MultiverSeg: Scalable Interactive Segmentation of Biomedical Imaging Datasets with In-Context Guidance

Supplementary Material

Contents

A. Code

Code and pre-trained weights are available at https://multiverseg.csail.mit.edu.

B. MultiverSeg Method

B.1. Architecture

CrossConv. We implement the CrossConvolutional layer slightly differently from [11]. To avoid duplicate convolutions on the context features v_i in Eq. 2, we partition weights θ_z channel-wise into $\{\theta_{z_1}, \theta_{z_2}\}$ and implement $z_i = LN(A(\text{Conv}(q, \theta_{z_1}) + \text{Conv}(v_i, \theta_{z_2})))$ where q is the target feature map and v_i is the feature map corresponding to context set entry i. We zero out the bias terms in $\text{Conv}(\cdot, \theta_{z_2})$ such that the computation is equivalent to $z_i = LN(A(\text{Conv}(q||v_i; \theta_z)))$.

Network. We implement $f_\theta(\cdot)$ using an encoder with 5 encoder CrossBlock stages and a decoder with 4 CrossBlock stages. Each stage has 256 output features and LeakyReLU non-linearities after each convolution. We use bilinear interpolation for upsampling and downsampling.

The CrossBlock mechanism requires at least one context set entry. If the context set is empty, we use a dummy context set entry consisting of an image and segmentation with uniform value of 0.5.

C. Data

C.1. Datasets

We build on large dataset gathering efforts like MegaMedical [11, 71, 83] to compile a collection of 79 open-access biomedical imaging datasets for training and evaluation, covering over 54k scans, 16 image types, and 713 labels.

Division of Datasets. The division of datasets and subjects for training, model selection, and evaluation is summarized in Tab. 1. The 79 datasets were divided into 67 training datasets (Tab. 3 and 12 evaluation datasets (Tab. 2). Data from 9 (out of 12) of the evaluation datasets were used for model selection and final evaluation. The other 3 evaluation datasets were completely held-out from model selection and only used in the final evaluation.

Division of Subjects. We split each dataset into 60% train, 20% validation, and 20% test by subject. We used the "train" splits from the 67 training datasets to train MultiverSeg models. We use the "validation" splits from the 67 training datasets and 9 validation datasets to select the best model checkpoint. We report final evaluation results across 12 held-out "test" splits of the 12 evaluation datasets to maximize the diversity of tasks and modalities in our evaluation set (Tab. 2). No data from the 9 validation datasets or 3 test datasets were seen by MultiverSeg during training.

Task Definition. We define a 2D segmentation task as a combination of (sub)dataset, axis (for 3D modalities), and label. For datasets with multiple segmentation labels, we consider each label separately as a binary segmentation task. For datasets with sub-datasets (e.g., malignant vs. benign lesions) we consider each cohort as a separate task. For multi-annotator datasets, we treat each annotator as a separate label. For instance segmentation datasets, we considered all instances as a single label.

3D Datasets. For 3D modalities, we use the slice with maximum label area ("maxslice") and the middle slice ("midslice") for each volume for training of MultiverSeg. For the 3D evaluation datasets (BTCV Cervix [45], ACDC [7], SCD [70], SpineWeb [88], COBRE [2], TotalSegmentator [82]) we evaluated the slice with the maximum label area for each subject, as in [83]. We also considered evaluating on the middle slice, as in [11, 71, 84] and saw similar trends on the validation data. However, we opted for evaluation on maxslices because for our 3D test datasets (COBRE, TotalSegmentator) some labels do not appear in the midslices. Due to the large number of tasks in COBRE and TotalSegmentator, we only consider coronal slices from these datasets for evaluation.

Data Processing and Image Resolution. We rescale image intensities to [0,1], padded square with zeros. For training, we resized images to 128^2. In our final evaluations, we use images resized to 256^2. We show additional evaluations on 128^2 sized images in Appendix E.5.

