Medical Image Segmentation Using Interactive Refinement

by

Diksha Goyal

A Thesis Presented in Partial Fulfillment
of the Requirement for the Degree
Master of Science

Approved April 2021 by the
Graduate Supervisory Committee:

Jianming Liang, Chair
Yalin Wang
Hemanth Kumar Demakethepalli Venkateswara

ARIZONA STATE UNIVERSITY

May 2021

ABSTRACT

Image segmentation is an important and challenging area of research in computer vision with various applications in medical imaging. Image segmentation refers to the process of partitioning an image into meaningful parts having similar attributes. Traditional manual segmentation approaches rely on human expertise to outline object boundaries in images which is a tedious and expensive process. In recent years, Deep Convolutional Neural Networks have demonstrated excellent performance in tasks such as detection, localization, recognition and segmentation of objects. However, these models require a large set of labeled training data which is difficult to obtain for medical images. To solve this problem, interactive segmentation techniques can be used to serve as a trade-off between fully automated and manual approaches. This allows a human expert in the loop as a form of guidance and refinement together with deep neural networks.

This thesis proposes an interactive training strategy for segmentation, where a robot-user is utilized during training to mimic an actual annotator and provide corrections to the predicted masks by drawing scribbles. These scribbles are then used as supervisory signals and fed to the network; which interactively refines the segmentation map through several iterations of training. Further, the conducted experiments using various heuristic click strategies demonstrate that user interaction in the form of curves inside the organ of interest achieve optimal editing performance. Moreover, by using the popular image segmentation architectures based on U-Net as base models, segmentation performance is further improved; signifying that the accuracy gain of the interactive correction conform to the accuracy of the initial segmentation map.

To my brother, Dushyant

ACKNOWLEDGEMENTS

First and foremost, I thank my family for their immense love and constant support throughout my academic career. Thank you for your relentless belief in me without which this journey would not have been possible.

I would like to express my sincere gratitude to my advisor, Dr. Jianming Liang, for his guidance and continuous support throughout my thesis. His motivation and enthusiasm helped me to work harder and sharpen my skills. Further, I thank Dr. Yalin Wang and Dr. Hemanth Kumar Demakethepalli Venkateswara for accepting to be part of my thesis committee, reviewing my work and providing me valuable comments.

I would like to thank the members of Dr. Liang's lab: Zongwei Zhou, Vatsal Sodha, Dr. Ruibin Feng, Md Mahfuzur Rahman Siddiquee, Mohammad Reza Hosseinzadeh Taher, Shivam Bajpai and Fatemeh Haghighi for providing insightful suggestions and having discussions with me.

This work has utilized the GPUs provided partially by the ASU Research Computing and partially by the Extreme Science and Engineering Discovery Environment (XSEDE) funded by the National Science Foundation (NSF) under grant number ACI-1548562.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

1.1 Background

Image segmentation consists of dividing an image into meaningful segments to distinguish different objects in an image. It plays an important role in medical imaging and computer aided diagnosis in order to separate various structures such as heart, spleen, knee, brain and blood vessel from images. This clinically useful information assist the radiologists in diagnosis, study of human anatomy, localization of pathology and treatment planning.

With the advent of deep learning, the performance of image segmentation algorithms has greatly increased. This success is attributed to the rise of neural networks with deeper architecture and the use of large annotated datasets. Collecting high-quality expert annotations demands an intensive and time-consuming labour which may not be manageable at large scales. Thus, semi-automatic segmentation methods which integrate user input to guide segmentation, appears to be an efficient alternative to mitigate the annotation effort.

There have been many different methods being proposed for interactive segmentation. In some earlier works such as GrabCut by Rother *et al.* (2004), GeoS by Criminisi *et al.* (2008), and GraphCut by Freedman and Zhang (2005), an energy functional is often minimized so that its local minimum is at the boundary of the object. In recent years, studies by Wang *et al.* (2018a), Jang and Kim (2019) have explored the interactive strategy by combining user interactions with CNNs. In these approaches users provide clicks, scribbles, points, extreme points, super-pixel an-

notations or bounding boxes as additional supervision to improve the results from automated approaches. We explore all these heuristics as well as how to encode them to minimize the number of interactions that the user provides at test time.

In this work, an interactive training strategy is proposed which improves the segmentation accuracy. The CNN is trained with user simulated inputs to edit the segmentation. Results on a prostate dataset from NCI-ISBI 2013 challenge show superior performance with the curve based user-interaction in comparison to other user feedback strategies. Moreover, using the recent state-of-the-art segmentation architecture nnU-Net by Isensee *et al.* (2018) as the base segmentation model, further performance improvement is observed using interactive training on Medical Segmentation Decathlon dataset.

The unique challenges posed by medical image analysis have suggested that retaining a human end user in deep learning enabled segmentation system, will speed-up annotations and be able to refine existing methods.

1.2 Terminology

Image segmentation approaches can be grouped into three categories.

- Manual Segmentation: Manual segmentation refers to a process where each pixel of the image is manually assigned to a class. This is done by a user by manually drawing the borders of object of interest or using techniques such as painting with a brush and marking the area covered by the object. Although this method is believed to be more accurate, it is tedious, time-consuming and prone to inter-observer and intra-observer variability.

- Fully Automatic Segmentation: Automatic segmentation methods can segment images without user interaction. The segmentation is performed using super-

vised techniques that train from manually-annotated labeled examples or unsupervised curve initialization approaches based on the minimization of the energy functional. Recently, deep learning techniques with convolutional neural networks (CNNs) have achieved best performance in many public benchmarks and challenges. Despite their success, these methods require a lot of labeled data to train, as well as a long training time and specialized hardware (GPU) to construct the model.

- Interactive Segmentation: Even though automatic image segmentation achieves impressive performance, it may still need to be refined to become accurate and robust enough for clinical use. Another approach which combines the automatic and manual approaches is called semi-automatic or interactive segmentation. This allows users to explicitly control the predictions using interactive input to indicate mis-segmentation.

1.3 System Design

In comparison to automatic segmentation, interactive segmentation provides user interaction to the network as an additional input or feedback. This way the image segmentation accuracy can be improved for practical applications since the model has a feedback control loop.

The design of such a system can be viewed as the following three modules/steps.

1. User Module

   The user provides input by placing scribbles and marking the correct boundary for the object.

2. Network Module

A segmentation architecture is adopted that takes images as input along with the user provided information in the form of clicks/scribbles.

3. Output Module

   Initial set of clicks along with the input image is passed to the network to get the initial prediction.

The user corrections are taken into account for any misclassified points and passed as input to Step 1. The above steps are performed iteratively until the user gets the satisfied result and then the process is terminated.

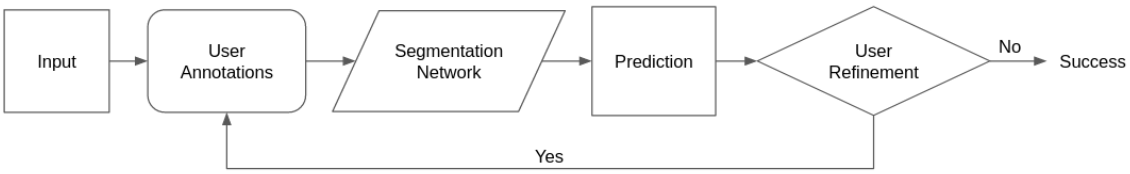

Figure 1.1: System Design

The process of an interactive segmentation strategy is shown in Figure 1.1. In this semi-automatic process, the knowledge provided by the user via interaction in the form of clicks/scribbles assists the system in the segmentation process. The system learns to use interactions to predict better boundaries of the object; as these user-interactions provide high-level information indicating the object and background regions. In each iteration, the prediction is updated until a satisfactory segmentation result is obtained.

