

MIDeepSeg: Minimally Interactive Segmentation of Unseen Objects from Medical Images Using Deep Learning

Xiangde Luo^a, Guotai Wang^{a,*}, Tao Song^b, Jingyang Zhang^c, Michael Aertsen^e, Jan Deprest^{e,f,g}, Sébastien Ourselin^d, Tom Vercauteren^d, Shaoting Zhang^{a,b}

^aSchool of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
^bSenseTime Research, Shanghai, China

^cSchool of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China

^dSchool of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

^eDepartment of Radiology, University Hospitals Leuven, Leuven, Belgium

^fDepartment of Obstetrics and Gynaecology, University Hospitals Leuven, Leuven, Belgium

^gInstitute for Women's Health, University College London, London, UK

Abstract

Segmentation of organs or lesions from medical images plays an essential role in many clinical applications such as diagnosis and treatment planning. Though Convolutional Neural Networks (CNN) have achieved the state-of-the-art performance for automatic segmentation, they are often limited by the lack of clinically acceptable accuracy and robustness in complex cases. Therefore, interactive segmentation is a practical alternative to these methods. However, traditional interactive segmentation methods require a large amount of user interactions, and recently proposed CNN-based interactive segmentation methods are limited by poor performance on previously unseen objects. To solve these problems, we propose a novel deep learning-based interactive segmentation method that not only has high efficiency due to only requiring clicks as user inputs but also generalizes well to a range of previously unseen objects. Specifically, we first encode user-provided interior margin points via our proposed exponentialized geodesic distance that enables a CNN to achieve a good initial segmentation result of both previously seen and unseen objects, then we use a novel information fusion method that combines the initial segmentation with only few additional user clicks to efficiently obtain a refined segmentation. We validated our proposed framework through extensive experiments on 2D and 3D medical image segmentation tasks with a wide range of previous unseen objects that were not present in the training set. Experimental results showed that our proposed framework 1) achieves accurate results with fewer user interactions and less time compared with state-of-the-art interactive frameworks and 2) generalizes well to previously unseen objects.

Keywords: Interactive image segmentation, convolutional neural network, geodesic distance, generalization

1. Introduction

Accurate and robust segmentation of organs or lesions from medical images plays an essential role in many clinical applications such as diagnosis and treatment planning (Zhao and Xie, 2013; Masood et al., 2015). Although automatic segmentation methods have been studied for many years, it remains challenging for them to obtain a consistently precise segmentation in cases with large anatomical variation and complex pathologies (Wang et al., 2018b). This is mainly due to the inherent limitations of medical images, such as low contrast, different imaging and segmentation protocols, and variations among patients (Wang et al., 2018b). In contrast, interactive segmentation methods, which leverage the user's knowledge and experience to obtain a more accurate and robust result, are more practical and widely used in clinical appli-

cations (Zhao and Xie, 2013; Masood et al., 2015; Wang et al., 2018b).

A desirable interactive segmentation tool should 1) achieve accurate segmentation results with as few user inputs as possible, leading to reduced burdens on the user; 2) have high efficiency so that the user can get real-time response, even when dealing with volumetric data; 3) generalize well to different objects so that it is ready-to-use for new objects or image modalities. However, existing interactive segmentation methods rarely satisfy all these often competing requirements. Many traditional interactive methods use low-level features (e.g., gray level or color distribution) for image segmentation (Hu et al., 2019), such as Graph Cuts (Boykov and Jolly, 2001), ITK-SNAP (Yushkevich et al., 2006), GeoS (Criminisi et al., 2008), Random Walks (Grady, 2006) and GrowCut (Vezhnevets and Konouchine, 2005). As low-level features cannot effectively distinguish the object from the background in many situations with low contrast (Hu et al., 2019), these methods often require a large amount of user inter-

*Corresponding author

Email address: guotai.wang@uestc.edu.cn (Guotai Wang)

actions and long user time to obtain reliable results. To reduce the amount of annotations required from the user to build an adequate foreground/background model, machine learning has been widely used to perform interactive segmentation. For example, SlicSeg (Wang et al., 2016b) and DyBaORF (Wang et al., 2016a) use an Online Random Forest (ORF) to segment the placenta from Magnetic Resonance Imaging (MRI) volume. GrabCut (Rother et al., 2004) uses Gaussian Mixture Models (GMMs) to estimate the foreground and background distributions. It obtains an initial result by a user-provided bounding box around the region of interest and allows additional interactions for refinement. Wang et al. (2014) used active learning to actively select candidate regions for querying the user to obtain much informative user feedback and thus reduced user interactions. These algorithms perform better than traditional methods without machine learning, but they are limited by the use of hand-crafted features (Wang et al., 2018a,b). As a result, they still require a considerable amount of user interactions for accurate segmentation.

Recently, with the ability to learn high-level semantic features automatically, deep learning with Convolutional Neural Networks (CNNs) has achieved state-of-the-art performance for image segmentation (Shen et al., 2017; Litjens et al., 2017). To take advantage of the good representation ability of CNNs and overcome the limited accuracy and robustness of the automatic CNNs, some deep learning-based interactive segmentation tools (Wang et al., 2018b; Hu et al., 2019; Wang et al., 2018a; Rajchl et al., 2016; Xu et al., 2016; Maninis et al., 2018; Sakinis et al., 2019) have recently been proposed. The methods of Hu et al. (2019), Xu et al. (2016) and Maninis et al. (2018) are designed to segment 2D RGB images interactively and lack evaluation on medical images with low contrast and ambiguous boundaries. Castrejon et al. (2017) and Acuna et al. (2018) integrated reinforcement learning and graph neural networks into a unified polygon-based interactive segmentation framework, where the user is allowed to drag a point on the polygon for refinement, but its ability to deal with objects with complex shapes and 3D medical images is limited.

In contrast, DeepIGeoS (Wang et al., 2018b), IF-Seg (Sakinis et al., 2019), DeepCut (Rajchl et al., 2016) and BIFSeg (Wang et al., 2018a) are specially designed to segment medical images. DeepCut (Rajchl et al., 2016) uses a set of user-provided bounding box as sparse annotations to train CNNs for the segmentation of fetal brain and lung from fetal MRI. Roth et al. (2019, 2020) combined extreme points (Maninis et al., 2018) with random walkers (Grady, 2006) for weakly supervised 3D medical image segmentation. Even though this method and DeepCut (Rajchl et al., 2016) reduced the annotation cost significantly, it was designed for weakly supervised model training over a large dataset rather than interactively editing a single segmentation result at test time. Raju et al. (2020) further used extreme points in a user-guided domain adaptation method for pathological liver segmen-

tation. DeepIGeoS (Wang et al., 2018b) performs user-friendly interactive segmentation by combining CNNs and user-provided scribbles, where a CNN is used to obtain an initial segmentation and another CNN accepts additional user interactions for refinement. However, DeepIGeoS can only deal with objects present in the training set and lacks of adaptability to previously unseen objects. Following Xu et al. (2016), IF-Seg (Sakinis et al., 2019) takes user clicks and the raw image as input for interactive medical image segmentation. Despite the fact that the framework is easy to use, its generalizability was only validated with a single previously unseen structure, and the ability to deal with various unseen objects in different modalities was not shown. BIFSeg (Wang et al., 2018a) exploits user-provided bounding boxes and image-specific fine-tuning to segment some unseen objects, but it is limited by dealing with only few unseen objects in the same image modality or similar context and requiring time-consuming fine-tuning for each test image. Therefore, novel interactive frameworks for medical image segmentation with higher efficiency and generalizability is highly desirable.

Besides, a practical problem for CNN-based interactive segmentation methods is to effectively encode user interactions, as different encoding strategies have a large impact on the interactive segmentation performance. Most of existing works encode user interactions by transforming them into a cue map, such as Euclidean distance map (Hu et al., 2019; Xu et al., 2016; Li et al., 2018; Benenson et al., 2019; Hao et al., 2019), Gaussian heatmap (Maninis et al., 2018; Wang et al., 2019), and iso-contours derived from user clicks (Khan et al., 2019). However, these encoding methods do not take the image context information into account. In contrast, the geodesic distance transform is spatially smooth and contrast-sensitive to encode user interactions (Criminisi et al., 2008; Bai and Sapiro, 2009; Price et al., 2010). DeepIGeoS (Wang et al., 2018b) uses geodesic distance transform with a specially designed threshold to deal with user-provided interactions. However, it is time-consuming to find an appropriate threshold value to truncate the generated geodesic distance map when dealing with different objects. We assume that a context-aware and parameter-free encoding method is helpful for improving the segmentation accuracy and generalizability.

To tackle the above-mentioned challenges, we propose a new generalizable framework for more intelligent and accurate interactive segmentation of 2D and 3D medical images, which aims at not only obtaining high performance and efficiency for segmentation of previously seen objects, but also achieving high generalization to a range of previously unseen objects. Our method takes advantages of CNNs and only requires few clicks as user interactions. We present a new way to encode user interactions based on Exponentialized Geodesic Distance (EGD) transform, which is context-aware and parameter-free and helps to improve the segmentation obtained by the CNN. We also propose an information fusion method that efficiently fuses addi-

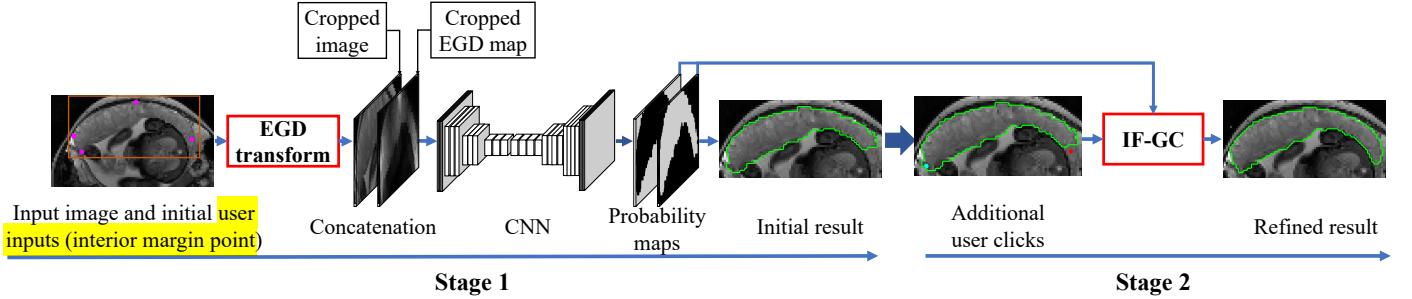


Figure 1: Pipeline of the proposed Minimally Interactive Deep learning-based Segmentation framework (MIDeepSeg). Stage 1: User-provided interior margin points are encoded by Exponentialized Geodesic Distance (EGD) maps to guide a CNN to obtain an initial result. Stage 2: Refining the initial segmentation based on additional user clicks and our proposed Information Fusion followed by Graph Cut (IF-GC). Note that this framework is ready to use for segmentation of previously unseen objects without the need of extra fine-tuning or re-training.

