No Prompting Frozen Foundation Models: Interactive Medical Volume Segmentation using Continual Test Time Adaptation of Compact Models

Kushal Borkar
International Institute of Information Technology, Hyderabad
Hyderabad, India
kushal.borkar@research.iiit.ac.in

Abhilaksh Singh Reen
Indian Institute of Technology, Delhi
Delhi, India
abhilakshsinghreen@gmail.com

C.V. Jawahar
IIIT-Hyderabad
Hyderabad, India
jawahar@iiit.ac.in

Chetan Arora
Indian Institute of Technology Delhi
Delhi, India
chetan@cse.iitd.ac.in

Abstract

Automated segmentation of medical image volumes promises to reduce costly medical experts' time for annotation. However, using machine learning for the task is challenging due to variations in imaging modalities and scarcity of patient data. While interactive image segmentation methods and foundational models incorporating user-provided prompts to refine segmentation masks have shown promise, they overlook crucial sequential information between the slices in 3D medical image volumes and videos, resulting in discontinuities in the segmentation results. This paper proposes a new framework that dynamically updates model parameters during inference in a test time training framework using user-provided scribbles. Our framework preserves acquired knowledge from the previous slices of the current medical volume and the training dataset via student-teacher learning. We evaluate our method on diverse CT, MRI, and microscopic cell datasets. Our framework significantly reduces user annotation time by a factor of 6.72×. Compared to other interactive segmentation methods, we reduce the time by a factor of 2.64×. Our method also outperforms prompting foundation models for segmentation by achieving a dice score of 0.9 in 3-4 interactions compared to 5-8 user interactions for the foundation model, significantly reducing annotation time for the CT and MRI volumes.

CCS Concepts

• **Computing methodologies → Image segmentation**; *Visual inspection*; • **Human-centered computing → Interaction design process and methods**.

ICVGIP 2024, December 13–15, 2024, Bengaluru, India
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1075-9/24/12
https://doi.org/10.1145/3702250.3702275

Keywords

Medical Image Analysis, Image Segmentation, Human-Computer Interaction, Continual Learning

ACM Reference Format:
Kushal Borkar, Abhilaksh Singh Reen, C.V. Jawahar, and Chetan Arora. 2024. No Prompting Frozen Foundation Models: Interactive Medical Volume Segmentation using Continual Test Time Adaptation of Compact Models. In *Indian Conference on Computer Vision Graphics and Image Processing (ICVGIP 2024), December 13–15, 2024, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3702250.3702275

1 Introduction

Medical image segmentation is essential to isolate regions of interest for medical diagnosis, modeling, and intervention tasks. Approaches based on deep learning for semantic and instance segmentation have seen success in medical imaging [37]. However, these methods require extensive data and depend heavily on the availability of detailed voxel-level segmentation, which requires considerable time, labor, and expertise [34]. To address this, IIS has emerged as a critical technique, enabling the efficient extraction of segmentation masks for specific objects with minimal user effort. Consequently, there has been a surge in research focused on developing and refining IIS techniques, exploring various types of user interactions, such as bounding boxes [23, 41], polygons [1], clicks [8, 20–22, 31, 32], scribbles [2, 39], and their combinations [44].

While current IIS techniques can segment regions of interest based on the user's input (clicks or bounding boxes), they rely on developing more effective backbone networks or refinement modules built upon the backbone to boost their segmentation performance. Recent vision foundation models for segmentation use interactive cues to work well on unseen data without fine-tuning. These models, trained on extensively large image datasets, segment potentially unseen regions in the target data by taking user interactions through clicks, bounding boxes, or scribbles [17]. SAM and fine-tuned SAM variations [23] perform well but may not consistently perform satisfactorily for low contrast scenarios in medical imaging when highly accurate segmentation boundaries are needed. These methods demand precise clicks or bounding boxes; erroneous clicks or bounding boxes may yield unapproved segmentation masks. When annotating a Region of Interest (ROI) in medical images, it is essential to note that scribbles are more accurate in locating ROIs than clicks or bounding boxes. Scribbles are particularly useful when annotating irregularly shaped ROIs, often in medical image annotation.

The typical procedure in a regular 3D segmentation process for CT/MRI volumes involves annotating each slice individually. However, conventional IIS and prompting foundation models pose challenges for such a scenario, as the annotators must manually provide the inputs for each slice again. This is inefficient and does not exploit the similarity between adjacent slices, as the model does not retain knowledge beyond the current slice. We address this gap by proposing a novel framework to combine ideas from continual learning with interactive segmentation in a continual test time adaptation [36] framework. We capitalize on this observation: the segmentation masks between close slices tend to overlap, and using the given set of corrections for the current refined segmentation mask, we use the continual adaptation method to adapt model parameters for the slices and subsequent CT/MRI volumes.

Our approach leverages an iterative process that combines a teacher-student architecture with test time training to adapt any pre-trained, off-the-shelf model on the fly without requiring architectural modifications. The teacher and the student have the same architecture and weight initialization. We use test-time augmentations and low-variance thresholding to generate an averaged pseudo-segmentation mask for a slice. The student model is updated in a test-time training style based on user-annotated scribbles and the pseudo-segmentation mask. In contrast, the teacher is updated through an exponential moving average (EMA) of the student's parameters. Unlike prior approaches [35], which restore the weights after each slice and end up deleting all accumulated knowledge, in our proposal, the teacher model serves as the repository of knowledge across the slices. The student model focuses on quick adaptation for the current slice. This leads to a balanced knowledge accumulation from previous and current slices. We note that a naive approach that does not use a teacher model but keeps on accumulating the knowledge in a student model across slices is not viable, as test time adaptation (TTA) will lead to overfitting of a student model on the current slice, and deteriorate the performance on the next slice. Hence, using the teacher model for regularization is essential for knowledge preservation without overfitting.