1.4 Comparison with InterCNN

- In this work, we have used nnU-Net as the base segmentation model to show that even though the prediction from base segmentation model can influence

the percentage gain, user-interaction can refine segmentation further for higher accuracy.

- In InterCNN, users needed to supply clicks as region scribbles around the objects of interest to guide segmentation; we have formalized user input via various representations.

- We demonstrate the performance of our proposed framework on Prostate dataset from NCI-ISBI 2013 Challenge and Heart, Spleen, Pancreas, Hippocampus dataset from Medical Segmentation Decathlon that shows the generalizability of our approach.

Chapter 2

RELATED WORKS

There have been several approaches being used for semantic segmentation. In this section, we review some of these state-of-the-art techniques.

2.1 Active Contour Models

Kass *et al.* (1988) proposed snakes algorithm to segment the images by means of energy minimization. In this approach, an initial contour is deformed along the boundary of an object in response to internal forces, external image forces and user defined constraints. These models are very sensitive to noise and the initial curve, which limits their practical applications. Other well known examples are active contour without edge (ACWE) by Chan and Vese (2001), geodesic active contours by Yezzi *et al.* (1997) and fast global minimization-based active contour model (FGM-ACM) by Bresson *et al.* (2007).

2.2 Neural Networks for Semantic Segmentation

With the introduction of deep learning algorithms, convolutional neural networks (CNNs) have significantly improved performance for segmentation tasks. Various classification architectures like Alexnet by Krizhevsky *et al.* (2012), VGG by Simonyan and Zisserman (2014), GoogLeNet by Szegedy *et al.* (2015) and ResNet by He *et al.* (2016) have been adapted to create semantic segmentation networks.

The most well known architecture U-Net by Ronneberger *et al.* (2015), uses the encoder-decoder architecture where the input image is down-sampled and then up-sampled to get image segmentation. U-Net's novel architecture has skip connections

which are designed to forward feature maps from down-sampling path to up-sampling path to avoid losing high-resolution information.

Various extensions of U-Net have been developed, a W-shaped network is proposed for 2D medical image segmentation task by Chen *et al.* (2018), Çiçek *et al.* (2016) proposed 3D U-Net architecture that deals with 3D volumetric data directly, Milletari *et al.* (2016) introduced a similar architecture, V-Net, which employs residual connections to design a deeper network for end to end segmentation of prostate cancer, Zhou *et al.* (2018) proposed U-Net++ with the integration of additional convolution layers in the form of dense skip connections in U-Net and tested their method on a variety of medical datasets. Some of the other relevant works includes H-DenseUNet by Li *et al.* (2018) for liver and tumor segmentation, PDV-Net by Hatamizadeh *et al.* (2018) for fast and automatic segmentation of pulmonary lobes from chest CT images, ASDNet by Nie *et al.* (2018) who designed attention model to segment prostate images with higher accuracy, Attention U-Net by Oktay *et al.* (2018), Gated-UNet by Schlemper *et al.* (2019) and nnU-Net by Isensee *et al.* (2018) which have also leveraged the attention concept into medical image segmentation.

2.3 Interactive Segmentation

As discussed previously, neural networks have been used in an effective way for performing semantic segmentation. However, supervised training of such models require large amount of high quality labels and acquiring such labelled data is tedious and often incur high costs. In order to reduce the cost of labeling, semi-automated or interactive methods have been proposed. Interactive approaches allows human-computer interaction to obtain more accurate segmentation.

Grabcut by Rother *et al.* (2004) and Graphcut by Freedman and Zhang (2005) are classic interactive segmentation models, which segment objects by gradually updating

the appearance model. More recently, building on advances in deep learning, CNN models have been extensively used for interactive segmentation. DeepIGeoS proposed by Wang *et al.* (2018b) uses geodesic distance transforms of scribbles as an additional input to the CNN for interactive segmentation of foetal MRI images and brain tumor images. Sakinis *et al.* (2019) uses point clicks that are modelled as Gaussian kernels and put them as input to an FCN for segmenting medical images. In BIFSeg, Wang *et al.* (2018a) proposed image-specific fine-tuning and incorporates bounding boxes and scribble based interaction. Here, users first draw a bounding box, the area inside this bounding box is considered as input to CNN to obtain an initial result, thereafter which users perform an image-specific fine-tuning to make CNN provide better segmentation results. Deep extreme points (DEXTR) by Roth *et al.* (2019), requires the user to click on the extreme boundary points of an object, generating initial segments via a random walker algorithm and then train a fully-supervised segmentation network.

Castrejon *et al.* (2017) proposed Polygon-RNN which predicts vertices of a polygon that are iteratively corrected. Several improvements to Polygon-RNN is done in Polygon-RNN++ by Acuna *et al.* (2018) where a better learning algorithm is proposed to train the model using reinforcement learning. Furthermore, Curve-GCN by Ling *et al.* (2019) represents object as a graph and use Graph Convolutional Network (GCN) for predicting the locations of all vertices simultaneously. In Curve-GCN, N control points are first initialized along a circle. These current coordinates are concatenated with features extracted from the corresponding location and propagated via a GCN to predict a location shift for each node. When human-in-the-loop, the annotator iteratively moves wrong control points onto their correct locations. Similarly, Pixel2Mesh by Wang *et al.* (2018c) also exploited a GCN to predict vertex locations of a 3D mesh.

2.4 CNNs with ACM

In recent years, researchers have also combined techniques like Active Contour Models (ACM) with deep learning approaches. One approach is to use these models as a post-processing step to improve an initial segmentation map. A different approach is to formulate new loss functions inspired by ACM principles. Hatamizadeh *et al.* (2019a) have proposed Deep Active Lesion Segmentation (DALS), an automated segmentation framework that utilises a level-set ACM formulation with a per-pixel-parameterized energy functional and a novel multiscale encoder-decoder CNN that learns an initialization probability map along with parameter maps for the ACM. In another work, Marcos *et al.* (2018) proposed Deep Structured Active Contours (DSAC), combining ACMs and CNNs for segmenting aerial images. Chen *et al.* (2019) proposed leveraging traditional active contour energy minimization into CNNs via a new loss function that combines the geometrical information with region similarity thus achieving better results than others. In an extended work, an end to end backpropagation trainable, fully-integrated FCN-ACM combination was introduced by Hatamizadeh *et al.* (2019b) in Deep Convolutional Active Contours (DCAC).

Chapter 3

METHODOLOGY

In this thesis, we present a two-step deep learning training framework. First, a supervised learning-based segmentation network is trained which takes original image as input and outputs an initial segmentation mask. Second, an interaction network is trained that utilizes the input image, the prediction from previous step, and a user interaction in form of scribble for segmentation refinement.

3.1 Base Segmentation network

We use a U-Net (Figure 3.1) based neural network architecture to predict an initial segmentation mask as shown in Figure 3.2. We also train a 3D nn-UNet by Isensee *et al.* (2018) as it is a robust and self-adapting framework with the ability to dynamically adapt to the details of the datasets (median patient size, input patch size, batch size, etc.) and amount of available GPU memory.

3.2 Interactive Segmentation network

We utilize the architecture based on InterCNN by Bredell *et al.* (2018) which allows for the network to have two additional inputs, user edits in the form of scribbles and most recent prediction.