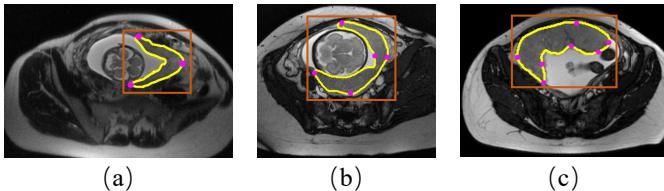


Figure 2: Simulation of interior margin points on training images for different shapes of placenta. Fuchsia: simulated clicks on placenta edge. Brown: interior margin points-derived relaxed bounding box. Yellow: ground truth.

tional user clicks with the initial segmentation to obtain a refined segmentation. Differently from existing interactive medical image segmentation frameworks (Wang et al., 2018a,b; Rajchl et al., 2016), our method is more efficient as it only works on a sub-region of the image and does not need to train an additional CNN on the fly for the refinement. Moreover, we validate the effectiveness of this framework with a large range of previously seen and unseen objects. The superiority of our method over existing interactive segmentation methods is validated with five types of 2D unseen objects and four types of 3D unseen objects from different types of image contexts and modalities.

2. Methods

The proposed Minimally Interactive Deep learning-based Segmentation framework is referred to as MIDeepSeg and illustrated in Fig. 1. It consists of two stages. In the first stage, the user provides few clicks near the boundary (i.e., interior margin point) of the target object. These points are used to infer a relaxed bounding box to crop the input image. Based on the cropped image, all user-provided interior margin points are converted to a cue map based on our proposed EGD transform. Then, the cue map is concatenated with the cropped input image as the input of a CNN to obtain an initial segmentation result. In the second stage, the user provides some additional clicks to indicate mis-segmented regions, and a refined result is obtained by our proposed Information Fusion followed by Graph Cuts (IF-GC). At

test time, the refinement step can run several times until the result is accepted by the user. After training with a small set of objects, our framework is ready to use for the segmentation of previously unseen objects without the need for fine-tuning or re-training that is time-consuming and requires additional annotations.

2.1. User interaction based on interior margin points

Many existing CNN-based interactive segmentation frameworks use scribbles (Wang et al., 2018a) or bounding boxes (Rajchl et al., 2016) or both (Wang et al., 2018b) as interactive cues. They need the user to drag the cursor carefully, which requires a lot of user's efforts (Papadopoulos et al., 2017; Maninis et al., 2018). Using clicks as user interactions is a more user-friendly and effective way as demonstrated by previous works (Xu et al., 2016; Maninis et al., 2018; Wang et al., 2019; Papadopoulos et al., 2017). Recently, Maninis et al. (2018) proposed a framework that only needs the user to provide clicks for extreme points (i.e., left-, right-, top- and bottom-most pixels) of an object for RGB image segmentation, which reduces the amount of user interactions substantially. However, in medical images, the accurate extreme points are hard and time-consuming to find, which increases the burden on the user, since the target organs or lesions have a large variability in the size and shape across different patients or imaging protocols, especially in 3D volumetric data. In addition, for irregular and concave shapes, extreme points are not enough to capture the main shape of the object (as shown in Fig. 2), which can limit the performance of the CNN.

To alleviate these limitations, we propose to use interior margin points as user interactions, where the user only needs to provide some clicks that are in the inner side and close the boundary of the target. Compared with DEXTR (Maninis et al., 2018) that uses at most four extreme points and optionally with one extra point, our interior margin points can provide more shape information for different types of organs with complex and irregular shapes. In addition, putting clicks exactly on the object boundary and even extreme points is hard for users at test

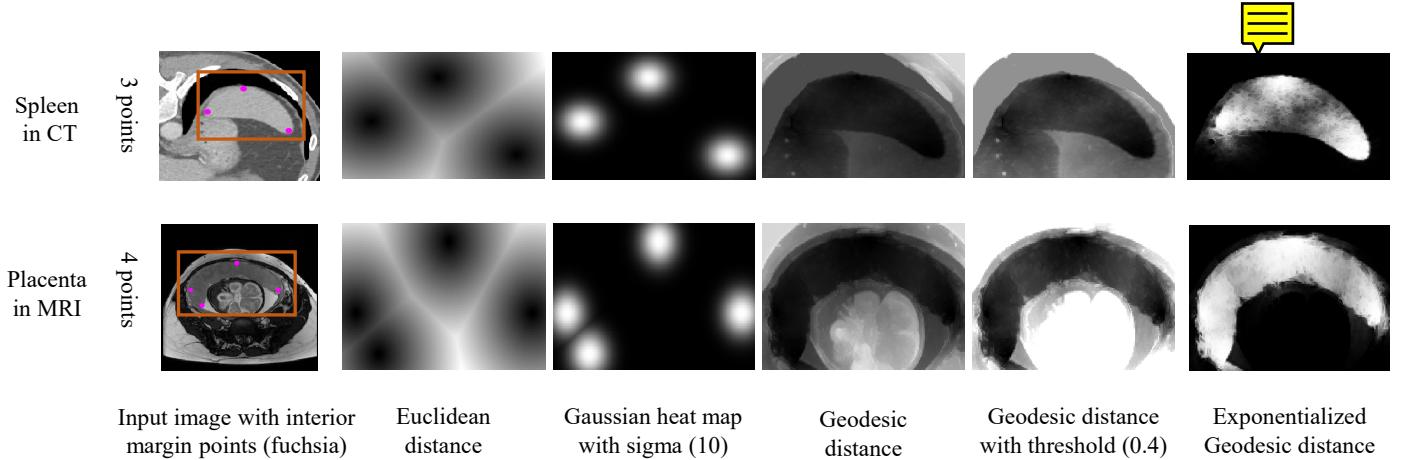


Figure 3: Visual comparison of different cue maps generated from user-provided interior margin points. (Fuchsia: interior margin points. Brown: inferred relaxed bounding box)

time, and relaxing the clicks to inner side of the boundary makes the interactions more friendlier and convenient to implement, which tolerates inaccurate clicks. We relax these points towards the inside region because an exponentialized geodesic distance transform of these interior margin points can be a good approximation of the saliency map of the target object, as shown in Fig. 3. Therefore, interior margin points bring potential advantages in guiding CNNs to deal with different types of unseen objects as well.

During training, all interior margin points for each object were automatically simulated based on the ground truth mask and edge detector (Harris et al., 1988). The interior margin points are generated based on two rules: First, these points should be located in the object and near the boundary. Second, a relaxed bounding box determined by these points should cover the entire object region. Therefore, we simulate user interactions for a training image in two steps. 1) To ensure that the relaxed bounding box covers the whole region of interest, few points on the ground truth boundary (three or four for 2D objects, five or six for 3D objects) close to the extreme points (Maninis et al., 2018) of the target object are selected. Then, we randomly sample n points from remaining boundary points of the target to provide more shape information, where n is a random number from 0 to 5. 2) To simulate real user clicks that may not be accurately positioned on the object boundary, all these points obtained in step 1 are slightly moved towards the inner side of the boundary by several pixels/voxels to obtain our interior margin points. We moved simulated points towards the inner side of the target object as the users are asked to put the interior margin points in the inner side of the boundary as well. Then the bounding box determined by these points is relaxed by several pixels/voxels to include some background region. Examples of simulated 2D interior margin points and relaxed bounding boxes on training images are shown in Fig. 2. In test stage, the user is required to provide the interior margin points in such a way

that they satisfy the above two rules. The relaxed bounding box determined by user interactions is expanded with a small margin to include some contextual information.

2.2. Exponentialized geodesic distance transform

It is critical for CNN-based interactive methods to encode user interactions efficiently. A desirable encoding method should take image context into account and can be combined with CNNs directly without any manually designed parameters. However, existing interaction encoding methods such as Euclidean distance transform (Xu et al., 2016; Li et al., 2018), Gaussian heatmap (Maninis et al., 2018; Wang et al., 2019), iso-contours (Khan et al., 2019) and geodesic distance transform (Wang et al., 2018b) do not have these merits at the same time. To deal with this problem, we propose a context-aware and parameter-free encoding method: Exponentialized Geodesic Distance (EGD) transform, which is a combination of geodesic distance transform and exponential transform.

Suppose \mathcal{S}_s represent the set of pixels/voxels belonging to the simulated interior margin points in the training stage or user-provided interior margin points in the testing stage. Let i be a pixel/voxel in the input image I , then the unsigned EGD from i to \mathcal{S}_s is:

$$EGD(i, \mathcal{S}_s, I) = \min_{j \in \mathcal{S}_s} e^{-\mathcal{D}_{geo}(i, j, I)} \quad (1)$$

$$\mathcal{D}_{geo}(i, j, I) = \min_{p \in \mathcal{P}_{i,j}} \int_0^1 \|\nabla I(p(n)) \cdot v(n)\| dn \quad (2)$$

where $\mathcal{P}_{i,j}$ is the set of all paths between pixels/voxels i and j . p is one feasible path and it is parameterized by $n \in [0, 1]$. $v(n) = p'(n)/\|p'(n)\|$ is a unit vector that is tangent to the direction of the path. Note that the EGD here is defined for scalar images but can easily be extended to vector-valued (i.e., multi-channel or multi-modal) images. Fig. 3 shows some examples of cue maps obtained by different encoding methods applied to some interior margin points. It can be observed that EGD-based cue map

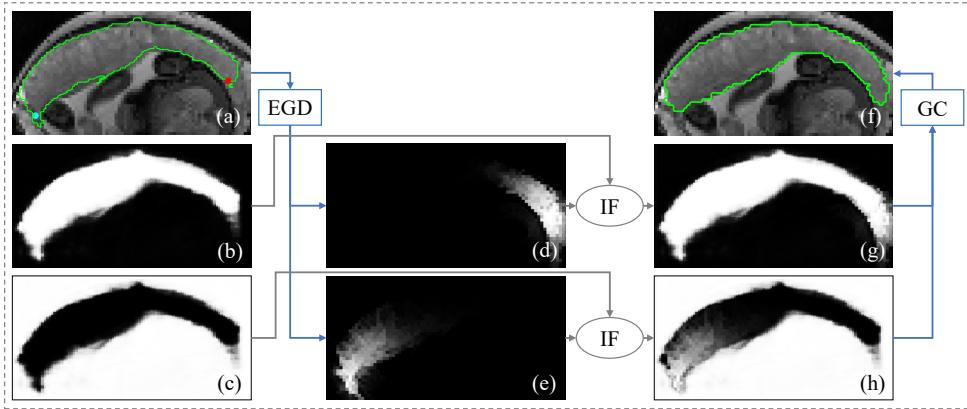


Figure 4: Illustration of refinement by information fusion. (a) The user provides clicks to indicate under-segmentation(red) and over-segmentation(cyan) regions. (b) and (c) are initial segmentation foreground and background probability maps obtained by CNN in the first stage, respectively. (d) and (e) are cue maps based on foreground and background refinement clicks and EGD transformation, respectively. (g) and (h) are calibrated foreground and background probability maps, respectively. (f) is refined segmentation result. (IF: Information Fusion; EGD: Exponentialized geodesic distance transformation; GC: Graph Cut)

differentiates the foreground from the background better than those based on the other encoding methods. Therefore, it has the potential to provide more shape, position and context information to guide the CNN to obtain a good initial segmentation result.