Data Sampling. During training, we sample image, segmentation pairs hierarchically – by dataset and modality, axis, and then label – to balance training on datasets of different sizes.

C.2. Synthetic Task Generation

We introduce a new approach for constructing synthetic tasks from real images. Given a single image x_0, we construct a set of images $\{x'_i, y'_i\}_{i=1}^{m+1}$ representing a synthetic task. We then partition this set into a target example and context set of size m for training.

Related Work. Although previous work found that training on a mix of real and synthetic segmentation *labels* based on image superpixels is useful for improving generalization in interactive segmentation [83], we do not use such data here. That approach cannot be directly applied to MultiverSeg because it does not produce semantically consistent labels across multiple images.

Method. To build a synthetic task from an image, we first generate a synthetic label and then perform aggressive augmentations to create a set of images corresponding to the same synthetic task (Fig. 1).

Given an image x_0, we first generate a synthetic label y_{synth} by applying a superpixel algorithm [21] with scale parameter $\lambda \sim U[1, \lambda_{max}]$ to partition the image into a multi-label mask of k superpixels $z \in \{1, \ldots, k\}^{n \times n}$. We then randomly select a superpixel $y_{synth} = \mathbb{1}(z = c)$ as a synthetic label.

To generate a set of $m + 1$ images representing the same task, we duplicate (x_0, y_{synth}), $m + 1$ times and apply aggressive augmentations to vary the images and segmentation labels [11, 86].

Implementation. MultiverSeg was trained with $p_{synth} = 0.5$. We use a superpixel algorithm [21] with $\lambda \sim [1, 500]$. Tab. 4 lists the data augmentations.

C.3. Data Augmentation

Tab. 5 shows the within-task augmentations and task-augmentations used to train MultiverSeg [11, 71].

D. Experimental Setup

D.1. Baselines

We provide additional details on the baselines. We summarize the capabilities of our method and baselines in Tab. 6.

SAM. We evaluated SAM [40] (ViT-b) in both "single-mask" and "multi-mask" mode on our validation data, and average results were better using "single-mask" mode. We report final results for SAM on the test data using "single-mask" mode.

UniverSeg. Previous work found that ensembling UniverSeg predictions across multiple randomly sampled context sets improved Dice score [11]. We report results *without* ensembling to accurately reflect the mean Dice of predictions given a fixed size context set.

OnePrompt. OnePrompt [84] is a medical image segmentation model that can perform in-context segmentation of a target image given a single context example with scribble, click, bounding box or mask annotation on the context image. OnePrompt can also be used for interactive segmentation by using the same image as both the context image and the target image. We do not compare to OnePrompt because the pre-trained model weights are not publicly available. Recreating the data processing and retraining the model was beyond our computational capacity. For reference, the OnePrompt model required 64 NVIDIA A100 GPUs to train [84].

LabelAnything. LabelAnything [18] is an in-context segmentation model designed for few-shot multi-label segmentation of natural images. LabelAnything takes as input a target image to segment and a context set of images with multi-label mask, click, or bounding box annotations. We do not compare to LabelAnything because the pre-trained model weights are not publicly available. As with OnePrompt, recreating the data handling and retraining the model from scratch was beyond our computational capacity.

D.2. Inference

Image Resolution. MultiverSeg, ScribblePrompt, and UniverSeg, which were all developed and trained on 128^2 sized images, and output predictions at the same resolution. SAM was trained with 1024^2 sized inputs and predicts segmentations at 256^2 resolution. For each method, we resized the inputs to the method's training input size using bilinear interpolation before performing inference and then resized the output (as needed) to the evaluation resolution.

Table 1. **Dataset split overview**. Each dataset was split into 60% train, 20% validation and 20% test by subject. Data from the "train" splits of the 67 training datasets were used to train the models. The MultiverSeg models did not see any data from the validation datasets or test datasets during training. Data from the "validation" split of the 9 validation datasets was used for MultiverSeg (MVS) model selection and experimenting with different evaluation methods of baselines. We report final results on the held-out test splits of 12 evaluation datasets: data from the "test" splits of the 9 validation datasets and the "test" splits of the 2 test datasets. To train the fully-supervised nnUNet baselines, we used the training and validation splits of the 12 evaluation datasets.