3.2.1 User Interaction

User guidance is provided by generating clicks which acts as a guidance signal to the network. Two types of user emulated inputs are generated:

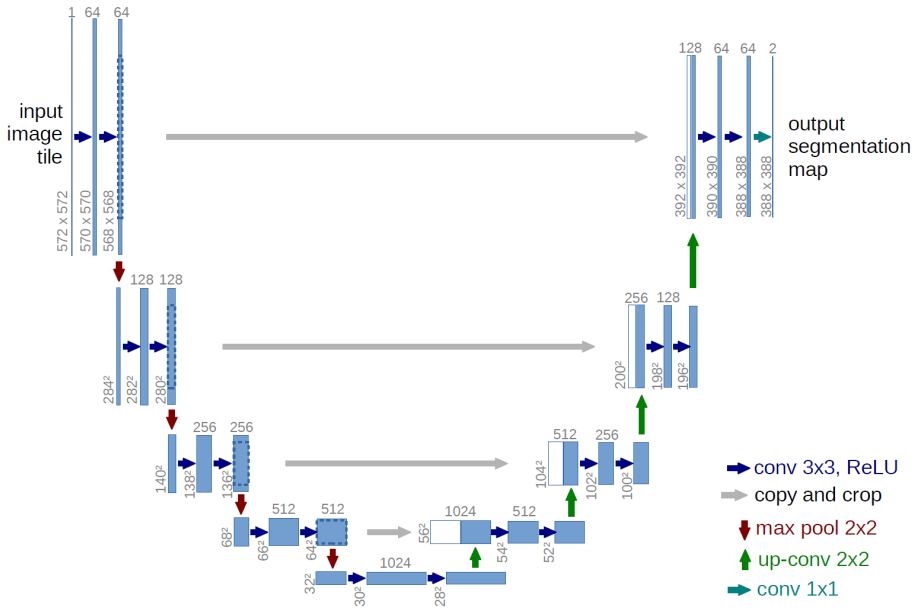

Figure 3.1: Unet Architecture From Ronneberger *et al.* (2015)

1. Foreground Clicks: These are placed within the area of interest to guide the network towards predicting foreground.

2. Background Clicks: These are placed in the false positive areas which have been incorrectly segmented as foreground regions.

Let S_f and S_b denote foreground and background clicks respectively. These interactions are given to the network in the form of an image with the same spatial aspect ratio as input image. All the pixels in the scribble image have zero value except for pixels corresponding to foreground and background clicks. In case of multi-class segmentation, clicks are generated corresponding to each class and are then combined together to get the final scribble image.

3.2.2 Training Strategy

The initial predictions (P_0) are received from a base segmentation network. These predictions are then compared with ground truth (G). The mislabelled pixels are

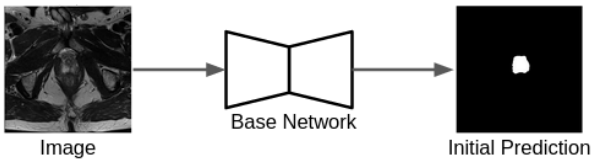

Figure 3.2: Base Segmentation Network (BSeg). Used for the Generation of Initial Prediction Mask and Scribbles.

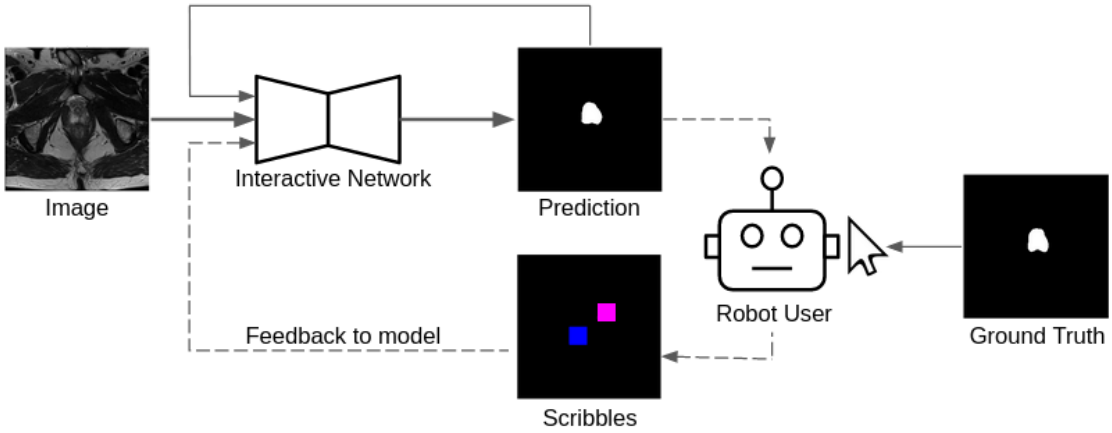

Figure 3.3: Overview of Our Interactive Segmentation Network (IntSeg). The Foreground Click (Pink) and Background Click (Blue) Constitute Our Encoding for Foreground and Background Correction to Create Scribbles. The Scribbles and Previous Prediction Are Concatenated with the Input Image to Form a 3-channel Input for the CNN. The Network Is Trained Iteratively Using the Simulated User Edits to Improve Segmentation Accuracy.

identified and the scribble image (S_0) is generated by the emulated user model. The input images (I) along with the initial predictions (P_0) and scribbles (S_0) are fed to the IntSeg network. Subsequently, the IntSeg network generates a new prediction (P_k) and corresponding scribble (S_k), which are then fed to the model at the next interaction (k). During each interaction (k), cross entropy loss is computed and the network weights are updated through back-propagation. This is done iteratively over multiple rounds (K).

The scribbles used during training should ideally be provided by users. However,

this is not feasible and hence we emulate user clicks by simulating the expected annotator behavior. Moreover, we train with a certain annotator noise model, by perturbing the position of clicks as we expect some inconsistent inputs from the annotators.

Similarly, at test time, we iteratively sample the clicks for correcting the predicted mask with the same annotator noise model, as used in training.

3.2.3 Simulating Annotations

First, mislabelled pixels are identified based on the prediction by comparing it with the ground truth mask. For example, Figure 3.4 (e) is the difference map which shows incorrectly labelled pixels. The black and white region shows false positive R_{fp} and false negative R_{fn} regions respectively. We have simulated various annotator behaviour by using different types of user-inputs or clicks.

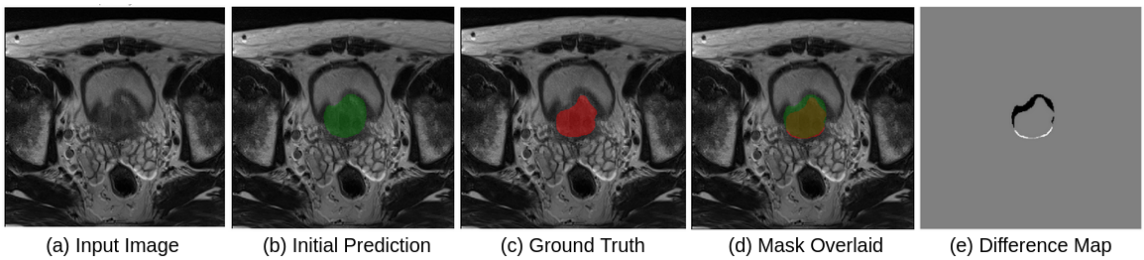

Figure 3.4: Example Shows (a) Input Prostate Image (b) Initial Prediction from Base Segmentation Network (c) Ground Truth Mask (d) Ground Truth and Initial Prediction Overlaid Together (e) Difference Map to Show Clearly the False Positive and False Negative Regions

1. Region Clicks: A 5x5 region/patch is placed randomly in the incorrectly predicted area for R_{fp} and R_{fn} correction (Figure 3.5 (a)).