2.3. Initial segmentation based on cue map and CNN

In this paper, we focus on designing an efficient and general framework to deal with seen and unseen objects from different types of images. Therefore, our framework does not rely on a specific design of CNN structure. To demonstrate its utility, we use adapted 2D-UNet (Ronneberger et al., 2015) and 3D-UNet (Çiçek et al., 2016) for 2D and 3D segmentation, respectively. We replace the batch normalization layers with instance normalization layers that has a better adaptability to different kinds of images, and reduce the feature channel numbers by four times to balance the performance, memory cost, and time consumption. In the training stage, all interior margin points and relaxed bounding boxes are automatically simulated based on the ground truth label, as described in Section 2.1. Then all interior margin points are converted into a cue map that is concatenated with the cropped input image as the input of the CNN, as shown in Fig. 2. In the testing stage, the user is asked to provide interior margin points for a given target. Then, the CNN can give an initial segmentation result. To correct the mis-segmentation, we use a refinement stage with information fusion between the initial segmentation and additional user clicks, as described in the following.

2.4. Refinement based on information fusion between initial segmentation and additional user clicks

For deep learning-based interactive segmentation, it is important to support refinement of an initial segmentation. Existing methods either require an additional model for refinement (Wang et al., 2018b; Castrejon et al., 2017;

Acuna et al., 2018; Zhou et al., 2019a; Liao et al., 2020) or need to fine-tune the pre-trained model for a specific image (Wang et al., 2018a). However, these refinement methods are time and memory consuming, and not ready-to-use for unseen objects. In addition, Chen et al. (2017) and Kamnitsas et al. (2016) used CRF (Lafferty et al., 2001) to refine CNN’s prediction automatically. However, these CRF-based (Lafferty et al., 2001) refinement methods (Chen et al., 2017; Kamnitsas et al., 2016) were not designed for interactive segmentation. Differently from these methods, we propose an efficient and simple refinement method based on a novel method for information fusion between initial segmentation and additional user interactions, which generalizes better to previously unseen objects without extra fine-tuning and re-training. Fig. 4 shows an illustration of our information fusion method.

In the refinement stage, the user is asked to provide some additional clicks to indicate mis-segmented foreground and background regions, respectively. To efficiently encode these new interactions, we use the proposed EGD transform again to get two additional interaction-derived cue maps: E^f and E^b are cue maps based on EGD of user-provided foreground and background clicks for refinement, respectively. Note that, we do not reuse the initial EGD map obtained in the first stage directly, but combine the initial interior margin points with refinement clicks for calculating the new EGD maps in the refinement step. The values of E^f and E^b are in the range of [0, 1] and represent the similarity between each pixel and foreground/background clicks. Let P^f and P^b denote the initial foreground and background probability map obtained by the CNN, and i represent a pixel/voxel in the input image I . The information fusion strategy is proposed to refine the P^f and P^b according E^f and E^b . Specifically, we aim to automatically emphasize on E^f and E^b when pixel i is close to the refinement clicks, otherwise P^f and P^b tend to keep unchanged. We define user-calibrated foreground

(R_i^f) and background (R_i^b) probability for pixel i as:

$$E_i^f = \frac{e^{-D_i^f}}{e^{-D_i^f} + e^{-D_i^b}} \quad (3)$$

$$E_i^b = \frac{e^{-D_i^b}}{e^{-D_i^f} + e^{-D_i^b}} \quad (4)$$

$$R_i^f = (1 - \alpha_i) * P_i^f + \alpha_i * E_i^f \quad (5)$$

$$R_i^b = (1 - \alpha_i) * P_i^b + \alpha_i * E_i^b \quad (6)$$

$$\alpha_i = e^{-\min(D_i^f, D_i^b)} \quad (7)$$

where $\alpha_i \in [0, 1]$ is an automatic and adaptive weighting factor. When i is close to the clicks, α_i is close to 1.0, and R_i^f (R_i^b) is more affected by E_i^f (E_i^b). If no clicks are provided for the foreground (background), we set the corresponding D_i^f or D_i^b to a constant value. Let C^f and C^b denote the clicks for foreground and background, respectively, so the entire set of clicks is $C = C^f \cup C^b$. Let c_i denote the user-provided label of a pixel in the clicks, then we have $c_i = 1$ if $i \in C^f$ and $c_i = 0$ if $i \in C^b$. We integrate R^f and R^b into a Conditional Random Field (CRF) to get the refined segmentation:

$$E = \sum_i \phi(y_i|I) + \lambda \cdot \sum_{i,j} \psi(y_i, y_j|I) \quad (8)$$

$$\text{subject to : } y_i = c_i \text{ if } i \in C \quad (9)$$

where ϕ and ψ are the unary and pairwise energy terms, respectively. λ specifies a relative weight between ϕ and ψ . In this paper:

$$\phi(y_i|I) = -(y_i \log(r_i) + (1 - y_i) \log(1 - r_i)) \quad (10)$$

$$\psi(y_i, y_j|I) \propto \exp\left(-\frac{(I_i - I_j)^2}{2\sigma^2}\right) \cdot \frac{1}{dist_{ij}} \quad (11)$$

where r_i denotes the value of pixel i in R^f , and $y_i = 1$ if i belongs to the foreground and 0 otherwise. I_i and I_j mean the intensity of pixel i and j in image I , respectively. $dist_{ij}$ is the Euclidean distance between pixel/voxel i and j . σ is a parameter to control the effect of intensity difference. In this paper, the CRF (Lafferty et al., 2001) problem in Eq. (8) is submodular and can be solved by Graph Cut through max-flow/min-cut (Boykov and Jolly, 2001).

2.5. Implementation details

We implemented the U-Net and 3D U-Net for 2D and 3D images segmentation by Pytorch (Paszke et al., 2019), respectively. The training was on a Ubuntu(16.04) desktop with an Intel Core i7 CPU and one GTX 1080Ti NVIDIA GPU and 120 GB memory. We used the DICE loss function and Adam algorithm for optimization, with mini-batch size of 4, weight decay 10^{-4} . For 2D segmentation, we totally trained 300 epochs for network convergence. The learning rate was kept as 10^{-4} for the first 150 epochs and then halved for every 30 epochs. For 3D segmentation, we totally trained 2000 epochs for network convergence. The learning rate was kept as 10^{-4} for the first 1000 epochs and then halved for every 200 epochs. Each image/volume was cropped based on the relaxed bounding box derived from the interior margin points firstly and then normalized by the mean value and standard deviation of the cropped image. To boost the generalizability to unseen objects, we used several data augmentation methods during the training stage, including random rotation, random scaling, random flipping in space and intensity. Following DeepIGeoS (Wang et al., 2018b), we used open source code to compute geodesic distance¹ and solve Eq. (8) based on max-folw², respectively.

In this paper, all testing processes with user interactions were performed on a Ubuntu(16.04) desktop with an Intel Core i7 CPU and a GTX 1080Ti NVIDIA GPU. Following the practice of DeepIGeoS (Wang et al., 2018b) and BIF-Seg (Wang et al., 2018a), the values of λ in Eq. (8) was 5 and σ in Eq. (11) was 0.1 based on a grid search with the validation data. But for specific cases, it also allows the user to set these two parameters manually, like many existing works (Boykov and Jolly, 2001; Rother et al., 2004; Criminisi et al., 2008). We developed two PyQt GUIs for user interactions on 2D images and 3D volumes respectively. (See supplementary videos ³)

3. Experiments and results

3.1. Comparison methods and evaluation metrics

To investigate the performance of different encoding methods with the same interior margin points in the first stage of our segmentation method, we compared our EGD with Euclidean distance transform, Gaussian distance transform and geodesic distance transform, which are referred to as EGD, Eucl, Gauss and Geos respectively. In addition, we compared them with segmentation based on the bounding box without encoding of interactions, which is referred to as BBox. All these methods were based on the same CNN structure. For fair comparison, Eucl, Gauss and Geos were implemented with their

¹geodesic distance: <https://github.com/taigw/GeodisTK>.

²max-flow: <https://vision.cs.uwaterloo.ca/code/>.

³<https://www.youtube.com/watch?v=eq-tqlJnckE>

respective optimal parameters for encoding user-provided interactions. (see the supplementary document)

MIDeepSeg was also compared with several existing interactive segmentation methods. In 2D cases, in addition to traditional methods like Graph Cuts (Boykov and Jolly, 2001), Random Walks (Grady, 2006) and SlicSeg (Wang et al., 2016b), we also compared recent deep learning-based methods including DeepIGeoS (Wang et al., 2018b), DIOS (Xu et al., 2016), DeepGrabCut (Xu et al., 2017) and DEXTR (Maninis et al., 2018), where the same 2D network structure was used as our 2D version of MIDeepSeg. For 3D segmentation, we compared MIDeepSeg with ITK-SNAP (Yushkevich et al., 2006) and 3D Graph Cuts (Boykov and Jolly, 2001), as well as 3D versions of DeepIGeoS (Wang et al., 2018b), DIOS (Xu et al., 2016), DeepGrabCut (Xu et al., 2017) and DEXTR (Maninis et al., 2018) that used the same 3D network as MIDeepSeg for 3D segmentation. Graph Cuts, SlicSeg, Random Walks, DeepIGeoS and DIOS allow the user to refine the results multiple times. DeepGrabCut just allows the user to draw a bounding box at the beginning and does not support further interactions for refinement. DEXTR takes the extreme points as the user interactions and allows the user to refine the results once. Graph Cuts, SlicSeg, Random Walks, and ITK-SNAP are traditional interactive segmentation methods without the need of training with an annotated dataset and have a high generalization. In contrast, DeepIGeoS, DIOS, DeepGrabCut, and DEXTR are deep learning-based methods and require labeled data to train, and DeepIGeoS cannot deal with unseen objects. Two users respectively used these interactive frameworks to segment each test image until the result was visually acceptable, and we reported the average results of the two users achieved. The segmentation results were compared with ground truth label which was annotated by experienced radiologists manually. For quantitative evaluation, we used the Dice similarity coefficient and the average symmetric surface distance (ASSD).

$$Dice = \frac{2 \cdot |R_p \cap R_g|}{|R_p| + |R_g|} \quad (12)$$

where R_p and R_g denote the region segmented by an algorithm and the ground truth label, respectively.

$$ASSD = \frac{1}{|S_p| + |S_g|} \left(\sum_{i \in S_p} d(i, S_g) + \sum_{i \in S_g} d(i, S_p) \right) \quad (13)$$

where S_p and S_g represent the set of surface points of the result provided by an algorithm and ground truth label, respectively. $d(i, S_p)$ is the shortest Euclidean distance between the point i and the surface S_g . To investigate the efficiency of these methods, we listed the user time and amount of user interaction points of each segmentation task.

Object	Modality	Training	Testing	DataSet
Placenta	MRI (T2)	532 slices	176 slices	Ours
Spleen	CT	235 slices	159 slices	BTCV, TCIA ³
Kidney	MRI (T1)	No	100 slices	CHAOS
Kidney	MRI (T2)	No	100 slices	CHAOS
Kidney	CT	No	100 slices	CHAOS
Spleen	MRI (T1)	No	100 slices	CHAOS
Spleen	MRI (T2)	No	100 slices	CHAOS
Spleen	CT	No	100 slices	CHAOS
Prostate	MRI (T2)	No	72 slices	MSD
Fetal brain	Ultrasound	No	60 slices	HC18

Table 1: Datasets used for training and testing the 2D interactive segmentation framework. Note that for spleen, BTCV and TCIA are training set and testing set, respectively.