We extensively evaluate our method on five publicly medical image datasets, where anatomical differences among individuals and ambiguous boundaries are present. With the plain backbone of the UNet Network trained on medical image datasets with CT and MRI images(along with other backbones), our method significantly reduces user annotation time compared to full-human annotations (factors of 3.05 to 6.72). Compared to interactive methods with frozen parameters, our method reduces user interaction time by 1.22 to 1.93x on CT and MRI Volumes. Our method achieves a dice score exceeding 0.9 in 3-4 interaction loops, which outperforms the previous best interactive segmentation method by 2-3 interaction loops, resulting in a 20-36% reduction in user interactions. While comparing our framework with the foundational models for segmentation, our method takes 3-4 user interactions compared to 5-8 user interactions for the foundation model on CT and MRI Volumes to reach the dice score of 0.9. Our main contributions are:

(1) We propose a new framework for interactive segmentation using the TTA of a teacher-student architecture.
(2) The proposed framework retains past and current knowledge captured through user interactions, making it especially useful for volume segmentation tasks.
(3) Our framework can incorporate arbitrary segmentation models as its backbone, thus making it future-proof.
(4) Our method achieves state-of-the-art performance on all publicly available datasets in our experiments, outperforming conventional IIS methods and prompting foundational models.

2 Related Work

2.1 3D Medical Image Segmentation

Efficient encoder-decoder architectures like UNet [27] and UNet++ [45] have demonstrated success in medical image segmentation. These architectures have been extended to 3D for volumetric segmentation tasks, with examples including VNet, a 3D fully convolutional neural network (CNN) proposed by [25], and a 3D extension of UNet by [9]. ConResNet by [43] introduced inter-slice context residual learning for improved performance. Recent advancements have incorporated transformers alongside CNNs in several volume segmentation methods [14, 40]. UNETR [14] leverages a transformer-based encoder to learn sequential representations of the input volume, effectively capturing global information across multiple scales. CoTr [40] proposes efficient bridging between CNNs and transformers. Yet, the generalization of medical image segmentation techniques remains challenging due to the inherent challenges of medical images, such as low tissue contrast, highly variable and irregular shapes of segmentation targets, diverse imaging and segmentation protocols, and variations across patients.

2.2 Interactive Image Segmentation (IIS)

IIS takes user inputs like clicks or bounding boxes to improve segmentation results. Fueled by their success in a variety of applications, deep learning-based segmentation approaches have been used in conjunction with various interaction techniques, including bounding boxes [41], polygons [1], clicks [32], and scribbles [39]. Xu et al. [41] introduced a click simulation strategy, which was adopted in subsequent research. Methods like f-BRS [31] and RiTM [32] emphasized the importance of modern backbones and addressed limitations in existing inference-time optimization by proposing an iterative training procedure. FocalClick [8] tackled the issue of mask refinement destroying correct parts of the image. iSegFormer [20] leverages a Swin Transformer backbone specifically for medical image segmentation. PseudoClick [21] reduces human annotation costs by estimating the following click location. None of the above interactive segmentation methods exploit contextual information between the slices for medical image volumes.

2.3 Continual Learning (CL) for Volume Segmentation

CL deals with acquiring new knowledge over time while retaining previously learned information (or avoiding catastrophic forgetting