2. Region Clicks within the largest connected component: For both R_{fp} and R_{fn} regions, the largest incorrect cluster region is selected and the user emulated

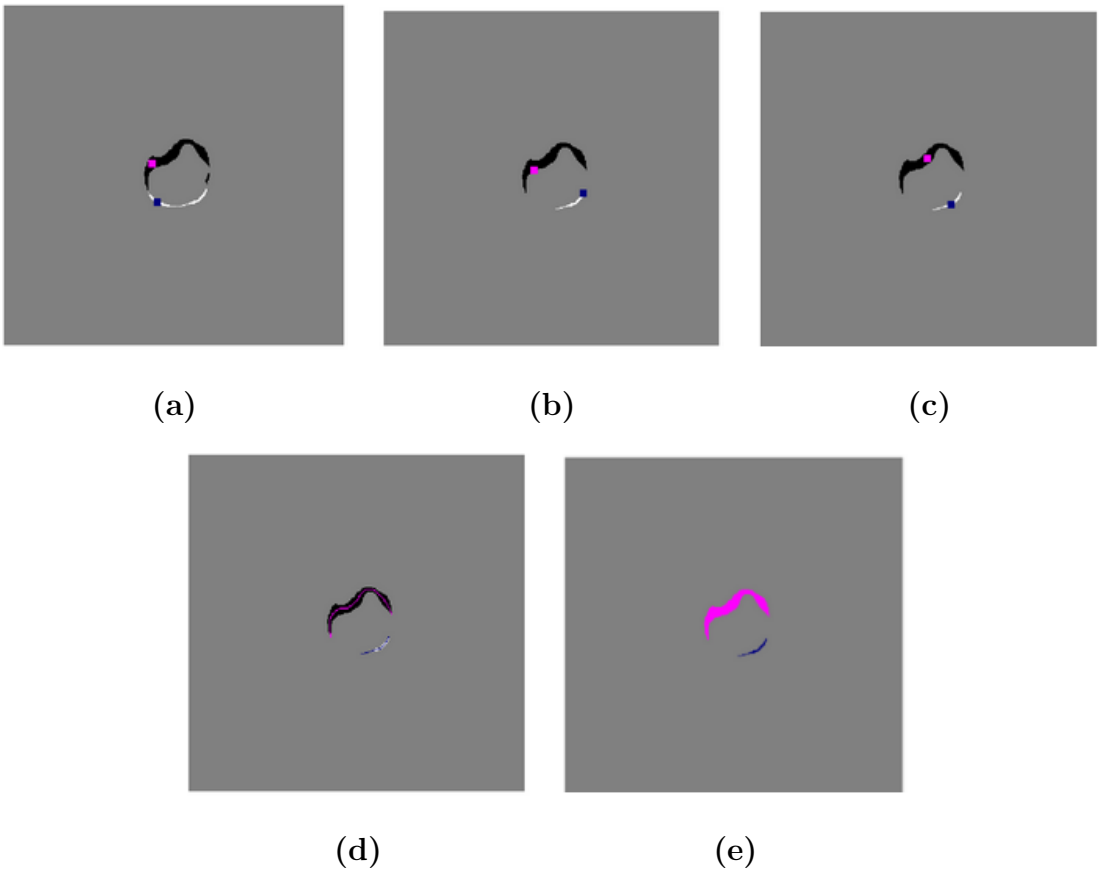

Figure 3.5: Different Types of Annotation Methods. Here the Markings in Pink and Blue Color Shows the Scribble/Clicks Used for Foreground and Background Correction Respectively.

clicks are placed in that region (Figure 3.5 (b)). This error region is the largest connected group of pixels of ground truth mask that has been mislabelled.

3. Region Clicks at the center of the largest connected component: The user annotators tend to correct the error by clicking at the center of the incorrect region. To replicate this behaviour, the largest incorrect cluster region is selected and erosion is performed to get the center of the cluster. A 5x5 region is placed at the center of this cluster (Figure 3.5 (c)).

4. Curves: We believe that scribbles are an efficient way for annotators to create corrections in the error regions. It is particularly favoured since user can simply

draw a curve by dragging the cursor inside the error region instead of precisely drawing the object boundaries.

The user drawn curves are emulated by utilizing image skeletonization. For R_{fp} and R_{fn} regions, the largest incorrect cluster region is selected and is skeletonized to 1 pixel width to match the expected behaviour of the user drawn curve (Figure 3.5 (d)). These scribbles are then smoothed using erosion and dilation to introduce some amount of noise.

5. Full region: The complete largest connected component is selected as scribble for R_{fp} and R_{fn} correction (Figure 3.5 (e)). This user interaction strategy is practically not feasible in a realistic use case as it requires annotators to mark the complete incorrect cluster. Therefore, this simulation is used just for a comparison.

Chapter 4

EXPERIMENTS

4.1 Data

In this work, we have utilized training datasets (as they include ground truth annotations) from public challenges, namely Medical Segmentation Decathlon and NCI-ISBI 2013 Challenge.

1. NCI-ISBI 2013: We adopt the T-2 weighted MRIs of the Prostate dataset from the NCI-ISBI 2013 challenge which in total contains 60 volumes from different patients. We utilize the 29 subjects which have multi-class ground truth segmentations, consisting of 2 labels namely central gland and peripheral zone.

   We randomly divide the dataset into 4 groups D1-D4. D1 contains 15 patient data and is used for training the base model, D2 consists of 23 patients (including D1) and is used for training interactive model, D3 is used for validation and contains 1 patient, G4 contains the remaining five patients which are used for testing. For benchmarking against other approaches, we kept the same dataset split for the base segmentation and interactive segmentation algorithms.

2. Medical Segmentation Decathlon: The medical decathlon challenge (MSD) provides ten different tasks on 3D CT/MR image segmentation. These tasks are selected to cover a large proportion of the dataset variability in the medical domain. We have used four public datasets from the challenge.

   - Heart: The dataset includes 20 MRI scans covering the entire heart acquired during a single cardiac phase (free breathing with respiratory and

ECG gating). Images were obtained on a 1.5T Achieva scanner.

- Spleen: The spleen dataset consists of 41 portal venous phase CT scans from patients undergoing chemotherapy treatment for liver metastases at Memorial Sloan Kettering Cancer Center.
- Pancreas: The MSD Pancreas Tumors dataset is labeled with both pancreatic tumors and normal pancreas regions. The training set contains 282 portal venous phase CT cases.
- Hippocampus: The dataset contains 263 training samples which have been used to segment two neighbouring small structures i.e. anterior and posterior hippocampus.

Dataset	Modalities	Total Classes	No. of Samples
Prostate (NCI-ISBI)	MRI(T2)	2	29
Heart	MRI	1	20
Spleen	CT	1	41
Pancreas	CT	2	282
Hippocampus	MRI	2	260

Table 4.1: Properties of Different Datasets from Medical Segmentation Decathlon (2018) and NCI-ISBI (2013). For the 4 MSD Datasets, All Models (Base and Interactive) Are Trained from Scratch and Evaluated Using Five-fold Cross-validation on the Training Set.

4.2 Implementation details

Training with NCI-ISBI Prostate dataset. We trained the base U-Net network for 400 epochs. The learning rate was set as 0.0001 and the images were randomly

flipped horizontally/vertically, rotated, resized and cropped before feeding into the network. In addition, Adam optimizer and Cross Entropy loss was used for training.