3.2. Interactive segmentation of 2D images

3.2.1. Data

Firstly, we validate the propose pipeline with two 2D applications: segmentation of placenta and spleen from fetal MRI and abdomen CT, respectively. Specifically, the placenta data were collected from clinical MRI scans of 30 pregnancies in the second trimester, and were acquired in axial view with pixel size between 0.7422 mm × 0.7422 mm and 1.582 mm × 1.582 mm and slice thickness 3 - 4 mm. Each slice was resampled with a uniform pixel size of 1 mm × 1 mm. We used 532 slices from 18 volumes, 111 slices from 4 volumes and 176 slices from 8 volumes for training, validation and testing, respectively. The ground truth was manually delineated by an experienced Radiologist. For the spleen data, we randomly selected 235 slices of spleen from 47 volumes (5 slices per volume) in BTCV (Marsh, 2013) data set for training, and selected 159 slices from 53 volumes (3 slices per volume) in TCIA⁴ data set for testing. Secondly, to validate the generalizability of our method, we apply our model trained only with placenta in MRI to four types of organs from a variety of modalities that were not present in the training set: 1) Kidney in CT, T1-weighted and T2-weighted MRI in the CHAOS⁵ training set. We randomly selected 100 slices for these three cases respectively. 2) Spleen in CT, T1-weighted and T2-weighted MRI in the CHAOS training set. We also randomly selected 100 slices for these three cases respectively. 3) Prostate in T2-weighted MRI from MSD⁶ Task05 dataset, where we randomly selected 72 slices from 24 cases. 4) Fetal brain in ultrasound images from HC18⁷ dataset, where we randomly selected 60 slices. Information of the training and testing set is listed in Tabel 1. To deal with different organs at different scales, we resized the cropped sub-region and the cue map to 64 × 64 as the input of CNN.

3.2.2. Initial segmentation based on EGD-based cue map and 2D CNN

Fig. 5 shows some examples of the initial segmentation of placenta from MRI and spleen from CT with user-

⁴<https://zenodo.org/record/1169361#.YETa43UzYUE>

⁵<https://chaos.grand-challenge.org/>

⁶<http://medicaldecathlon.com/>

⁷<https://hc18.grand-challenge.org/>

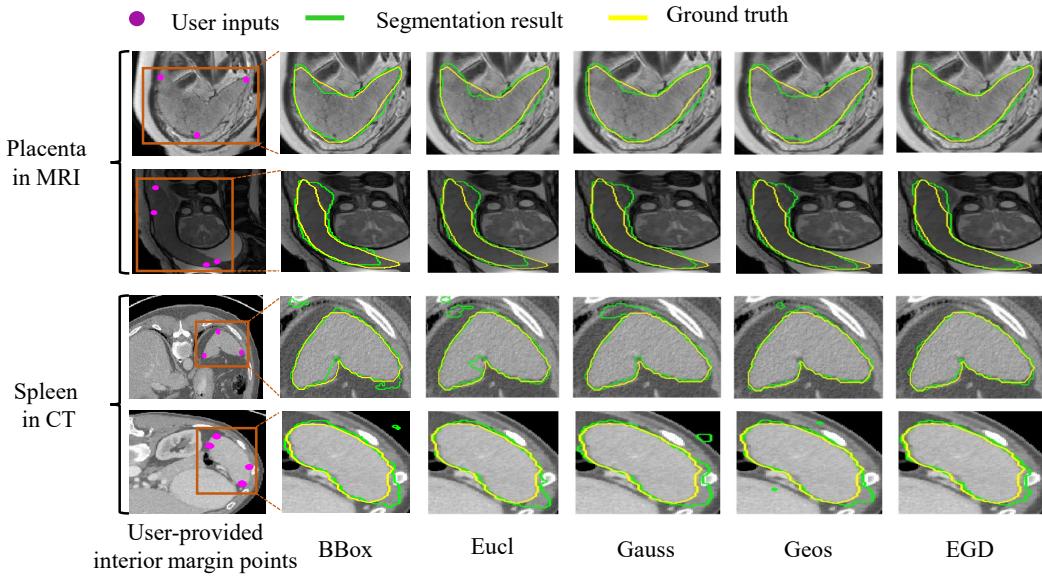


Figure 5: Visual comparison of different encoding methods for placenta and spleen segmentation, in the first stage of our method. The first column shows the input image with user-provided interior margin points (fuchsia). The other column show the initial results.

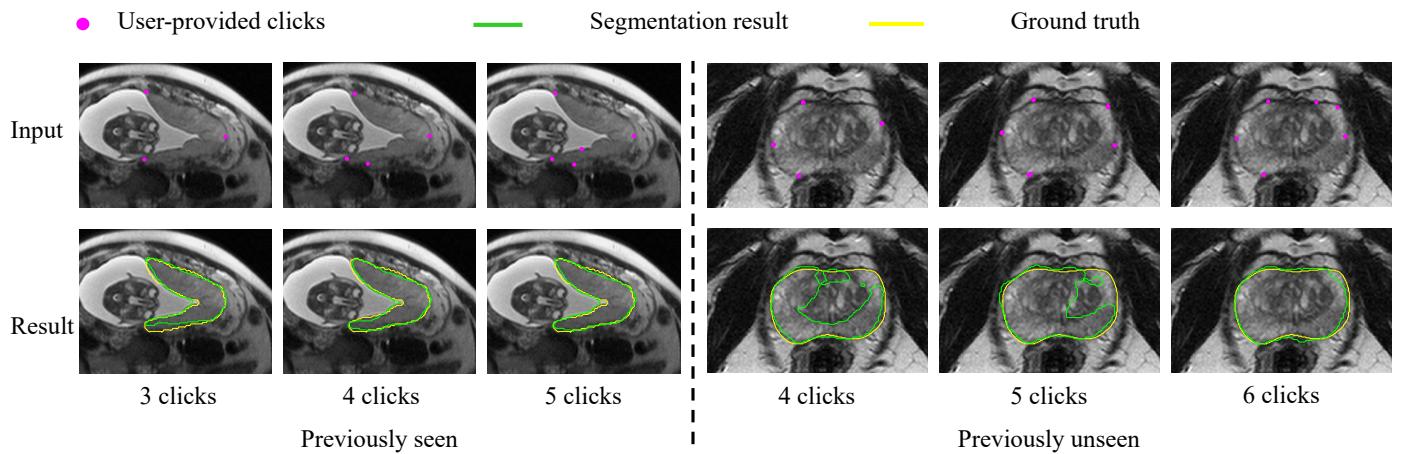


Figure 6: Effect of different number of start interior margin points for segmentation of a placenta (seen object) and prostate (unseen object) with a complex shape. The first row shows the input image with different numbers of interior margin points. The second row shows the segmentation results.

Method	“Placenta” from MRI		“Spleen” from CT		Time (s)
	Dice (%)	ASS D (pixels)	Dice (%)	ASSD (pixels)	
Bbox	85.53±7.16	4.37±3.05	91.36±4.69	3.76±1.71	-
Eucl	87.56±5.98	3.42±2.30	93.58±6.98	2.29±1.13	0.001
Gauss	87.91±6.18	3.56±2.43	93.22±3.32	2.29±1.52	0.001
Geos	87.17±6.38	3.62±1.01	94.02±7.23	2.13±0.94	0.003
EGD	88.10±4.47*	3.33±2.19	95.08±3.23*	2.25±1.28	0.004

Table 2: Quantitative comparison of different encoding methods for placenta and spleen segmentation with the same set of interior margin points. The results are based on the initial segmentation of our framework. * denotes p -value < 0.05 when comparing with the second place method.

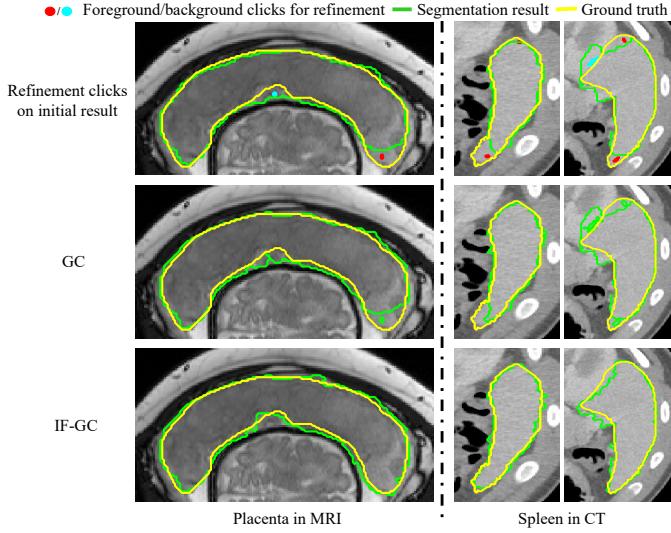


Figure 7: Visual comparison of GC and IF-GC. The first row shows the user clicks for refining initial segmentation result. The other rows show the refined results by GC and IF-GC, respectively. The results are based on the same set of user clicks for refinement. (GC: naive Graph Cuts, IF-GC: information fusion followed by Graph Cut)

provided interior margin points, respectively. We compared the proposed EGD with BBox, Eucl (with a threshold), Gauss (with a sigma), Geos (with a threshold) with same user-provided interior margin points, respectively. Note that, the parameters of Eucl, Gauss and Geos were respectively optimized for comparison, and more details of these optimal parameters are listed in Section 8 (appendix). It can be observed that the EGD transform can guide CNN to obtain more accurate segmentation results than the other encoding methods. Table 2 lists the quantitative evaluation results of different encoding methods for placenta and spleen. It can be observed that our context-aware and parameter-free encoding method of EGD consistently outperforms the others. The computation time for EGD in 2D is less than 0.05s, which gives real-time response. Fig. 6 shows the effect of different number of interior margin points for initial segmentation of challenging cases with complex shapes.

3.2.3. Interactive refinement by 2D information fusion between initial segmentation and additional clicks

Fig. 7 shows examples of interactive refinement of placenta and spleen segmentation using different refinement methods. The first row shows the initial segmentation obtained in stage 1 of our framework. Based on the initial segmentation, we further use additional clicks to obtain refined results. We compared the refined results between naive Graph Cuts (GC) and information fusion followed by Graph Cuts (IF-GC) using the same set of user clicks. Following the implementation in BIFSeg (Wang et al., 2018a), the naive Graph Cuts takes the initial segmentation probability map and the user interactions (background and foreground seeds) as inputs and is solved by max-flow. A

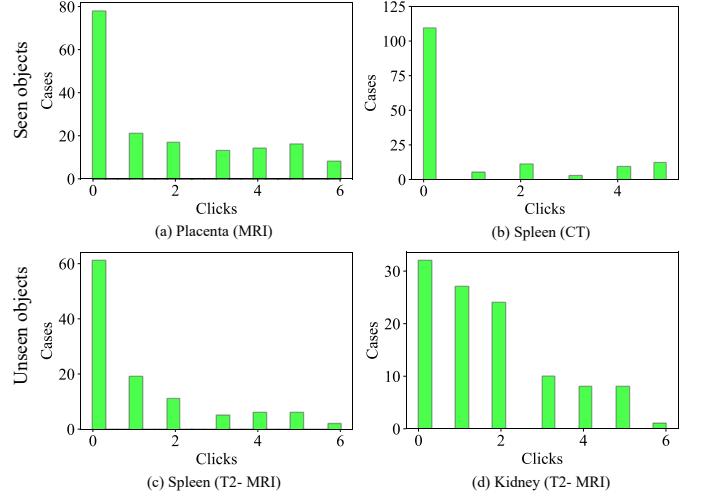


Figure 8: Histogram of number of refinement clicks required by MIDeepSeg for different objects. Placenta (MRI) and spleen (CT) in the first row are seen objects, while spleen (T2-MRI) and Kidney (T2-MRI) in the second row are previously unseen objects.

python implementation is publicly available in the SimpleCRF toolkit⁸. The performance on placenta and spleen segmentation is listed in Table 3, where the first two rows show that our method in the first stage already largely outperformed automatic segmentation with the same network structure. The last two rows demonstrate that our IF-GC achieved higher accuracy than naive Graph Cuts with the same set of user clicks for refinement in the second stage.

