., Westerlund, T., 2020. Multi-scale supervised 3D U-net for kidneys and kidney tumor segmentation. ArXiv preprint arXiv:2004.08108.

Zhao, F., Xie, X., 2013. An overview of interactive medical image segmentation. Ann. BMVA 2013 (7), 1–22.

Zhou, T., Li, L., Bredell, G., Li, J., Konukoglu, E., 2021. Quality-aware memory network for interactive volumetric image segmentation. In: Medical Image Computing and Computer Assisted Intervention. Springer, pp. 560–570.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 3–11.

Zhou, T., Wang, W., Konukoglu, E., Van Gool, L., 2022. Rethinking semantic segmentation: A prototype view. In: Proceedings of the IEEE/CVF Computer Vision Pattern Recognition. pp. 2582–2593.

Zhou, Y., Xie, L., Shen, W., Wang, Y., Fishman, E.K., Yuille, A.L., 2017. A fixed-point model for pancreas segmentation in abdominal CT scans. In: Medical Image Computing and Computer Assisted Intervention. pp. 693–701.

Zhu, Z., Liu, C., Yang, D., Yuille, A., Xu, D., 2019. V-nas: Neural architecture search for volumetric medical image segmentation. In: 3DV. pp. 240–248.