Training with MSD dataset. We used nnU-Net framework (base model) for training on MSD dataset that adapts itself to any given dataset without user intervention. Hence, all the hyperparameters tuning and design choices such as the U-Net architecture, dice loss, data augmentation were automatically determined by nnU-Net.

Interactive segmentation network is trained on both datasets for K number of interactions per batch. Hence, predictions from each batch are updated iteratively with the respective scribbles and fed into the network for K interactions. We train our model with the Adam optimizer for 80 epochs with learning rate 0.0001 and apply data augmentation by vertical or horizontal flipping, cropping and random rotation. For pre-processing, all the images were normalized by mean value and standard variation of the training set.

4.3 Metrics

For quantitative evaluation, we measured the Dice score as $\frac{2|R_g \cap R_p|}{|R_g|+|R_p|}$, where R_p and R_g are the regions predicted by model and the ground truth. A robot user is used to simulate user-annotations during testing up to K interactions. Figures 5.1 – 5.3 reported the average segmentation accuracy (mean Dice score) across the first 10 iterated clicks on each dataset.

Chapter 5

RESULTS

5.1 Base Segmentation Results

Once BSeg is trained, we evaluate it on the samples reserved for testing purposes. Figure 5.2 and Figure 5.3 shows the initial segmentation performance at interaction 0 using nnU-Net as the base segmentation model. Figure 5.1 shows the results on Prostate dataset using a U-Net architecture.

5.2 Interactive Segmentation Results

We utilize the initial prediction from BSeg and train interactive network from scratch. Figure 5.1 - Figure 5.3 shows a clear gain in performance with the number of interactions using IntSeg.

Table 5.1 summarizes the number of clicks needed for each annotation method (scribble) to reach a certain performance. It can be observed that our skeleton method (curve) outperforms other approaches on all the datasets.

5.3 Ablation Study

To further validate our results, we conduct ablation study on the heart and spleen dataset. We define the performance by carrying out a percentile study of the data distribution. Various performance points are *p_max* (max performance), *p_25* (25th percentile), *p_50* (50th percentile), *p_75* (75th percentile), and *p_min* (min performance).

The entire test dataset (population) is divided into four equal splits (sub-population)

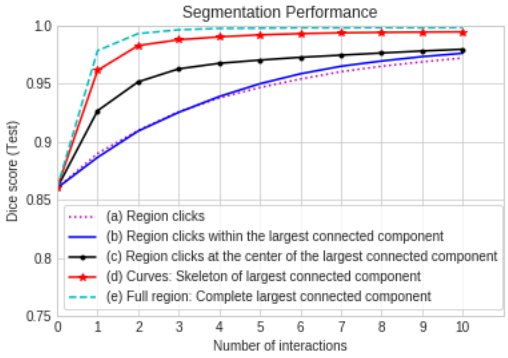

Figure 5.1: Comparison of Different Types of Annotation Strategy on Prostate Dataset. Simple Strokes (Curves) Behave Better than the Alternatives.

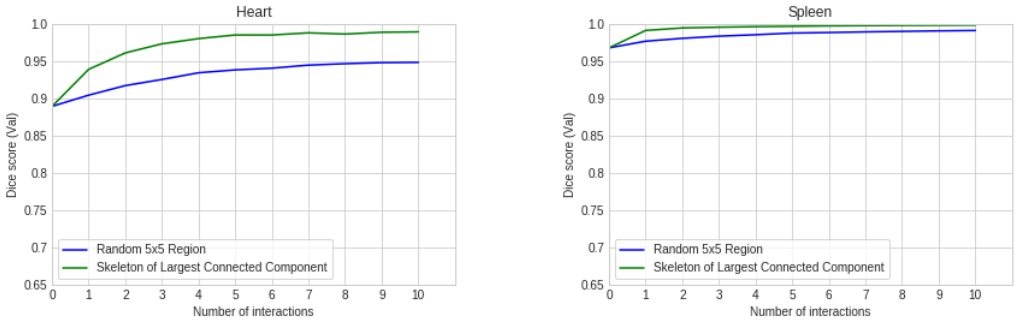

Figure 5.2: Comparison of mDice Scores Vs Interactions on (a) Heart and (b) Spleen Dataset Using Region (5x5) and Skeleton Scribble. Both Type of Clicks Bring Improvements in Every Interaction

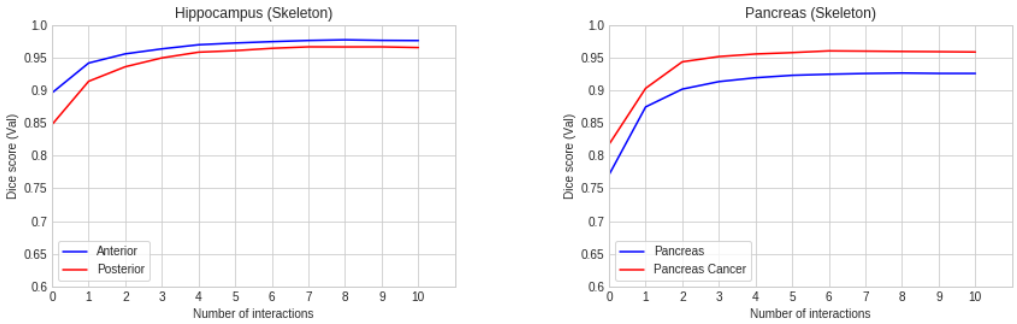

Figure 5.3: Comparison of mDice Scores Vs Interactions on (a) Hippocampus and (b) Pancreas Dataset Using Skeleton Scribble. The Performance at Interaction 0 Is Obtained Using nnU-Net as Base Model.

Dataset	Scribble	0	1	2	3	4
Prostate	Region (5x5)	0.8598	0.8896	0.9095	0.9256	0.9373
Prostate	Centroid (5x5)	0.8598	0.9264	0.9512	0.9624	0.9672
Prostate	Curve	0.8598	0.9611	0.9824	0.9874	0.9899
Heart	Region (5x5)	0.8889	0.9039	0.9168	0.9250	0.9340
Heart	Curve	0.8889	0.9385	0.9606	0.9728	0.9797
Spleen	Region (5x5)	0.9673	0.9763	0.9801	0.9830	0.9849
Spleen	Curve	0.9673	0.9907	0.9940	0.9950	0.9958
Pancreas	Curve	0.7706	0.8743	0.9015	0.9130	0.9189
Pancreas Cancer	Curve	0.8170	0.9027	0.9432	0.9512	0.9551
Anterior Hippocampus	Curve	0.8957	0.9413	0.9553	0.9630	0.9692
Posterior Hippocampus	Curve	0.8473	0.9134	0.9357	0.9492	0.9579

Table 5.1: Results of Different Proposed Scribble Type in Interaction Network on Five 3D Segmentation Tasks Across Organs, Diseases, and Modalities for Interactions 0 to 4.

based on the performance percentile, as *[p_min, p_75]*, *(p_75, p_50]*, *(p_50, p_25]*, *(p_25, p_max]*. We then emphasize on the images with inaccurate segmentation and indicate the helpfulness of user-annotations in interactive training strategy.

Figures 5.4 and 5.5 shows that the average performance gain after one and two interactions for *[p_min, p_75]* and *(p_75, p_50]* sub-population. Results suggest that the IntSeg model learns to use guidance information given to the model through scribbles, as the poorly segmented p_50% of images show a remarkable improvement in performance. Since, the improvement of performance is significant, it reflects the importance of user-clicks.