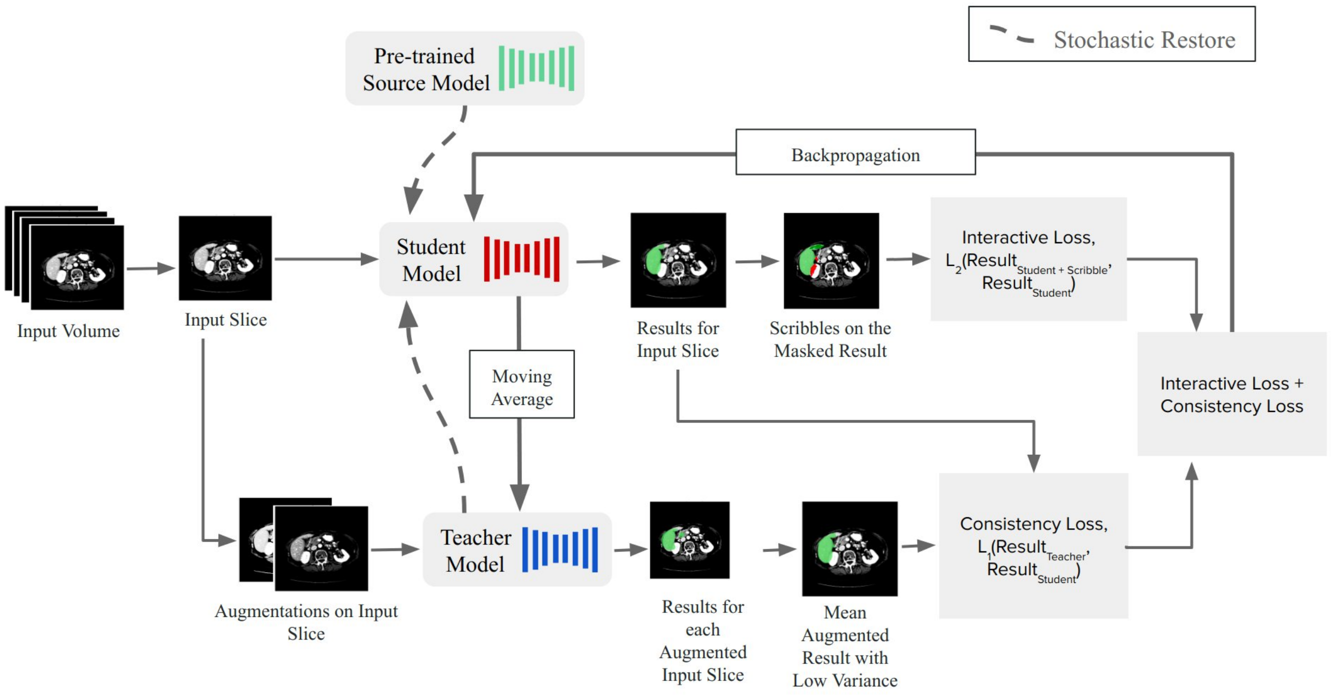

Stochastic Restore
Pre-trained Source Model
Backpropagation
Input Volume
Input Slice
Student Model
Results for Input Slice
Scribbles on the Masked Result
Interactive Loss, L$_2$(Result$_{\text{Student + Scribble}}$, Result$_{\text{Student}}$)
Moving Average
Interactive Loss + Consistency Loss
Augmentations on Input Slice
Teacher Model
Results for each Augmented Input Slice
Mean Augmented Result with Low Variance
Consistency Loss, L$_1$(Result$_{\text{Teacher}}$, Result$_{\text{Student}}$)

Figure 1: Overview of the proposed framework. Our framework can use various pre-trained network architectures depending on the application. Therefore, we do not provide the detailed architecture of any specific model. The first step in our framework is to obtain an initial segmentation using a pre-trained deep-learning network. The user can then review the segmentation and add scribbles where they want corrections made.

[13]). Indeed, CL for image classification is more popular [3, 7, 12, 42], but its application in segmentation tasks is also gaining traction [6, 11, 12, 24]. Many CL techniques assume access to target data during training; whereas others use a subset of the source data for test-time adaptation. Another strategy for adapting the model during inference (test time) involves leveraging prediction entropy [18, 35] or self-training with pseudo-labels [36]. Previously, during student training, their weights were either restored with the teacher at each step or reset using updated teacher parameters [36]. However, this method could lead to the student adopting erroneous weights from the teacher or source network.

3 Methodology

Interactive segmentation of medical images involves iteratively annotating and refining the output segmentation masks from an ML model. This can be accomplished for a CT or MRI volume by repeating the procedure slice by slice using specialized interactive image segmentation techniques or prompting foundation models. However, this is inefficient as the approach overlooks the relationship between adjacent slices.

We present an interactive segmentation method that can continuously adapt model parameters for the new target medical image (3D CT/MRI volume) using any pre-trained, off-the-shelf model. We describe the problem formulation for 3D medical image interactive segmentation and how we connect it to continual adaptation. This is followed by our methodology section, where we show how to adapt a pre-trained model to a new target medical image using interactive segmentation and continual adaptation techniques.

3.1 Problem Definition

Continuous adaptation of a pre-trained interactive segmentation model can be formally defined as follows. Let $f_{\theta_0}(x)$ be the pre-trained segmentation model with parameters θ_0 learned from training on the source training data D_s. During testing, the source training data is usually not available due to privacy concerns. At this stage, the model encounters test data D_t which is not available during annotation. Our objective is to improve the segmentation performance of a pre-trained model on an unseen medical image volume while maintaining the model's performance on the previously trained data. Note that maintaining performance on the previously trained data is a proxy objective to ensure that the model does not overfit on the current slice, leading to lower accuracy and, hence, more efforts in interactive segmentation for future slices.

3.2 Proposed Framework

Our approach proposes an iterative teacher-student architecture with interactive segmentation, where the student is supposed to adapt to a new domain quickly, meanwhile, the teacher accumulates the newly learned knowledge without catastrophic forgetting. The overview of our method can be found in Figure 1.

Method		Time Taken for the Datasets (in min.)				
		CHAOS	LiTS	AMOS	CHAOS	DSB
		(84 CT Sl.)	(104 CT Sl.)	(243 CT Sl.)	(70 MRI Sl.)	(Cell Image)
Manual Annot. Time		63.00	84.00	140.00	57.00	11.00
Annot. Time$^{(prev.\ res)}$		32.00	39.00	79.00	39.00	-
GrabCut [28]	User Time	66.10	67.00	89.00	61.30	8.00
	Machine Time	30.00	26.00	36.00	22.00	5.00
f-BRS [31]	User Time	52.18	58.43	100.65	40.3	8.00
	Machine Time	5.35	6.34	15.00	9.52	3.00
RiTM [32]	User Time	48.82	54.31	110.75	42.30	8.00
	Machine Time	4.79	5.78	13.44	8.92	2.00
iSegFormer [20]	User Time	35.23	38.55	-	-	3.00
	Machine Time	5.21	4.66	-	-	1.32
PseudoClick [21]	User Time	30.44	35.45	67.22	40.11	3.00
	Machine Time	5.01	5.45	12.86	8.44	1.12
FocalClick [8]	User Time	25.33	31.50	-	-	4.00
	Machine Time	4.32	5.21	-	-	0.78
Our Method	User Time	9.24	24.41	40.23	36.72	3.00
(UNet as Backbone)	Machine Time	4.45	5.50	12.85	8.47	1.40

Table 1: For each method, we report the total time (in minutes) to reach a Dice Score of 0.95 for the CT Volume. Our method reaches the target with the least User Time among all the methods. Here, 'Manual Annot. Time' gives the time taken when the user performs the annotation from scratch on each slice. On the other hand 'Annot Time (prev. res)' give the time taken for the annotation when the annotator copies the annotation of the last slice and makes the necessary changes to fit it on the current slice. The term 'SL' is used to represent the number of slices. "84 CT SL" means that we have 84 Slices of the CT Modality.