		Split within each dataset by subject		
Dataset Group	No. Datasets	Training Split (60%)	Validation Split (20%)	Test Split (20%)
Training Datasts	67	MVS training	MVS model selection	Not used
Validation Datasets	9	nnUNet training	MVS and baselines model selection, nnUNet training	Final evaluation
Test Datasets	3	nnUNet training	nnUNet training	Final evaluation

Table 2. **Evaluation datasets**. We assembled the following set of datasets to evaluate MultiverSeg and baseline methods. For the relative size of datasets, we include the number of unique scans (subject and modality pairs) and labels that each dataset has. These datasets were unseen by MultiverSeg during training. Three datasets were completely held-out from model selection. The validation splits of the other 9 datasets were used for selecting the best model checkpoint. We report final results on the test splits of these 12 datasets.

Dataset Name	**Description**	**Scans**	**Labels**	**Modalities**
ACDC [7]	Left and right ventricular endocardium	99	3	cine-MRI
BTCV Cervix [45]	Bladder, uterus, rectum, small bowel	30	4	CT
BUID [3]	Breast tumors	647	2	Ultrasound
COBRE [2, 17, 22]	Brain anatomy	258	45	T1-weighted MRI
DRIVE [79]	Blood vessels in retinal images	20	1	Optical camera
HipXRay [29]	Ilium and femur	140	2	X-Ray
PanDental [1]	Mandible and teeth	215	2	X-Ray
SCD [70]	Sunnybrook Cardiac Multi-Dataset Collection	100	1	cine-MRI
SCR [80]	Lungs, heart, and clavicles	247	5	X-Ray
SpineWeb [88]	Vertebrae	15	1	T2-weighted MRI
TotalSegmentator [82]	104 anatomic structures (27 organs, 59 bones, 10 muscles, and 8 vessels)	1,204	104	CT
WBC [89]	White blood cell cytoplasm and nucleus	400	2	Microscopy

D.3. Metrics

Averaging. When reporting average performance for a dataset or across multiple datasets, we averaged metrics hierarchically by subject, label, axis, modality, subdataset, and then dataset.

Confidence Intervals. For Experiment 1, we calculate 95% confidence intervals over results from 200 simulations with different random seeds. For Experiment 2, we calculate 95% confidence intervals by bootstrapping over subjects with 100 runs.

E. Experiment 1: Evaluation

E.1. Setup

We illustrate the process of segmenting a set of images using MultiverSeg in Fig. 3

Procedure. For all methods, we interactively segment a seed image to 90% Dice using ScribblePrompt. This first image was randomly sampled (for each simulation round) from the training split. Since the number of interactions and the prediction for this seed image is the same for all methods, we exclude it from the reported results.

We report the number of interactions to achieve 90% Dice for each of the *next* 18 images from the held-out test split of our evaluation tasks. We conduct 200 rounds of simulations, randomly sampling 18 test images (without replacement) from each task and sequentially segmenting

Table 3. **Train datasets**. We train MultiverSeg on the following datasets. For the relative size of datasets, we have included the number of unique scans (subject and modality pairs) that each dataset has.