We further investigated the number of refinement clicks for segmentation of different objects using MIDeepSeg and plotted the histogram of refinement click number in Fig. 8. We can find that a large number of testing cases do not need additional clicks to achieve accurate results and just a few challenging cases need more than 4 clicks for refinement.

3.2.4. Comparison with other interactive methods

We compared MIDeepSeg with DeepIGeoS (Wang et al., 2018b), Graph Cuts (Boykov and Jolly, 2001), Random Walks (Grady, 2006), SlicSeg (Wang et al., 2016b), DIOS (Xu et al., 2016), DeepGrabCut (Xu et al., 2017) and DEXTR (Maninis et al., 2018) for placenta and spleen segmentation, respectively. Fig. 9 shows a visual comparison between these methods for 2D placenta segmentation. The first row shows the initial interactions and the initial segmentation results, and the second row shows the final results and all user interactions after refinement. It shows that MIDeepSeg can get a good result with only fewer user clicks, while the others need more interactions. The quantitative comparison of these methods based on placenta and spleen results as presented in Table 4. It shows that MIDeepSeg achieves higher accuracy than the other interactive segmentation methods and it needs less user

⁸<https://github.com/Hilab-git/SimpleCRF>

Method	“Placenta” from MRI			“Spleen” from CT		
	Dice (%)	ASSD (pixels)	Dice (%)	ASSD (pixels)		
Auto	79.76±15.33	8.94±11.2	90.09±10.2	10.50±13.9		
Stage 1 result	88.10±4.47	3.33±2.19	95.08±3.23	2.25±1.28		
Refined by GC	88.41±5.33	3.14±2.39	95.46±3.19	2.06±1.16		
Refined by IF-GC	89.21±4.37*	2.87±1.89*	95.79±3.07*	1.84±0.86		

Table 3: Quantitative comparison of different refinement methods for placenta and spleen segmentation with the same set of clicks. GC: naive Graph Cuts; IF-GC: information fusion followed by Graph Cuts. * denotes significant difference from GC (p -value < 0.05).

Method	“Placenta” from MRI				“Spleen” from CT			
	Dice (%)	ASSD (pix)	Times (s)	Points (pix)	Dice (%)	ASSD (pix)	Times (s)	Points (pix)
Graph Cuts	87.02±5.20	3.12 ±0.42	30.1±10.9	265.0±103.6	95.27±4.36	1.30 ±0.42	21.2±7.7	335.1±91.7
Random Walks	87.02±4.58	2.95 ±2.66	33.9±34.6	374.3±114.2	95.51 ±1.59	1.45±2.66	20.1±7.9	218.4±69.0
SlicSeg	87.63±5.71	3.00 ±0.39	25.8±11.5	189.3±81.2	95.18±4.70	1.23±0.39	20.1±8.2	254±77.5
DeepIGeoS	87.96±5.16	3.89 ±2.74	12.0±8.0	90.6 ± 95.2	96.39±2.22	1.71±2.74	6.1±4.8	31.1±52.4
DeepGrabCut	86.74±7.03	4.18±2.89	4.2±2.8	2.0±0	92.54±3.36	2.43±1.56	3.8±1.5	2.0±0
DIOS	87.48±6.31	4.03±2.52	15.3±13.0	12.4±5.7	94.85±2.79	2.06±1.47	7.6±2.7	7.8±4.7
DEXTR	88.77±4.83	3.07 ±2.25	8.9±3.7	7.2±3.3	94.18±3.25	2.67±1.36	5.9±3.9	5.6±2.7
MIDeepSeg	89.63±4.15 *	2.69±1.75 *	6.40±3.1	5.75±2.1	96.93±1.43 *	1.18±0.44 *	4.76±2.0	4.85±1.6

Table 4: Quantitative comparison of 2D placenta and spleen segmentation by different interactive methods in terms of Dice, ASSD, user time and number of interaction points. * denotes p -value < 0.05 when comparing with the second place method.

time and a smaller number of interaction points than the others except for DeepGrabCut. Note that DeepGrabCut does not allow additional user interactions for refinement, which caused the lowest accuracy among the compared methods. This demonstrates that our method is very efficient to obtain highly accurate segmentation results. (Also see supplementary video part A)

3.2.5. Deal with previously unseen 2D objects

To investigate the performance and generalizability of MIDeepSeg on previously unseen objects, we compared MIDeepSeg with existing methods with good generalizability to different objects: Graph Cuts (Boykov and Jolly, 2001), Random Walks (Grady, 2006), SlicSeg (Wang et al., 2016b), DIOS (Xu et al., 2016), DeepGrabCut (Xu et al., 2017) and DEXTR (Maninis et al., 2018). For MIDeepSeg, DIOS, DeepGrabCut and DEXTR, we used the model that was only trained with placenta images (T2-weighted MRI) to segment four previously unseen organs (i.e., kidney, spleen, prostate and fetal brain) in a range of modalities, as listed in Table 1. Fig. 10 shows examples of segmentation results of previously unseen objects by MIDeepSeg. The first row shows the initial interactions and initial segmentation results. In the second row, all interactions and final segmentation results are presented. It can be observed that MIDeepSeg can obtain a good result on unseen organs with only few user clicks. The quantitative comparison of these methods based on final results as presented in Fig. 11. It shows that MIDeepSeg takes noticeably less user time and interactions with similar or higher accuracy compared with the other interactive segmentation methods. What is more, it can be observed that MIDeepSeg can deal with different types of previously unseen image modalities and organs very well without any additional training or fine-tuning. We further studied the number of refinement clicks for Kidney (T2-MRI) and spleen (T2-

MRI) segmentation using MIDeepSeg and plotted the histogram of refinement click number in Fig. 8. We can find that although these objects are not present in the training set, our method requires no or only few clicks for refinement to obtain accurate results. (See supplementary video part B)

3.3. Interactive segmentation of 3D volumes

3.3.1. Data

Firstly, we validated the performance of MIDeepSeg on 3D brain tumor core segmentation from contrast-enhanced T1-weighted images. We used the BraTS2018⁹ training set which consists of 285 cases with four modalities: FLAIR, T1ce, T1 and T2. All images had been skull-stripped and resampled to an isotropic resolution of 1mm × 1mm × 1mm. We used 170 and 47 T1ce cases for training and testing, respectively. Manual segmentations were used as the ground truth.

Then, we validated the generalizability of MIDeepSeg with three tasks of segmentation of unseen objects: 1) Whole brain tumor in FLAIR from BraTS2018, from which we randomly selected 60 cases for testing. 2) Kidney in CT from KiTS2019¹⁰ dataset, where we randomly selected 15 cases (include 30 kidneys with or without tumor) to test. 3) Left ventricular in MRI from ACDC¹¹, where we randomly selected 30 cases. The testing data of KiTS and ACDC were resampled to an isotropic resolution of 1mm × 1mm × 1mm. All data sets for training and testing are listed in Tabel 5. To deal with 3D objects at different scales, we resized the cropped sub-region and cue map to 64 × 96 × 96.

⁹<https://www.med.upenn.edu/sbia/brats2018.html>.

¹⁰<https://kits19.grand-challenge.org>

¹¹<https://acdc.creatis.insa-lyon.fr/>

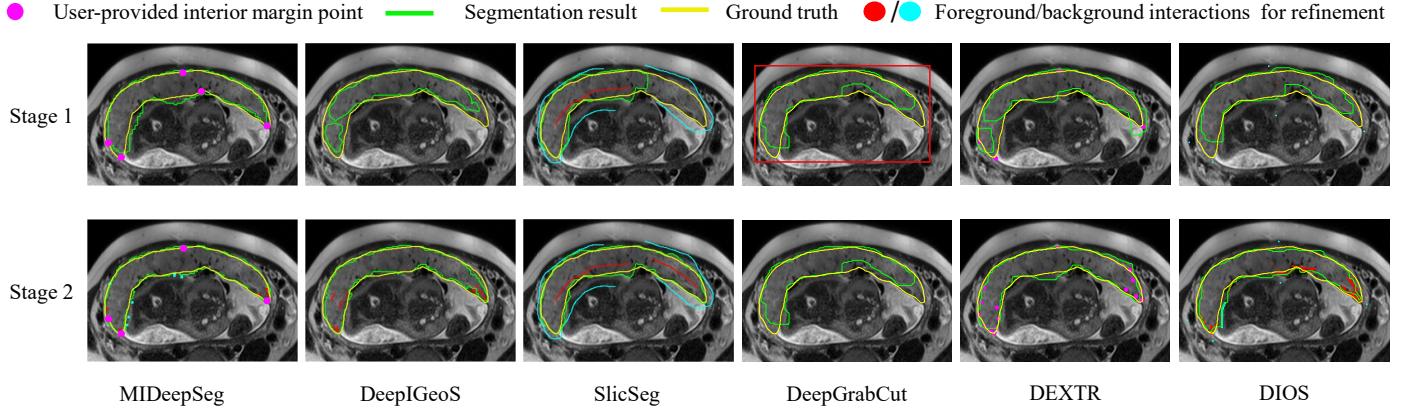


Figure 9: Visual comparison of MIDeepSeg and other interactive methods for 2D placenta segmentation. The first row shows the initial segmentation results with or without initial interactions. And the second row shows the final results after refinement.