We utilize the pre-trained segmentation model with parameters θ_0 that takes a slice from a medical volume x_t as input. We initialize the teacher model ($f_{\theta^T}(x_t)$) and the student model ($f_{\theta^S}(x_t)$) with the same pre-trained interactive segmentation weights (θ), i.e., $\theta^T = \theta_0$ and $\theta^S = \theta_0$. We generate multiple outputs, $\hat{y}_i^T$, using the teacher model by applying test-time augmentations [33]. We exclusively apply intensity-based transforms, including Gaussian noise, blurring, and pixel intensity inversion. However, traditional TTA methods are unable to adapt to domain shifts between the source and target data distributions [26, 30]. We overcome this challenge by modulating the number of augmentations. To adapt, we calculate the teacher model's confidence based on the softmax function applied to its output masks [29] and extract the pixel-wise confidence of the output masks. We then evaluate the ratio of the number of pixels with prediction confidence between 0.4 and 0.6 to the total number of pixels, which we denote as ratio($f_{\theta^T}(x_t)$). The range of 0.4 to 0.6 is chosen because the prediction confidences are probability values, and a value of 0.5 indicates that the model assigns equal probabilities to both classes. Thus, this range reflects uncertain predictions where the model is less confident about assigning a pixel to a specific class. This ratio regulates the number of augmentations. Equation 1 shows the number of augmentations used based on this ratio.

$$y'_{t^T} = \frac{1}{N}\sum_{j=1}^{N} f_{\theta^T}(\tilde{x}_t^j) \begin{cases} N = 4 & \text{for} \quad \text{ratio}(f_{\theta^T}(x_t)) \le 0.3 \\ N = 16 & \text{for} \quad 0.3 < \text{ratio}(f_{\theta^T}(x_t)) \le 0.7 \\ N = 32 & \text{for} \quad 0.7 < \text{ratio}(f_{\theta^T}(x_t)) < 1 \end{cases} \tag{1}$$

Here, y'_{t^T} represents the average segmentation mask from the teacher model, $f_{\theta^T}(\tilde{x}_t^j)$ denotes the augmented version of the target image x_t, and N signifies the number of augmentations applied, ($f_{\theta^T}(x_t)$) is the source pre-trained model's prediction confidence on the current input $\tilde{x}_t$. We apply temperature scaling [19] with a parameter $\tau > 1$ to make the obtained segmentation masks less prone to overconfidence. We hereby obtain the average segmentation mask, using various $\hat{y}_T^i$, denoted as $\hat{y}_T^a$.

For each slice, x_t, the student model generates an initial segmentation mask, $\hat{y}^S = f_{\theta^S}(x)$. We calculate the cross-entropy loss by comparing the student's segmentation mask and the teacher's mean segmentation mask at each pixel as follows:

$$\mathcal{L}_1 = -\text{sum}\left(\sum_c \hat{y}_T^a[:, c]\log\left(\hat{y}_S[:, c]\right)\right). \tag{2}$$

Here, we only focus on the pixels within the segmentation mask that exhibit low variance across the augmented versions. This prioritizes regions with consistent predictions across different augmentations, leading to a more reliable pseudo-segmentation mask.

We capture user input through a scribble mask m and label $\hat{y}_U$ to incorporate user feedback. Here, m is a n dimensional binary vector with 1 indicating user feedback provided for the pixel in the form of scribble, and 0 otherwise. $\hat{y}_U$ is a $n \times c$ dimensional matrix, with each row representing a 1-hot vector indicating the user provided label at the pixel, and zero vector at other pixels. To simulate *thick* user scribbles, we blur the mask m and copy the user-provided label on each non-zero pixel in m. For capturing multiple scribbles, we

Algorithm 1 Interactive Medical Image Segmentation using Scribbles through Continuous Adaptation

- **Input:** Input Image x_t, User Interaction (Scribbles) u_t
- **Output:** Refined Segmentation Mask M
- **Requirement:** Model $f_{\theta^0}(x)$ with the pre-trained weights θ_0, learning rate λ, hyperparameters α, τ, k
- **Initialization:** Teacher model ($f_{\theta^T}(x)$) and the Student model ($f_{\theta^S}(x)$) are initialized with pre-trained interactive segmentation $f_{\theta^0}(x)$

```
procedure $u_t = u_{t_{prev}} \cup$ New Scribbles
    for $i = 1 \rightarrow k$ do
        for $j = 1 \rightarrow n$ do    ▷ $n$ = number of augmentations based on 1
            $M_T^j \leftarrow f_{\theta^T}(\tilde{x}_t^j, \tau)$ ▷ perform inference using $\theta_T$ and scale it using $\tau$
            $\hat{M}_T^j \leftarrow M_T^j$ ▷ Only pixels variance less than 0.05 are allowed
        $M_T \leftarrow \frac{1}{n}\sum_{j=1}^{N} \hat{M}_T^j$    ▷ Average Teacher Segmentation Mask
        $M_S \leftarrow f_{\theta^T}(\tilde{x}_t)$    ▷ Student Segmentation Mask
        ▷ Determine $l_1$ from $\mathcal{L}_1$ using Equation 2
        $M \leftarrow M_S \cup u_t$
        ▷ Determine $l_2$ from $\mathcal{L}_2$ using Equation 3
        ▷ Determine $l_t$ from $\mathcal{L}_t$
        $\theta_{t+1}^S \leftarrow \theta_t^S - \lambda \frac{d}{d\theta^S}\mathcal{L}_t(M, M_S, M_T, \theta^S, \theta^T)$    ▷ Update Student Weights using Equation 4
        if $l_{best} \geq l_t$ then
            $\theta_{t+1}^T \leftarrow \alpha\,\theta_t^T + (1-\alpha)\,\theta_{t+1}^S$    ▷ Update Teacher Weights via EMA
            $l_{best} = l_t$
            Stochastically restore student weights $\theta^S$
```

repeat the above process for each scribble separately and then do an element-wise add for both m and $\hat{y}_U$, such that if there is an overlap between the two masks, we consider none. In an abuse of notation, we denote the added mask and label matrix using m and $\hat{y}_U$, respectively. The interactive loss, $\mathcal{L}_2$, is computed as follows:

$$\mathcal{L}_2 = -\mathrm{sum}\left(\sum_c m[:]\ \hat{y}_U[:, c] \log\left(\hat{y}_S[:, c]\right)\right). \tag{3}$$