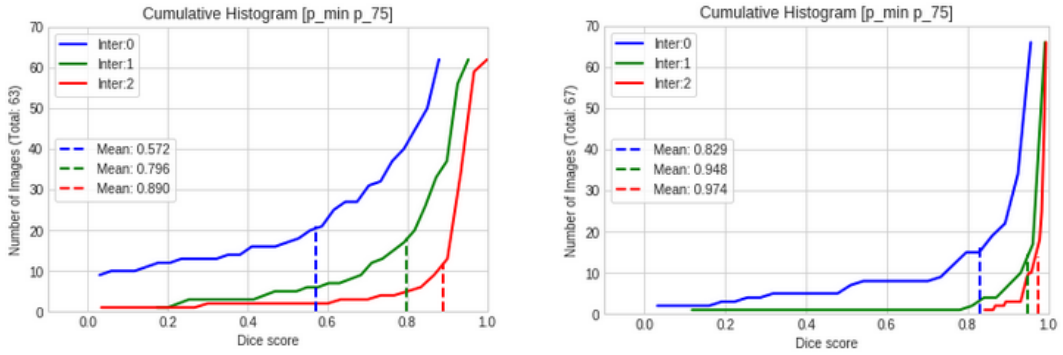

Figure 5.4: Cumulative Histogram Shows the Performance Improvement for Images *[p_min, p_75]* with Interactions 0-2 for Dataset (a) Heart and (b) Spleen. The Mean Dice Score Increased From 0.572 to 0.890 after 2 Interactions for *p_25%* of Images in Case of Heart Segmentation. For Spleen, the Average Performance Increase Is From 0.829 to 0.974.

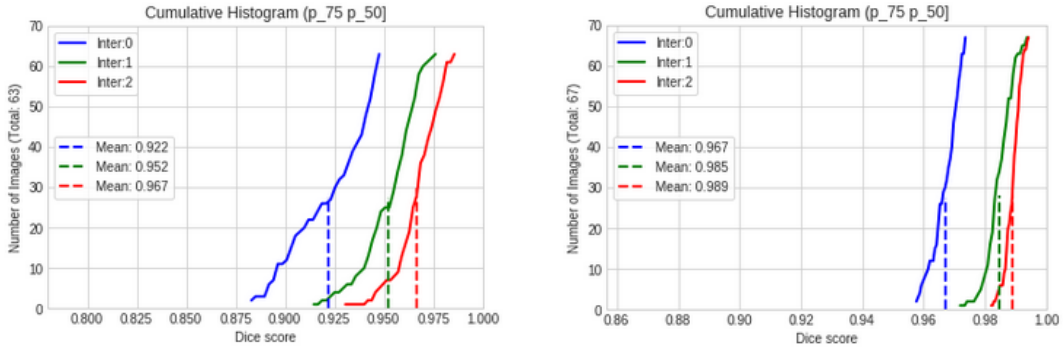

Figure 5.5: Cumulative Histogram Shows the Performance Improvement for Images *(p_75, p_50]* with Interactions 0-2 for Dataset (a) Heart and (b) Spleen. The Mean Dice Score Significantly Increased for *(p_75, p_50]* Sub-population.

5.4 Influence of the Annotation Strategy

The type of scribble used for correction is important and can influence the performance. Hence, we evaluate performance across different scribble inputs as shown in Figure 3.4 on the Prostate dataset from NCI-ISBI 2013 challenge.

As shown in Figure 5.1, we observe that the best performance is achieved when the largest incorrect cluster is considered a scribble. However, this may be impractical for annotators to mark the complete cluster. We also notice that skeleton approach

clearly increases the benefits of the annotations compared to other encoding; the mDice score is increased by 10% in just one interaction. Finally, other region based interactions have also been evaluated which rely on clicks; also bringing considerable improvement in performance. Through these experiments, we demonstrate the effectiveness of good specification of user interaction to minimize the number of clicks and maintain high quality segmentation.

5.5 Influence of the Iteration Training Parameter

IntSeg is trained for K number of iterations per batch. Hence, predictions from each batch are updated iteratively with the respective scribbles and fed into the interactive network for K iterations. K is varied from 5 to 15 to check the influence of the number of iterations during training.

Figure 5.6 shows the results of varying K on Prostate data. It is seen from figure that with lower K the performance improvement is lower in comparison to when K is 10 or higher. Although, the improvement is not substantial when K is higher than 10.

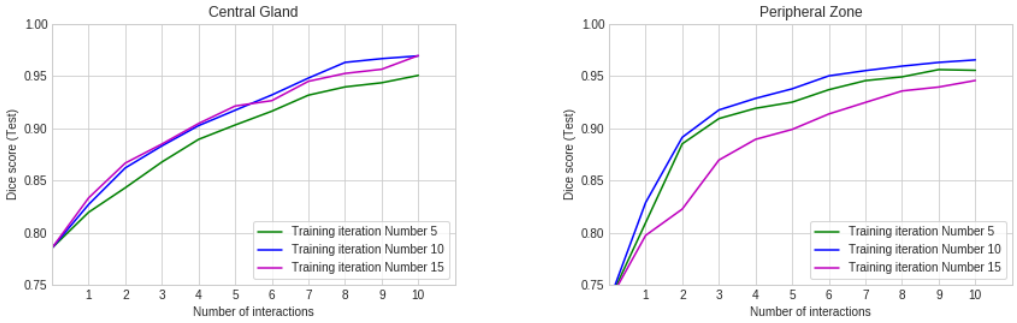

Figure 5.6: Segmentation Performance on Multi-class Prostate Dataset (a) Central Gland and (b) Peripheral Zone Using Region Scribble.

5.6 Influence of the Base Model

We compare the significance of the base model on the performance. This also allows us to study if the initial quality of the segmentation maps influences annotations to refine its prediction. Here, we use an empty mask as an initial prediction to train interactive network architecture. As we can see in Figure 5.7, the use of base model clearly improves the segmentation performance.

Figure 5.8 shows the decrease in average length of scribble for first 10 interactions. The significant variation in scribble length can be explained by their initial prediction. In case of blank mask, the annotator draws a curve having average length 109 per instance over two rounds of interactive segmentation. In comparison, our approach reaches 98% mean Dice score in an average curve length of 70 per instance. We thus present a trade-off between annotation effort and quality.

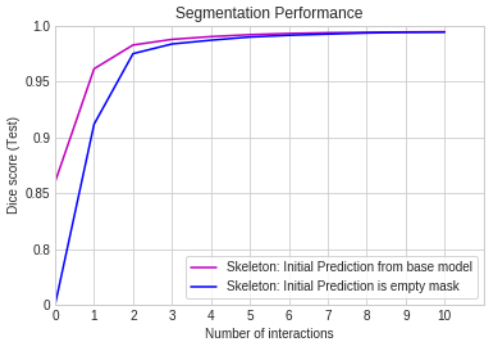

Figure 5.7: Comparison of Using Two Different Initial Segmentation Mask on the Prostate Dataset.

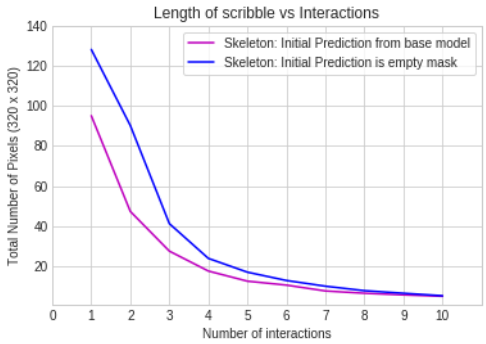

Figure 5.8: Influence of the Initial Segmentation Mask on the Average Curve Length per Interactions.