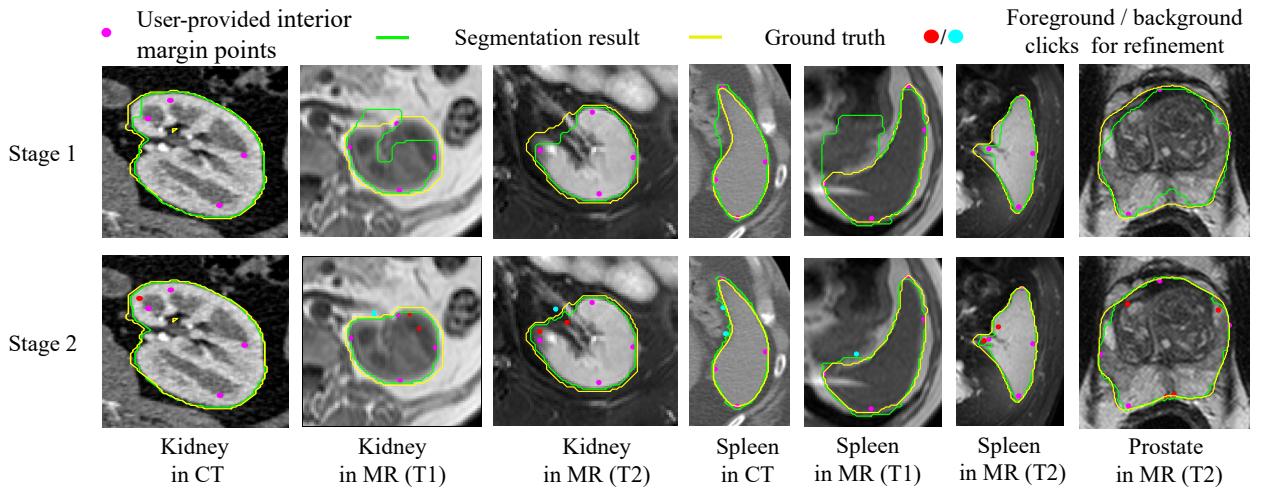


Figure 10: Some examples of 2D unseen organ segmentation results by MIDeepSeg. The first row shows the initial user interactions and the initial segmentation. The second row shows all user interactions and final segmentation results. Note that the model is only trained with placenta in T2 MRI.

Object	Modality	N_{train}	N_{test}	Dataset
Tumor core	MRI (T1ce)	170	47	BraTS2018
Whole tumor	MRI (FLAIR)	0	60	BraTS2018
Kidney	CT	0	30	KiTS
Left ventricular	MRI (T2)	0	30	ACDC

Table 5: Datasets used for training and testing for 3D experiments. N_{train} and N_{test} denote the number of volumes for training and testing, respectively.

3.3.2. Initial segmentation based on EGD-based cue map and 3D CNN

To validate our proposed EGD transform for interior margin points encoding in 3D volumes, we compared it with BBox, Eucl, Gauss, and Geos with their respectively optimized parameters, respectively. In this stage, the same set of interior margin points provided by the user were used for these methods. Fig. 12 shows the initial segmentation results obtained by CNN guided by different encoding methods. It shows that EGD transform can guide CNN to achieve more noticeable improvement from BBox

compared with the other encoding methods. Table 6 lists the quantitative evaluation results of different encoding methods for tumor core segmentation from T1ce images. It can be observed that our context-aware and parameter-free encoding method of EGD consistently outperforms the others with 87.00% in term of Dice and 1.46 mm in term of ASSD for tumor core, respectively. Despite that EGD takes more time (0.24s) for interaction encoding than the others, it is still very efficient in practice.

3.3.3. Interactive refinement by 3D information fusion between initial segmentation and additional clicks.

Based on the above initial segmentation obtained by our method, we further use additional clicks to obtain refined results. We compared the refined result between naive Graph Cuts (GC) and the proposed information fusion followed by Graph Cuts (IF-GC) with the same set of user refinement clicks. The performance on tumor core segmentation is listed in Table 7, showing that the information fusion achieves higher accuracy than the other variants.

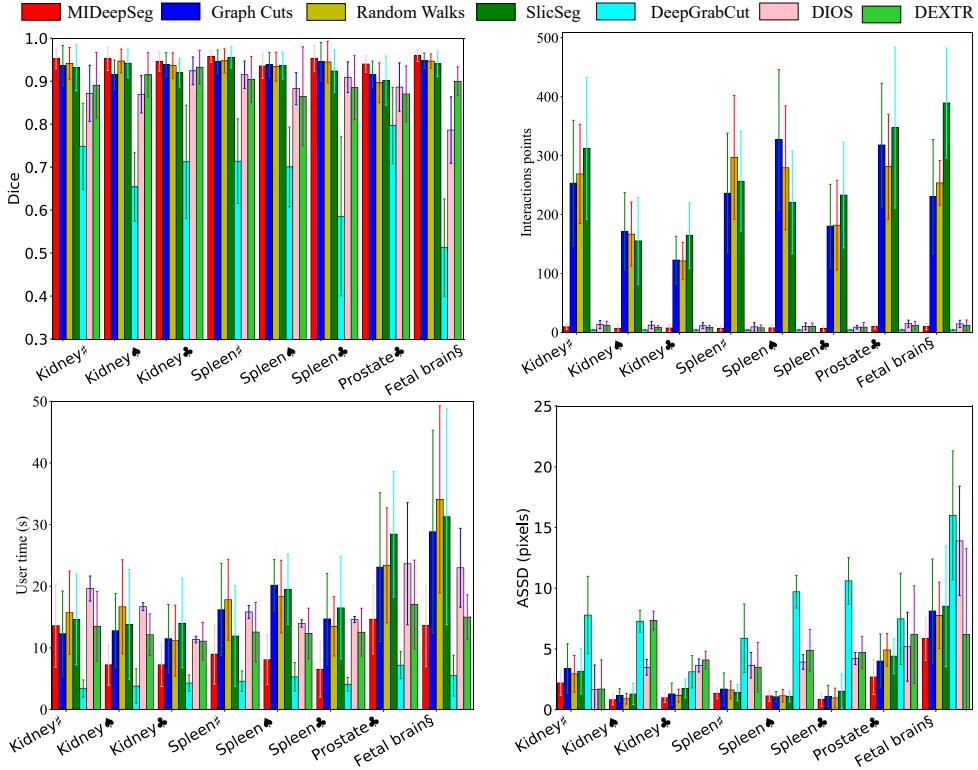


Figure 11: Dice, ASSD, user time and interaction points of different interactive segmentation methods for unseen objects. \sharp , \S , \spadesuit , \clubsuit denote CT images, ultrasound images, T1-weighted MR images and T2-weighted MR images, respectively. All these organs are previously unseen in the training set.

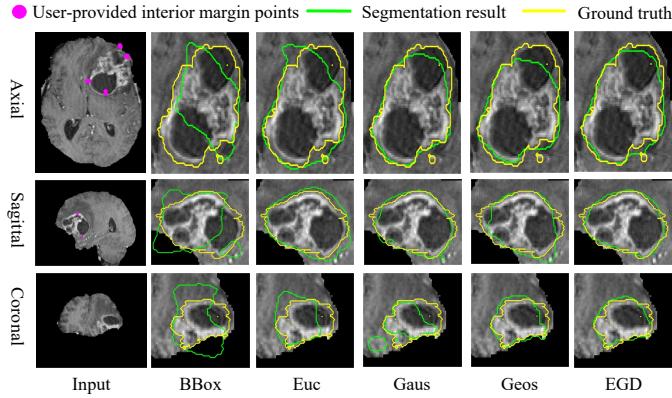


Figure 12: Visual comparison of different encoding methods for 3D tumor core segmentation, which is based on the initial segmentation obtained in the first stage. All these methods used the same interior margin points and inferred bounding box for the input image.

Fig. 13 shows an example of tumor core segmentation by different refinement methods. It can be observed that IF-GC refined the result more accurately than GC with the same set of clicks for refinement.

3.3.4. Comparison with other interactive methods

Fig. 14 shows a visual comparison between MIDeepSeg, 3D Graph Cuts (Boykov and Jolly, 2001), ITK-SNAP (Yushkevich et al., 2006), and 3D versions of DeepI-

Method	Dice (%)	ASSD (mm)	Time (s)
BBox	82.32 \pm 12.03	2.17 \pm 1.53	-
Eucl	85.25 \pm 9.78	1.71 \pm 1.20	0.05
Gauss	85.90 \pm 9.11	1.64 \pm 1.19	0.06
Geos	86.42 \pm 8.91	1.60 \pm 1.15	0.15
EGD	87.00\pm9.11*	1.46\pm1.14*	0.24

Table 6: Quantitative comparison of different encoding methods for 3D tumor core segmentation with the same set of interior margin points. The results are based on the initial segmentation (Stage 1) of our framework. * denotes p -value < 0.05 when comparing with the second place method.

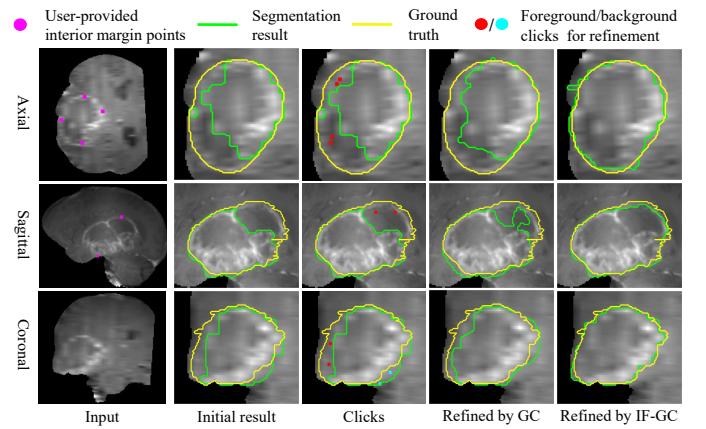


Figure 13: Visual comparison of different refinement methods for 3D tumor core segmentation. These refinement methods are compared with the same initial segmentation with the same set of clicks.

Method	Dice (%)	ASSD (mm)
Auto	78.08±13.56	2.78±2.22
Stage 1 result	87.00±9.11	1.46±1.14
Refined by GC	87.44±8.31	1.37±1.15
Refined by IF-GC	88.21±7.31*	1.28±0.94

Table 7: Quantitative comparison of different refinement methods for 3D tumor core segmentation with the same set of refinement clicks. The segmentation before refinement is obtained by MIDeepSeg in stage 1. GC: 3D Graph Cuts; IF-GC: information fusion followed by Graph Cuts.* denotes significantly higher performance than GC (p -value < 0.05).

GeoS (Wang et al., 2018b), DIOS (Xu et al., 2016), DeepGrabCut (Xu et al., 2017) and DEXTR (Maninis et al., 2018). It can be found that MIDeepSeg needs only few interior margin points as the initial interactions, but its initial segmentation is more accurate and it requires fewer user clicks to get an accurate final result. The quantitative comparison of these methods based on the final result is presented in Table 8. It shows that MIDeepSeg achieved significantly higher accuracy than the others. Additionally, MIDeepSeg takes 29s in average for the entire 3D interactive segmentation process for tumor core, which is less than the other methods except for DeepGrabCut. (See supplementary video part C)

3.3.5. Deal with previously unseen 3D objects

To investigate the generalizability of MIDeepSeg on previously unseen 3D objects, we used the 3D CNN model trained in the task of tumor core segmentation from T1ce images to deal with three previous unseen objects and modalities: whole tumor in FLAIR; kidney in CT and left ventricular in MRI, as listed in Table 5. Two users used MIDeepSeg and two existing methods with good generalizability including ITK-SNAP (Yushkevich et al., 2006) and 3D versions of Graph Cuts (Boykov and Jolly, 2001), DIOS (Xu et al., 2016), DeepGrabCut (Xu et al., 2017) and DEXTR (Maninis et al., 2018) to segment these objects. Fig. 15 shows some examples of 3D whole tumor, kidney and left ventricular segmentation using MIDeepSeg. It can be found that accurate results are obtained for different types of unseen objects by using MIDeepSeg with few clicks. Quantitative evaluation results are presented in Fig. 16. It shows that MIDeepSeg achieves similar or higher accuracy compared with 3D Graph Cuts, ITK-SNAP, DeepIGeoS, DIOS, DeepGrabCut and DEXTR. However, MIDeepSeg takes notable less user time to achieve the results. (See supplementary video part D)

4. Discussion

Though some recent works (Wang et al., 2018b; Zhou et al., 2019a; Wang et al., 2020; Liao et al., 2020) on deep learning-based interactive segmentation have shown good performance, it is a great challenge for current CNNs to generalize well on previously unseen object classes, as

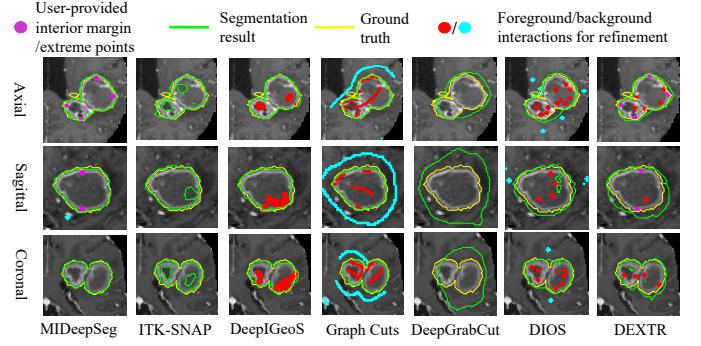


Figure 14: Visual comparison of 3D tumor core segmentation using MIDeepSeg, DeepIGeoS, 3D Graph Cuts and ITK-SNAP.