We define the cumulative loss, $\mathcal{L}_t$, as the sum of individual loss terms $\mathcal{L}_1$ and $\mathcal{L}_2$. We update student model parameters from $\theta_t^S \rightarrow \theta_{t+1}^S$ by performing standard stochastic gradient descent on $\mathcal{L}_t$. Following the student model update, we also update the teacher model. This ensures that the teacher retains valuable knowledge from previous training while progressively accumulating new information acquired through student adaptation. This knowledge transfer is achieved by updating the teacher model's parameters with an exponential moving average (EMA) of the student's parameters:

$$\theta_{t+1}^T = \alpha\,\theta_t^T + (1-\alpha)\,\theta_t^S. \tag{4}$$

Here α is a smoothing factor. Our final prediction for the slice x is the segmentation mask obtained from the updated teacher model.

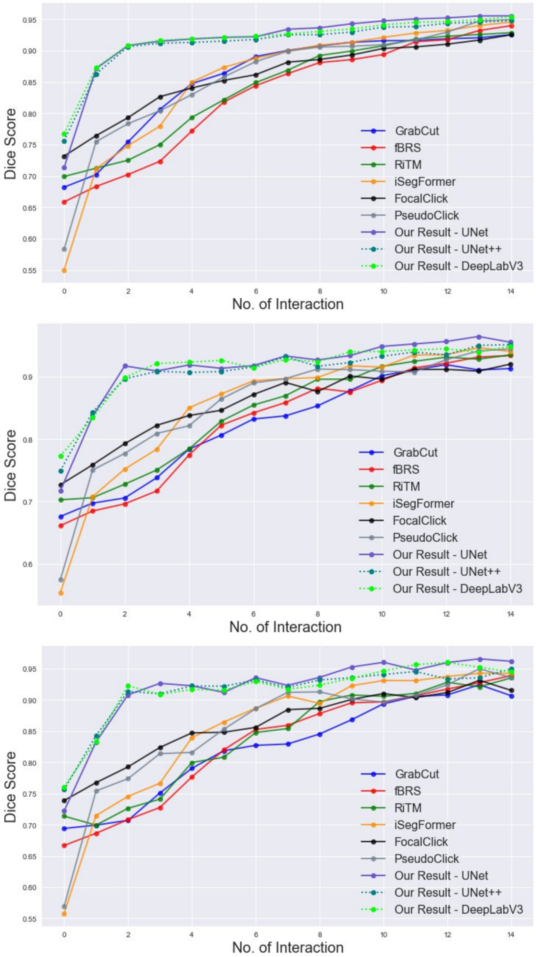
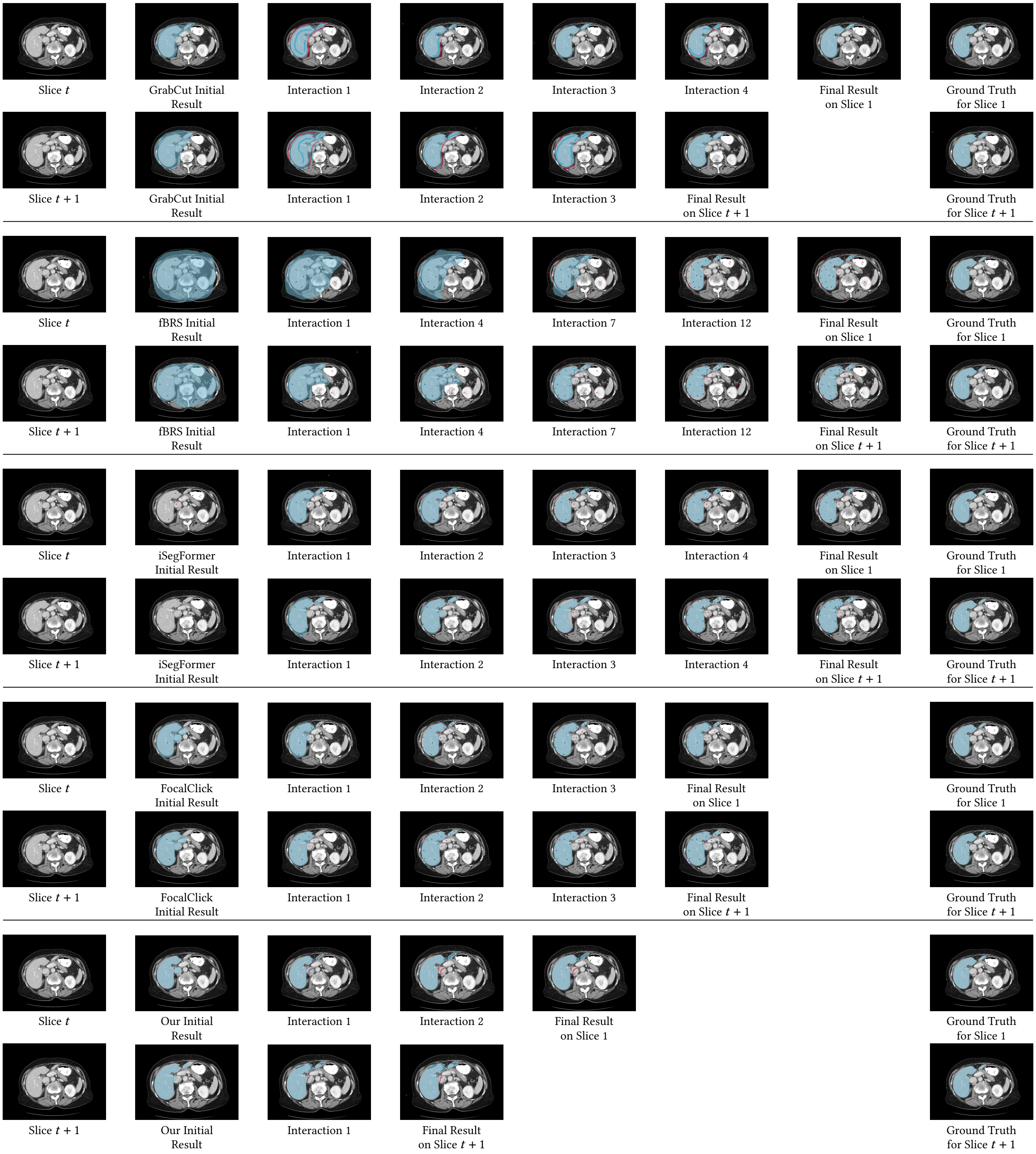