5.7 Qualitative Results

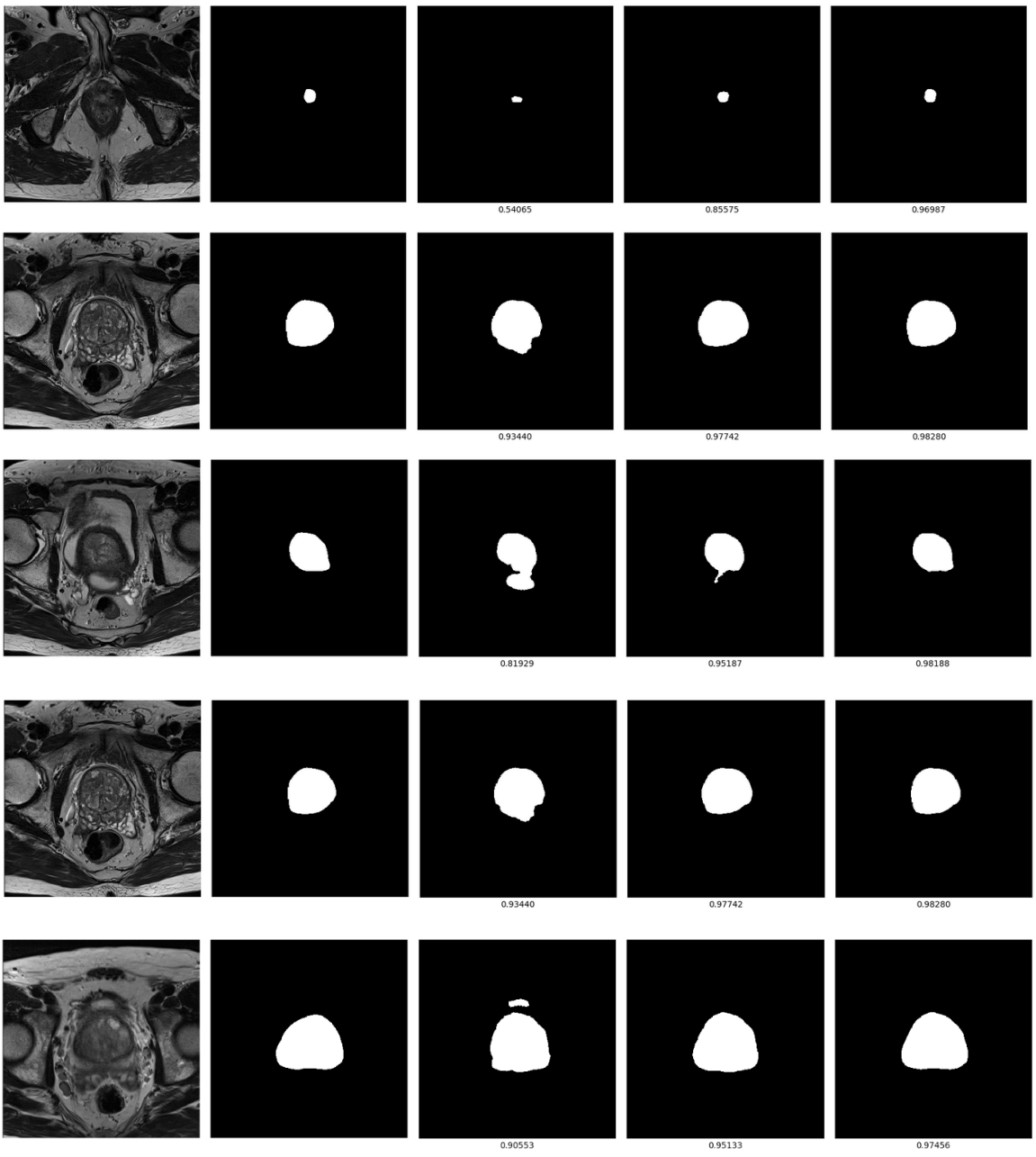

Figure 5.9: Qualitative Results of Prostate Segmentation. Binary Segmentation Is Performed by Combining the Labels of Central Gland and Peripheral Zone. Each Row Shows the Input Image, Ground Truth, Prediction from Base Model, Prediction at Interaction 1 and 2 Using Skeleton Scribble.

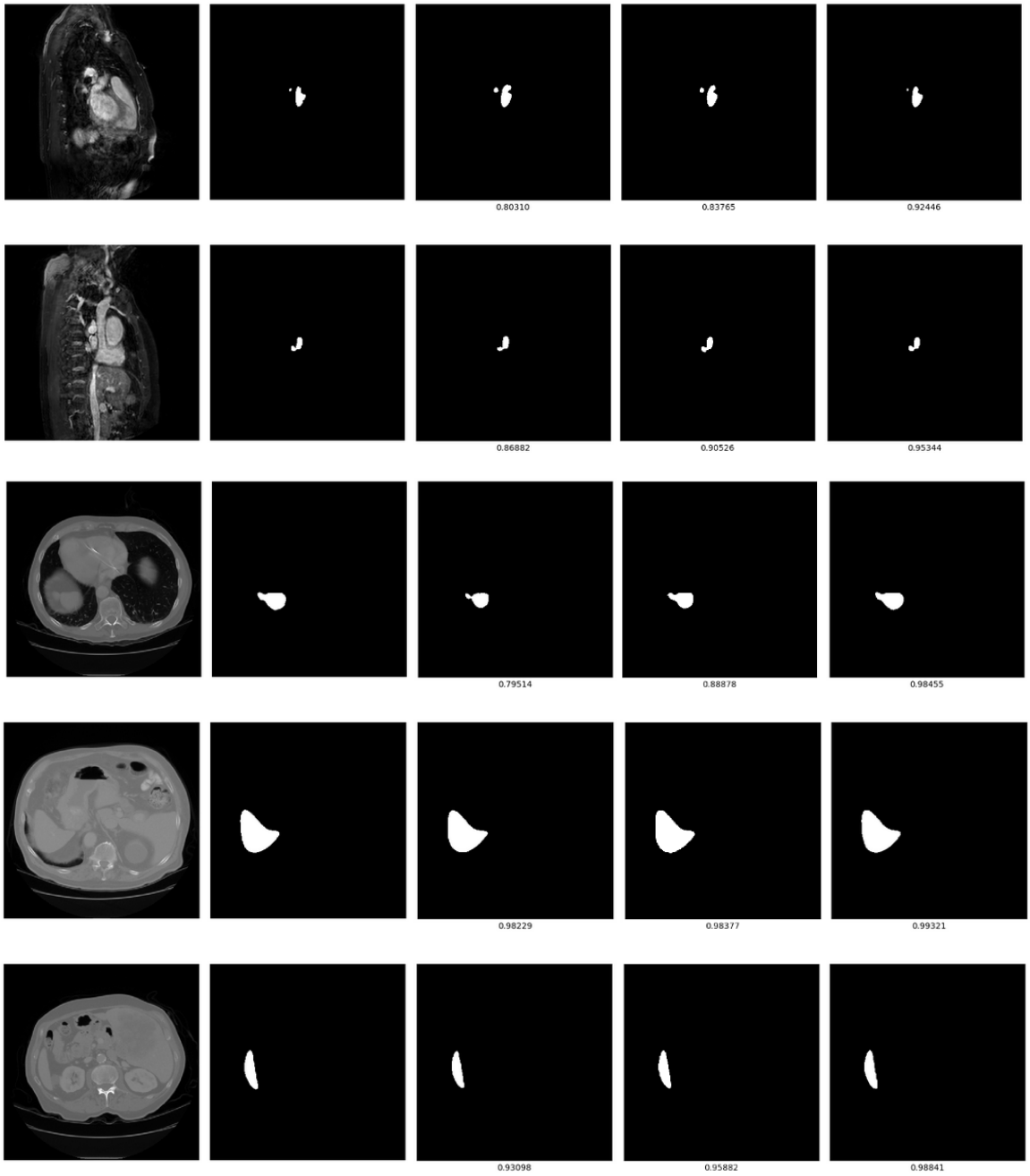

Figure 5.10: Visual Comparison of Two Annotation Strategies for Heart (Rows 1 and 2) and Spleen (Rows 3, 4 and 5) Segmentation. Each Row Shows the Input Image, Ground Truth, Prediction from Base Model, Prediction at Interaction 2 Using Region Scribble and Skeleton Scribble Respectively. It Can Be Observed That Skeleton Scribble Achieves a Larger Improvement of Accuracy from the Initial Segmentation When Compared with the Use of Region Scribble.