Method	Dice (%)	ASSD (mm)	Time (s)
3D Graph Cuts	78.91±14.98	3.46±5.10	99.4±36.7
ITK-SNAP	82.34±11.42	1.99 ±1.31	173.0±75.5
DeepIGeoS	82.47±10.78	3.34±3.81	82.2±44.7
DeepGrabCut	82.58±11.79	2.89±2.37	10.5±8.3
DIOS	83.34±10.47	2.57±1.79	67.5±23.6
DEXTR	86.39±9.03	1.59±1.11	34.7±18.6
MIDeepSeg	88.71±7.00*	1.24 ±0.88*	28.6 ±12.2

Table 8: Quantitative evaluation of 3D tumor core segmentation by different interactive methods in terms of Dice, ASSD and user time, respectively. * denotes p -value < 0.05 when comparing with the second place method.

they rely on annotated image to learn directly (Masood et al., 2015). For medical images, annotated images are very precious and scarce since accurate annotations require both expertise and time to obtain. This limits the performance of CNNs to deal with unseen objects that are not present in training set. Compared with traditional CNNs (Ronneberger et al., 2015; Çiçek et al., 2016) and transfer learning (Wang et al., 2018a; Tajbakhsh et al., 2016), the major advantage of ours proposed framework is that it can segment unseen objects without re-training or fine-tuning. Therefore, it reduces the burden for collecting and annotating data noticeably, and can be applied to segment or annotate unseen objects directly. Compared with DeepIGeoS (Wang et al., 2018b) and BIFSeg (Wang et al., 2018a), MIDeepSeg only requires few clicks as input and has a higher generalizability.

A big challenge for existing deep learning frameworks is that they hardly generalize on previously unseen objects, and they require additional re-training or fine-tuning for segmentation of new targets. BIFSeg (Wang et al., 2018a) uses image-specific fine-tuning to improve the generalization of CNN, but it requires fine-tuning for each test image, which is a time and memory consuming process. Based on our proposed interior margin points, EGD transform and information fusion, MIDeepSeg can deal with different types of unseen medical images without additional fine-tuning or training.

Despite the simple implementation, our EGD has not been proposed earlier for user interaction encoding, and it has two important differences from geodesic distances: First, EGD is parameter-free with higher generalizabil-

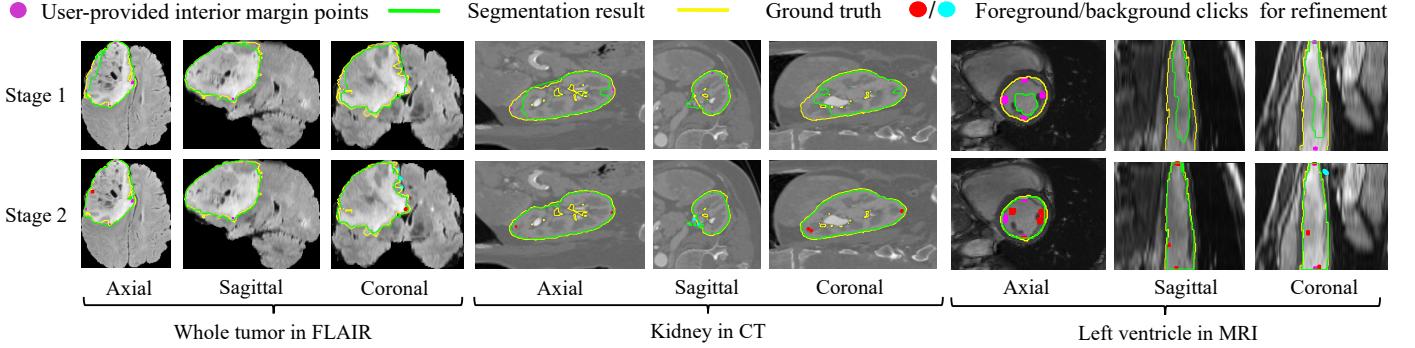


Figure 15: Three examples of segmentation of 3D unseen objects using MIDeepSeg. Note that only tumor core in T1ce images were used for training.

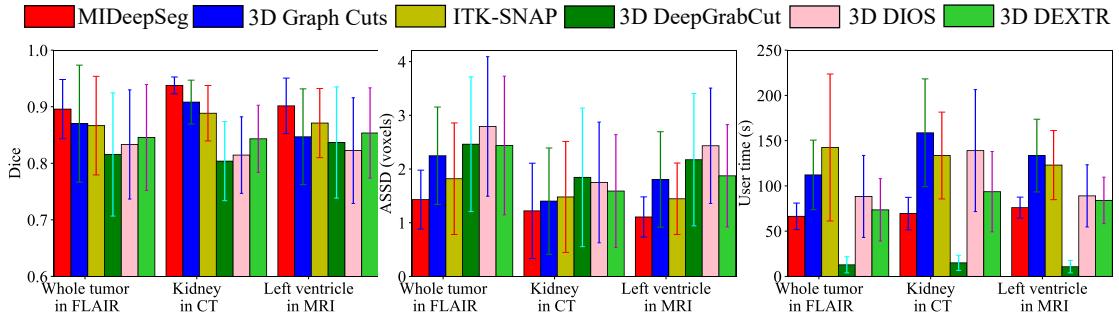


Figure 16: Dice score, ASSD and user time of different interactive methods for 3D unseen objects segmentation.

ity. The geodesic distance method (Wang et al., 2018b) requires a user-defined threshold to make sure that the interactions will affect a local region, which reduces its generalizability as different images may require different threshold values. In contrast, our EGD does not require such a parameter, and it can be applied to different images without some specific adjustment, making it a simple and effective method with a wider utility. Second, the EGD naturally outputs a probabilistic map, which gives can be used as the probability of a pixel belonging to the foreground or background indicated by the user interactions. This probabilistic view allow it to be seamlessly integrated in to a conditional random field formulation for refinement.

The computation time of our EGD listed in Table 2 and Table 6 show that it takes less than 0.05s and 0.25s for 2D and 3D images respectively, which is acceptable for fast response of user interactions. We also studied the computation time of other stages of our method: the inference time for 2D and 3D CNNs was 0.008s and 0.04s, respectively. The CRF optimization time was 0.015s and 0.5s for 2D and 3D images, respectively. The entire user time was around 8-12s for 2D cases (Fig. 11) and 60-80s for 3D cases (Fig. 16). Therefore, our method is efficient for interactive segmentation of unseen objects.

In our experiments, we found that our refinement method based on calibrated probability maps and Graph Cuts worked well in various cases for different organs in a range of modalities. The advantages include: 1) the refinement step is decoupled from the initial segmentation step

based on CNNs, thus is ready-to-use as a general refinement tool for interactively correcting segmentation results obtained by different networks and for unseen objects. 2) It is computational efficient, and allows real-time response of user interactions, which is highly desired for improving the user experience of interactive segmentation. 3) The user interactions are used as hard constraints, which ensures that points given by users will have their desired labels after refinement. A potential issue is that in complex cases a relatively large number of clicks are needed to obtain accurate results. However, in practice, our method is easy to use and efficient in dealing with different unseen objects, as shown by the experimental results.

A general problem for interactive segmentation is that the result may depend on knowledge and experience of the user, as the user refines the segmentation until it is visually acceptable, where the standard may be subjective. However, our method has some requirements on the user interactions: in the first stage, the interactions need to be given near the inner side boundary, and in the second stage the interactions are only given in incorrect regions, where for most cases, the incorrect region is small, leading to the range of clicks provided by different users limited. Therefore, inter-user variation of our method is small. As our method does not require the user to provide clicks exactly on the boundary or extreme points, our interior margin points tolerate inaccurate clicks, which is more user-friendly. As shown in Fig. 10, in the first column the interior margin points are not accurate and far

away from the boundary, and in the fourth column the top point is also inaccurate and even clicked in the background, but they still lead to good initial segmentation results. It further demonstrates the robustness and generalization of MDeepSeg.

Recently, some works (Sourati et al., 2019; Rupprecht et al., 2018; Zhou et al., 2019b; Song et al., 2018) used Fisher information, natural language, active learning and deep reinforcement learning to develop an intelligent interactive segmentation or annotation tool. In the future, it is of interest to use active learning (Top et al., 2011) and deep reinforcement learning (Liao et al., 2020) and uncertainty estimation (Wang et al., 2020) to provide guidance on user interactions for refinement, which has a potential to further improve the efficiency of interactive segmentation.

5. Conclusion

In this paper, we proposed a deep learning-based interactive framework with good generalizability to unseen objects for medical images segmentation and it only requires few clicks as user inputs. A novel context-aware and parameter-free encoding method was proposed to encode user interactions to guide CNN for a good initial segmentation. Based on the encoding method, we also proposed an effective refinement way for improving the accuracy of the segmentation results. The framework is designed to improve the generalizability to unseen objects, which is highly desired for deep learning-based models. Experiments on segmenting a wide range of previously seen and unseen organs or lesions from various 2D and 3D images show that: 1) Our interior margin points and EGD transform-based framework outperforms existing deep learning-based interactive segmentation tools in terms of accuracy and efficiency. 2) The proposed framework generalizes well on previously unseen objects. It could be used as an annotation tool to obtain segmentation masks of a range of objects more efficiently with high accuracy.

6. Acknowledgment

This work was supported by the National Natural Science Foundations of China [61901084 and 81771921] funding, key research and development project of Sichuan province, China [No. 20ZDYF2817]. This work also was supported by the Wellcome Trust [WT101957, 203148/Z/16/Z], and the Engineering and Physical Sciences Research Council (EPSRC) [NS/A000027/1, NS/A000049/1]. TV is supported by a Medtronic / Royal Academy of Engineering Research Chair [RCSR18194].