Figure 2: Improvement in Dice Score per user interaction: Our framework achieves the highest accuracy with minimum user interactions. 1st row: CHAOS(CT), 2nd row: LiTS (CT), and 3rd row: DSB (Cell Image)

This EMA approach helps mitigate catastrophic forgetting, a critical challenge in continual learning scenarios. The specific details of the procedure are outlined in Algorithm 1.

4 Experiments and Results

We have performed tests on datasets primarily composed of medical images. These datasets encompass the Combined (CT-MR) Healthy Abdominal Organ Segmentation (CHAOS) dataset, which is employed for the segmentation of four abdominal organs - spleen, right kidney, left kidney, and liver [16], and liver segmentation dataset - Liver Tumor Segmentation (LiTS) [4]. Moreover, we have utilized the Abdominal multi-organ segmentation (AMOS) dataset, which contains 15 abdominal organs, including the ones mentioned above, along with the gallbladder, esophagus, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, and prostate/uterus [15]. We use the Data Science Bowl dataset for microscopic cell segmentation [5].



Slice t | GrabCut Initial Result | Interaction 1 | Interaction 2 | Interaction 3 | Interaction 4 | Final Result on Slice 1 | Ground Truth for Slice 1

Slice $t + 1$ | GrabCut Initial Result | Interaction 1 | Interaction 2 | Interaction 3 | Final Result on Slice $t + 1$ | Ground Truth for Slice $t + 1$

Slice t | fBRS Initial Result | Interaction 1 | Interaction 4 | Interaction 7 | Interaction 12 | Final Result on Slice 1 | Ground Truth for Slice 1

Slice $t + 1$ | fBRS Initial Result | Interaction 1 | Interaction 4 | Interaction 7 | Interaction 12 | Final Result on Slice $t + 1$ | Ground Truth for Slice $t + 1$

Slice t | iSegFormer Initial Result | Interaction 1 | Interaction 2 | Interaction 3 | Interaction 4 | Final Result on Slice 1 | Ground Truth for Slice 1

Slice $t + 1$ | iSegFormer Initial Result | Interaction 1 | Interaction 2 | Interaction 3 | Interaction 4 | Final Result on Slice $t + 1$ | Ground Truth for Slice $t + 1$

Slice t | FocalClick Initial Result | Interaction 1 | Interaction 2 | Interaction 3 | Final Result on Slice 1 | Ground Truth for Slice 1

Slice $t + 1$ | FocalClick Initial Result | Interaction 1 | Interaction 2 | Interaction 3 | Final Result on Slice $t + 1$ | Ground Truth for Slice $t + 1$

Slice t | Our Initial Result | Interaction 1 | Interaction 2 | Final Result on Slice 1 | Ground Truth for Slice 1

Slice $t + 1$ | Our Initial Result | Interaction 1 | Final Result on Slice $t + 1$ | Ground Truth for Slice $t + 1$

Figure 3: Comparison on two LiTS slices. Left to right: original images, initial mask, Results after certain interactions, final segmentation mask, and Ground Truth. We have shown the results for Slice t and $t + 1$ (in this case, slices 55 and 56) for GrabCut, fBRS, iSegFormer, FocalClick, and Our Results for the consecutive slices. We have evaluated the same results for other datasets as well in our supplementary work