Chapter 6

CONCLUSION

A semi-automatic training strategy has been proposed which utilizes user-scribbles to guide the network to correct segmentation error. The user-model emulates an actual annotator and generates scribbles for training the network. The model continuously improves with each iteration from new information provided by the user scribble and updated prediction. Various user interactions were evaluated and it is found that the proposed skeleton based simulation scheme performs better than a region based scribble. Further, we observe that this requires far less user inputs compared with other scribbles and achieves higher accuracy in just two to three correction.

Also, we present an extensive ablation study to examine the significant performance gain for poorly segmented examples. Results suggest that the interaction network better handles the erroneous region in the form of skeleton feedback, to yield high quality segmentations.

Moreover, we show our models' strong generalization capabilities by evaluating our approach across different datasets and domains. Finally, we demonstrate that using interaction network on top of the state-of-the-art segmentation architecture, improves the prediction accuracy further compared to when the base model is a simple encoder decoder architecture.

REFERENCES

Acuna, D., H. Ling, A. Kar and S. Fidler, "Efficient interactive annotation of segmentation datasets with polygon-rnn++", in "Proceedings of the IEEE conference on Computer Vision and Pattern Recognition", pp. 859–868 (2018).

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

Bresson, X., S. Esedoḡlu, P. Vandergheynst, J.-P. Thiran and S. Osher, "Fast global minimization of the active contour/snake model", Journal of Mathematical Imaging and vision **28**, 2, 151–167 (2007).

Castrejon, L., K. Kundu, R. Urtasun and S. Fidler, "Annotating object instances with a polygon-rnn", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 5230–5238 (2017).

Chan, T. F. and L. A. Vese, "Active contours without edges", IEEE Transactions on image processing **10**, 2, 266–277 (2001).

Chen, W., Y. Zhang, J. He, Y. Qiao, Y. Chen, H. Shi and X. Tang, "W-net: Bridged u-net for 2d medical image segmentation", arXiv preprint arXiv:1807.04459 (2018).

Chen, X., B. M. Williams, S. R. Vallabhaneni, G. Czanner, R. Williams and Y. Zheng, "Learning active contour models for medical image segmentation", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 11632–11640 (2019).

Çiçek, Ö., A. Abdulkadir, S. S. Lienkamp, T. Brox and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation", in "International conference on medical image computing and computer-assisted intervention", pp. 424–432 (Springer, 2016).

Criminisi, A., T. Sharp and A. Blake, "Geos: Geodesic image segmentation", in "European Conference on Computer Vision", pp. 99–112 (Springer, 2008).

Freedman, D. and T. Zhang, "Interactive graph cut based segmentation with shape priors", in "2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)", vol. 1, pp. 755–762 (IEEE, 2005).

Hatamizadeh, A., S. P. Ananth, X. Ding, D. Terzopoulos, N. Tajbakhsh *et al.*, "Automatic segmentation of pulmonary lobes using a progressive dense v-network", in "Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support", pp. 282–290 (Springer, 2018).

Hatamizadeh, A., A. Hoogi, D. Sengupta, W. Lu, B. Wilcox, D. Rubin and D. Terzopoulos, "Deep active lesion segmentation", in "International Workshop on Machine Learning in Medical Imaging", pp. 98–105 (Springer, 2019a).

Hatamizadeh, A., D. Sengupta and D. Terzopoulos, "End-to-end deep convolutional active contours for image segmentation", arXiv preprint arXiv:1909.13359 (2019b).

He, K., X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 770–778 (2016).

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

Jang, W.-D. and C.-S. Kim, "Interactive image segmentation via backpropagating refinement scheme", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 5297–5306 (2019).

Kass, M., A. Witkin and D. Terzopoulos, "Snakes: Active contour models", International journal of computer vision **1**, 4, 321–331 (1988).

Krizhevsky, A., I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", Advances in neural information processing systems **25**, 1097–1105 (2012).

Li, X., H. Chen, X. Qi, Q. Dou, C.-W. Fu and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes", IEEE transactions on medical imaging **37**, 12, 2663–2674 (2018).

Ling, H., J. Gao, A. Kar, W. Chen and S. Fidler, "Fast interactive object annotation with curve-gcn", in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition", pp. 5257–5266 (2019).

Marcos, D., D. Tuia, B. Kellenberger, L. Zhang, M. Bai, R. Liao and R. Urtasun, "Learning deep structured active contours end-to-end", in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", pp. 8877–8885 (2018).

Medical Segmentation Decathlon, "Medical segmentation decathlon", `http://medicaldecathlon.com/`, Last accessed on 2019 (2018).

Milletari, F., N. Navab and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation", in "2016 fourth international conference on 3D vision (3DV)", pp. 565–571 (IEEE, 2016).

Nie, D., Y. Gao, L. Wang and D. Shen, "Asdnet: attention based semi-supervised deep networks for medical image segmentation", in "International conference on medical image computing and computer-assisted intervention", pp. 370–378 (Springer, 2018).

Oktay, O., J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas", arXiv preprint arXiv:1804.03999 (2018).

Ronneberger, O., P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in "International Conference on Medical image computing and computer-assisted intervention", pp. 234–241 (Springer, 2015).

Roth, H., L. Zhang, D. Yang, F. Milletari, Z. Xu, X. Wang and D. Xu, "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", pp. 42–50 (Springer, 2019).

Rother, C., V. Kolmogorov and A. Blake, "" grabcut" interactive foreground extraction using iterated graph cuts", ACM transactions on graphics (TOG) **23**, 3, 309–314 (2004).

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

Schlemper, J., O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images", Medical image analysis **53**, 197–207 (2019).

Simonyan, K. and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556 (2014).

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 1–9 (2015).

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

Wang, G., M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin *et al.*, "Deepigeos: a deep interactive geodesic framework for medical image segmentation", IEEE transactions on pattern analysis and machine intelligence **41**, 7, 1559–1572 (2018b).

Wang, N., Y. Zhang, Z. Li, Y. Fu, W. Liu and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images", in "Proceedings of the European Conference on Computer Vision (ECCV)", pp. 52–67 (2018c).

Yezzi, A., S. Kichenassamy, A. Kumar, P. Olver and A. Tannenbaum, "A geometric snake model for segmentation of medical imagery", IEEE Transactions on medical imaging **16**, 2, 199–209 (1997).

Zhou, Z., M. M. R. Siddiquee, N. Tajbakhsh and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation", in "Deep learning in medical image analysis and multimodal learning for clinical decision support", pp. 3–11 (Springer, 2018).