7. Supplementary Video

The supplementary video of this paper can be found at: <https://www.youtube.com/watch?v=eq-tqlJnckE>.

8. Appendix

In Section 3.2.2, we compared our parameter-free interaction encoding method EGD with several existing encoding methods: Euclidean distance transform, Gaussian heatmap and Geodesic distance transform. Each of these alternative methods relies on a hyper-parameter. Here we report the details of finding the optimal hyper-parameter values for these methods. The dataset and CNN structure were the same as those used in Section 3.2.2.

8.1. Effects of different threshold for Euclidean distance transforms

DeepIOS (Xu et al., 2016) and ISLD (Li et al., 2018) use euclidean distance to encode user interactions. The distance transforms are truncated by a threshold value for efficient representation. Then, the truncated encoding map is concatenated with the input image to guide the CNN to obtain a segmentation result. Following these works, we rescaled all encoding maps to [0, 1] and truncated them with four different threshold values: 0.2, 0.4, 0.6 and 0.8, respectively. We used their corresponding cue maps to guide CNN to segment placenta in MRI and spleen in CT, respectively. Table 9 lists the corresponding quantitative evaluation results. We found that the optimal threshold value was 0.6 for the placenta, and 0.4 for the spleen. We therefore used these optimal values for DeepIOS (Xu et al., 2016) and ISLD (Li et al., 2018) in the experiment.

8.2. Effects of different sigma for Gaussian distance transforms

DEXTR (Maninis et al., 2018) and DELSE (Wang et al., 2019) use Gaussian distance transforms to deal with user interactions. To generate the gaussian heat-map, we need to determine a proper sigma for each task. In this work, we used four different values of sigma (i.e., 3, 6, 9 and 12) to generate gaussian heatmap for placenta and spleen segmentation, respectively. Their corresponding heatmaps were concatenated with the input image to guide the CNN to achieve an initial segmentation. The results are listed in Table 10. It can be found the optimal sigma value was 9 for the placenta and the spleen. We therefore used these optimal values for DEXTR (Maninis et al., 2018) and DELSE (Wang et al., 2019) in the experiment.

8.3. Effects of different threshold for Geodesic distance transforms

DeepIGeoS (Wang et al., 2018b) encodes user interactions by geodesic distance transforms. Similarly to Euclidean distance transforms, we truncated the geodesic distance maps with different threshold values: 0.2, 0.4, 0.6 and 0.8, respectively. These geodesic distance maps were concatenated with the input image to guide the CNN to achieve an initial segmentation. The quantitative evaluation results are listed in Table 11. It can be observed the optimal threshold value was 0.6 for the placenta, and 0.4 for the spleen. We therefore used these values for DeepIGeoS (Wang et al., 2018b) in the experiment.

Threshold	“Placenta” from MRI		“Spleen” from CT	
	Dice (%)	ASSD (pixels)	Dice (%)	ASSD (pixels)
0.2	86.82±6.65	3.74±2.58	92.88±3.81	2.47±1.58
0.4	86.91±7.20	3.62±2.47	93.58±6.98	2.29±1.13
0.6	87.56±5.98	3.42±2.30	92.62±3.13	2.30±1.16
0.8	87.13±6.26	3.57± 2.29	93.04±5.00	2.33±1.08

Table 9: Quantitative evaluation of different threshold values in Euclidean distance transforms for placenta and spleen segmentation with the same set of real-users clicked points. The results are based on the initial segmentation of our framework.

σ	“Placenta” from MRI		“Spleen” from CT	
	Dice (%)	ASSD (pixels)	Dice (%)	ASSD (pixels)
3	86.96±6.89	3.74±2.58	92.76±4.16	2.38±1.52
6	87.78±6.15	3.62±2.47	92.91±6.98	2.21±1.45
9	87.91±6.18	3.42±2.30	93.22±3.32	2.29±1.51
12	86.97±6.79	3.57± 2.29	92.69±3.70	2.24±1.26

Table 10: Quantitative evaluation of different σ in Gaussian distance transforms for placenta and spleen segmentation with the same set of real-users clicked points. The results are based on the initial segmentation of our framework.

References

- Acuna, D., Ling, H., Kar, A., Fidler, S., 2018. Efficient interactive annotation of segmentation datasets with polygon-rnn++, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 859–868.
- Bai, X., Sapiro, G., 2009. Geodesic matting: A framework for fast interactive image and video segmentation and matting. *International Journal of Computer Vision* 82, 113–132.
- Benenson, R., Popov, S., Ferrari, V., 2019. Large-scale interactive object segmentation with human annotators, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11700–11709.
- 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 International Conference on Computer Vision, IEEE. pp. 105–112.
- Castrejon, L., Kundu, K., Urtasun, R., Fidler, S., 2017. Annotating object instances with a polygon-rnn, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5230–5238.
- Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 834–848.
- Ciçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3d u-net: learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 424–432.
- Criminisi, A., Sharp, T., Blake, A., 2008. Geos: Geodesic image segmentation, in: European Conference on Computer Vision, Springer. pp. 99–112.
- Grady, L., 2006. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis Machine Intelligence* , 1768–1783.
- Hao, J., Scott, L., Brian, C., Long, P., Ong, M.S.H., 2019. Multi-Seg : Semantically Meaningful , Scale-Diverse Segmentations from Minimal User Input, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 662–670.
- Harris, C.G., Stephens, M., et al., 1988. A combined corner and edge detector. *Alvey Vision Conference* 15, 10–5244.
- Hu, Y., Soltoggio, A., Lock, R., Carter, S., 2019. A fully convolutional two-stream fusion network for interactive image segmentation. *Neural Networks* 109, 31–42.
- Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A.V., Criminisi, A., Rueckert, D., Glocker, B., 2016. Deepmedic for brain tumor segmentation, in: International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries, Springer. pp. 138–149.
- Khan, S., Shahin, A.H., Villafruela, J., Shen, J., Shao, L., 2019. Extreme points derived confidence map as a cue for class-agnostic interactive segmentation using deep neural network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 66–73.
- Lafferty, J., McCallum, A., Pereira, F.C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: International Conference on Machine Learning, pp. 282–289.
- Li, Z., Chen, Q., Koltun, V., 2018. Interactive image segmentation with latent diversity, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 577–585.
- 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 Conference on Computer Vision and Pattern Recognition (CVPR).
- Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42, 60–88.
- 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 Conference on Computer Vision and Pattern Recognition, pp. 616–625.
- Marsh, H.E., 2013. Beyond thick versus thin: mapping cranial vault thickness patterns in recent homo sapiens. University of Iowa .
- Masood, S., Sharif, M., Masood, A., Yasmin, M., Raza, M., 2015. A survey on medical image segmentation. *Current Medical Imaging Reviews* 11, 3–14.
- Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V., 2017. Extreme clicking for efficient object annotation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4930–4939.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, pp. 8024–8035.
- Price, B.L., Morse, B., Cohen, S., 2010. Geodesic graph cut for interactive image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 3161–3168.
- 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

Threshold	“Placenta” from MRI		“Spleen” from CT	
	Dice (%)	ASSD (pixels)	Dice (%)	ASSD (pixels)
0.2	86.81±6.33	4.00±2.72	92.96±3.13	2.23±0.97
0.4	87.08±5.35	3.59±2.05	94.02±3.23	2.13±0.93
0.6	87.17±6.38	3.55±2.62	93.06±4.67	2.17±0.96
0.8	86.93±7.13	3.62± 2.45	92.91±3.33	2.23±0.92

Table 11: Quantitative evaluation of different thresholds in Geodesic distance transforms for placenta and spleen segmentation with the same set of real-users clicked points. The results are based on the initial segmentation of our framework.

- from bounding box annotations using convolutional neural networks. *IEEE Transactions on Medical Imaging* 36, 674–683.
- Raju, A., Ji, Z., Cheng, C.T., Cai, J., Huang, J., Xiao, J., Lu, L., Liao, C., Harrison, A.P., 2020. User-guided domain adaptation for rapid annotation from user interactions: A study on pathological liver segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 457–467.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 234–241.
- Roth, H., Zhang, L., Yang, D., Milletari, F., Xu, Z., Wang, X., Xu, D., 2019. Weakly supervised segmentation from extreme points, in: Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention. Springer, pp. 42–50.
- Roth, H.R., Yang, D., Xu, Z., Wang, X., Xu, D., 2020. Going to extremes: Weakly supervised medical image segmentation. arXiv preprint arXiv:2009.11988 .
- Rother, C., Kolmogorov, V., Blake, A., 2004. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics* 23, 309–314.
- Rupprecht, C., Laina, I., Navab, N., Hager, G.D., Tombari, F., 2018. Guide me: Interacting with deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8551–8561.
- 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 .
- Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering* 19, 221–248.
- Song, G., Myeong, H., Mu Lee, K., 2018. Seednet: Automatic seed generation with deep reinforcement learning for robust interactive segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1760–1768.
- Sourati, J., Gholipour, A., Dy, J.G., Tomas-Fernandez, X., Kurugol, S., Warfield, S.K., 2019. Intelligent labeling based on fisher information for medical image segmentation using deep learning. *IEEE transactions on medical imaging* 38, 2642–2653.
- Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging* 35, 1299–1312.
- Top, A., Hamarneh, G., Abugharbieh, R., 2011. Active learning for interactive 3d image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 603–610.
- Vezhnevets, V., Konouchine, V., 2005. Growcut: Interactive multi-label nd image segmentation by cellular automata. *Graphicon* 1, 150–156.
- Wang, B., Liu, K.W., Prastawa, K.M., Irima, A., Vespa, P.M., Van Horn, J.D., Fletcher, P.T., Gerig, G., 2014. 4d active cut: An interactive tool for pathological anatomy modeling, in: IEEE International Symposium on Biomedical Imaging, IEEE. pp. 529–532.
- 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: International Conference on 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 Transactions on Medical Imaging* 37, 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 Transactions on Pattern Analysis and Machine Intelligence* 41, 1559–1572.
- Wang, G., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., Klusmann, M., David, A.L., Deprest, J., Vercauteren, T., Ourselin, S., 2016a. Dynamically balanced online random forests for interactive scribble-based segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 352–360.
- Wang, G., Zuluaga, M.A., Pratt, R., Aertsen, M., Doel, T., Klusmann, M., David, A.L., Deprest, J., Vercauteren, T., Ourselin, S., 2016b. Slic-seg: A minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal mri in multiple views. *Medical Image Analysis* 34, 137–147.
- Wang, Z., Acuna, D., Ling, H., Kar, A., Fidler, S., 2019. Object instance annotation with deep extreme level set evolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7500–7508.
- Xu, N., Price, B., Cohen, S., Yang, J., Huang, T., 2017. Deep grabcut for object selection. arXiv preprint arXiv:1707.00243 .
- Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.S., 2016. Deep interactive object selection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 373–381.
- Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 31, 1116–1128.
- Zhao, F., Xie, X., 2013. An overview of interactive medical image segmentation. *Annals of the British Machine Vision Association* 2013, 1–22.
- Zhou, B., Chen, L., Wang, Z., 2019a. Interactive deep editing framework for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 329–337.
- Zhou, X., Zhuo, J., Krahenbuhl, P., 2019b. Bottom-up object detection by grouping extreme and center points, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 850–859.