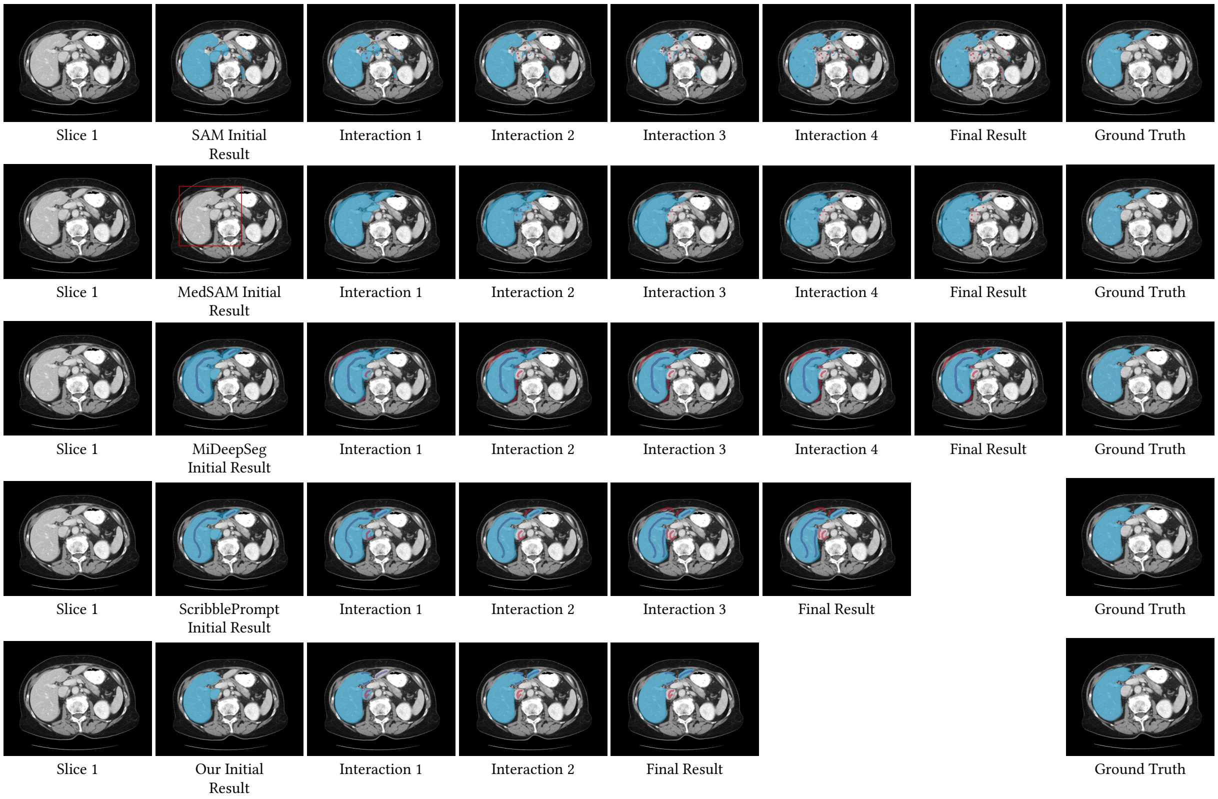


Figure 4: Comparison on LiTS slice for Foundational Models. Left to right: original images, initial mask, Results after each interaction, final segmentation mask, and Ground Truth. We have shown SAM, MedSAM, MiDeepSeg, ScribblePrompt and Our Result results. We have evaluated the same results for other datasets added in our supplementary work

For each dataset, we show the Dice Score [10] after k user interactions and the number of interactions to reach a particular target Dice Score on the slice.

Our experiments show comparison while utilizing various pretrained backbone models, such as UNet, UNet++, and DeepLabv3. This showcases the ability of our framework to use arbitrary models as the backbone.

Comparison with state of the art. We conducted a comprehensive evaluation of our interactive segmentation approach against other state-of-the-art methods on four medical image datasets: CHAOS (MRI Slice data), LiTS, AMOS, and the 2018 Data Science Bowl (microscopic cell data) (all primarily CT Volume data), along with the 95% confidence interval (CI) of segmentation obtained using manually provided scribbles and clicks over three interaction loops across different methods. Each interaction loop may involve multiple scribbles or clicks. Table 1 showcases the significant efficiency gains achieved by our framework. Our method significantly reduces the User's Time on all datasets, as tested by a medical expert. The first row displays the results when the medical expert annotated the slices in the volume from scratch (without external help). The second row shows the results of the experts' annotation aided by tools. Compared to complete human annotation, our method reduces user annotation time by a factor of 4.60 (CHAOS - CT), 2.80 (LiTS), 2.64 (AMOS), 6.72 (CHAOS - MRI), and 3.66 (DSB) on the respective datasets. Similarly impressive gains are observed when compared to human annotation aided by tools, with factors of 2.33 (CHAOS), 1.30 (LiTS), 1.46 (AMOS), and 1.11 (CHAOS - MRI). We compared the accuracy of segmentation after each user interaction for each technique. Figure 2 shows that our framework outperforms all other methods regarding accuracy and the number of iterations taken to achieve it. Additionally, our method can accomplish a dice score exceeding 0.9 with an average of less than four user iteration loops. Our method outperforms baseline methods for all the user interaction procedures at all numbers of interactions. Figure 3 shows the visual results. Here, we show the visual results for all interactions and the final results for segmenting two consecutive slices using all the mentioned methods.

Comparison with prompting foundation models. We compared our interactive segmentation method against existing generalized approaches, focusing on baselines like SAM [17], MedSAM [23], ScribblePrompt [38], and MIDeepSeg [22]. We evaluated (ViT-h) versions of the Segment Anything Model (SAM) [17] trained on natural images. SAM takes bounding boxes, clicks, and the logits of the previous prediction as input for segmentation. MedSAM [23] builds upon a ViT-B SAM model, further fine-tuned with bounding box prompts on a dataset of 1.5 million biomedical image segmentation pairs. MIDeepSeg [22] is an interactive segmentation framework designed for unseen tasks on medical images. It starts with

Methods	Dataset					
	CHAOS (CT)		LiTS (CT)		DSB (Cell)	
	Int. Dice	Final Dice	Int. Dice	Final Dice	Int. Dice	Final Dice
SAM [17]	0.628	0.741	0.573	0.772	0.498	0.521
MedSAM [23]	0.755	0.804	0.786	0.838	0.667	0.685
MIDeepSeg [22]		0.853	-	0.837	-	-
ScribblePrompt [38]	0.843	0.903	0.847	0.913	0.853	0.920
Our Method (UNet)	0.843	0.928	0.854	0.934	0.887	0.929
Our Method (UNet++)	0.881	0.933	0.898	0.912	0.867	0.931
Our Method (DeepLab-V3)	0.889	0.931	0.891	0.936	0.881	0.942

Table 2: User Interaction. We have effectively determined the mean Dice Score for segmentations predicted by different methods after ten rounds of user interaction. During each interaction loop, users can provide multiple scribbles, bounding boxes, and clicks to facilitate the segmentation process. We terminate the interaction loop if the method attains a dice score of 0.9.

interior margin points (positive clicks) and crops the image based on these points before utilizing a CNN for initial prediction. We evaluated the pre-trained MIDeepSeg model, initially developed for placenta T2 MRI segmentation tasks. Lastly, ScribblePrompt [38] leverages various user guidance modalities like scribbles, clicks, and bounding boxes for interactive medical image segmentation. ScribblePrompt inputs include positive and negative scribble and click prompts alongside bounding boxes and the predicted segmentation results to generate final predictions. Table 2 shows the comparison with prompting various foundational models. Our method outperforms the foundation models, achieving a dice score of 0.931, 0.936, and 0.942, compared to ScribblePrompt's scores of 0.903, 0.913, and 0.929, SAM's scores of 0.741, 0.772 and 0.521, MedSAM's scores of 0.804, 0.838 and 0.685, for CHAOS, LiTS, and the Data Science Bowl (DSB) dataset, respectively. We observed that the foundation models produce inadequate results after ten interactions, and performance stagnates. In our experiments with SAM and MedSAM, we observed that Clicks can often misguide the model and lead to the production of the output the user did not expect. SAM does not generalize well to click inputs (which they were not trained for). MedSAM has better predictions than other SAM baselines using the SAM architecture; however, it primarily relies on the initial bounding box as input and cannot use negative clicks. Scribbles, on the other hand, are an intuitive form of interaction that the algorithm can leverage to understand better the segmentation desired by the user. Figure 4 displays visual results for a slice's interactions and final segmentation using the foundation models.

5 Conclusion

This paper proposes a continual test time adaptation framework for interactive segmentation of medical image volumes. Our approach dynamically updates model parameters during inference based on user-provided scribbles and retains the knowledge gained from interactions on previous slices using the proposed continual adaptation framework. Consistently effective and efficient performance on unseen datasets with diverse modalities (CT, MRI, microscopic cells) validates the superiority of our method over the current state of the art, including prompting foundation models.

Acknowledgments

The authors would like to thank the I-Hub Foundation for Robotics, the Technology Innovation Hub of IIT Delhi, and IIIT Hyderabad for providing financial support for this work. The authors also thank CVIT for valuable discussions and support throughout the process.

References

[1] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. 2018. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 859–868.

[2] Eirikur Agustsson, Jasper RR Uijlings, and Vittorio Ferrari. 2019. Interactive full image segmentation by considering all regions jointly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11622–11631.

[3] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*. 139–154.

[4] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. 2023. The liver tumor segmentation benchmark (lits). *Medical Image Analysis* 84 (2023), 102680.

[5] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, et al. 2019. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature methods* 16, 12 (2019), 1247–1253.

[6] Fabio Cermelli, Dario Fontanel, Antonio Tavera, Marco Ciccone, and Barbara Caputo. 2022. Incremental learning in semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4371–4381.

[7] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2018. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420* (2018).

[8] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. 2022. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1300–1309.

[9] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 424–432.

[10] Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology* 26, 3 (1945), 297–302.

[11] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. 2021. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4040–4050.

[12] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. 2020. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XX 16*. Springer, 86–102.

[13] Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences* 3, 4 (1999), 128–135.

[14] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. 2022. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 574–584.

[15] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems* 35 (2022), 36722–36732.

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

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

[18] Jian Liang, Dapeng Hu, and Jiashi Feng. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*. PMLR, 6028–6039.

[19] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690* (2017).

[20] Qin Liu, Zhenlin Xu, Yining Jiao, and Marc Niethammer. 2022. iSegFormer: interactive segmentation via transformers with application to 3D knee MR images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 464–474.

[21] Qin Liu, Meng Zheng, Benjamin Planche, Srikrishna Karanam, Terrence Chen, Marc Niethammer, and Ziyan Wu. 2022. Pseudoclick: Interactive image segmentation with click imitation. In *European Conference on Computer Vision*. Springer, 728–745.

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

[23] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications* 15, 1 (2024), 654.

[24] Umberto Michieli and Pietro Zanuttigh. 2021. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1114–1124.

[25] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*. Ieee, 565–571.

[26] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. 2018. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8571–8580.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

[28] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. " GrabCut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)* 23, 3 (2004), 309–314.

[29] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems* 31 (2018).

[30] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. 2021. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1214–1223.

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

[32] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. 2022. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3141–3145.

[33] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*. PMLR, 9229–9248.

[34] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. 2020. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical image analysis* 63 (2020), 101693.

[35] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726* (2020).

[36] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. 2022. Continual Test-Time Domain Adaptation. In *Proceedings of Conference on Computer Vision and Pattern Recognition*.

[37] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. 2022. Medical image segmentation using deep learning: A survey. *IET Image Processing* 16, 5 (2022), 1243–1267.

[38] Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. 2023. ScribblePrompt: Fast and Flexible Interactive Segmentation for Any Medical Image. *arXiv e-prints* (2023), arXiv–2312.

[39] Jiajun Wu, Yibiao Zhao, Jun-Yan Zhu, Siwei Luo, and Zhuowen Tu. 2014. Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 256–263.

[40] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. 2021. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 171–180.

[41] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. 2017. Deep grabcut for object selection. *arXiv preprint arXiv:1707.00243* (2017).

[42] Shipeng Yan, Jiangwei Xie, and Xuming He. 2021. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3014–3023.

[43] Jianpeng Zhang, Yutong Xie, Yan Wang, and Yong Xia. 2020. Inter-slice context residual learning for 3D medical image segmentation. *IEEE Transactions on Medical Imaging* 40, 2 (2020), 661–672.

[44] Shiyin Zhang, Jun Hao Liew, Yunchao Wei, Shikui Wei, and Yao Zhao. 2020. Interactive object segmentation with inside-outside guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12234–12244.

[45] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2019. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging* 39, 6 (2019), 1856–1867.