Dataset Name	Description	Scans	Modalities
AbdominalUS [81]	Abdominal organ segmentation	1,543	Ultrasound
AMOS [35]	Abdominal organ segmentation	240	CT, MRI
BBBC003 [52]	Mouse embryos	15	Microscopy
BBBC038 [12]	Nuclei instance segmentation	670	Microscopy
BrainDev [26, 27, 43, 75]	Adult and neonatal brain atlases	53	Multimodal MRI
BrainMetShare[28]	Brain tumors	420	Multimodal MRI
BRATS [4, 5, 64]	Brain tumors	6,096	Multimodal MRI
BTCV Abdominal [45]	13 abdominal organs	30	CT
BUSIS [85]	Breast tumors	163	Ultrasound
CAMUS [46]	Four-chamber and Apical two-chamber heart	500	Ultrasound
CDemris [36]	Human left atrial wall	60	CMR
CHAOS [37, 38]	Abdominal organs (liver, kidneys, spleen)	40	CT, T2-weighted MRI
CheXplanation [73]	Chest X-Ray observations	170	X-Ray
CoNSeP	Histopathology Nuclei	27	Microscopy
CT2US [78]	Liver segmentation in synthetic ultrasound	4,586	Ultrasound
CT-ORG[72]	Abdominal organ segmentation (overlap with LiTS)	140	CT
DDTI [67]	Thyroid segmentation	472	Ultrasound
DukeLiver [58]	Liver segmentation in abdominal MRI	310	MRI
EOphtha [19]	Eye microaneurysms and diabetic retinopathy	102	Optical camera
FeTA [66]	Fetal brain structures	80	Fetal MRI
FetoPlac [6]	Placenta vessel	6	Fetoscopic optical camera
FLARE [55]	Abdominal organs (liver, kidney, spleen, pancreas)	361	CT
HaN-Seg [68]	Head and neck organs at risk	84	CT, T1-weighted MRI
HMC-QU [20, 39]	4-chamber (A4C) and apical 2-chamber (A2C) left wall	292	Ultrasound
I2CVB [47]	Prostate (peripheral zone, central gland)	19	T2-weighted MRI
IDRID [69]	Diabetic retinopathy	54	Optical camera
ISBI-EM [13]	Neuronal structures in electron microscopy	30	Microscopy
ISIC [15]	Demoscopic lesions	2,000	Dermatology
ISLES [31]	Ischemic stroke lesion	180	Multimodal MRI
KiTS [30]	Kidney and kidney tumor	210	CT
LGGFlair [10, 62]	TCIA lower-grade glioma brain tumor	110	MRI
LiTS [8]	Liver tumor	131	CT
LUNA [76]	Lungs	888	CT
MCIC [25]	Multi-site brain regions of schizophrenic patients	390	T1-weighted MRI
MMOTU [87]	Ovarian tumors	1,140	Ultrasound
MSD [77]	Large-scale collection of 10 medical segmentation datasets	3,225	CT, Multimodal MRI
MuscleUS [61]	Muscle segmentation (biceps and lower leg)	8,169	Ultrasound
NCI-ISBI [9]	Prostate	30	T2-weighted MRI
NerveUS [65]	Nerve segmentation	5,635	Ultrasound
OASIS [32, 59]	Brain anatomy	414	T1-weighted MRI
OCTA500 [48]	Retinal vascular	500	OCT/OCTA
PanNuke [23]	Nuclei instance segmentation	7,901	Microscopy
PAXRay [74]	92 labels covering lungs, mediastinum, bones, and sub-diaphram in Chest X-Ray	852	X-Ray
PROMISE12 [49]	Prostate	37	T2-weighted MRI
PPMI [16, 60]	Brain regions of Parkinson patients	1,130	T1-weighted MRI
QUBIQ [63]	Collection of 4 multi-annotator datasets (brain, kidney, pancreas and prostate)	209	T1-weighted MRI, Multi-modal MRI, CT
ROSE [57]	Retinal vessel	117	OCT/OCTA
SegTHOR [44]	Thoracic organs (heart, trachea, esophagus)	40	CT
SegThy [41]	Thyroid and neck segmentation	532	MRI, Ultrasound
ssTEM [24]	Neuron membranes, mitochondria, synapses and extracellular space	20	Microscopy
STARE [33]	Blood vessels in retinal images	20	Optical camera
ToothSeg [34]	Individual teeth	598	X-Ray
VerSe [53]	Individual vertebrae	55	CT
WMH [42]	White matter hyper-intensities	60	Multimodal MRI
WORD [54]	Abdominal organ segmentation	120	CT

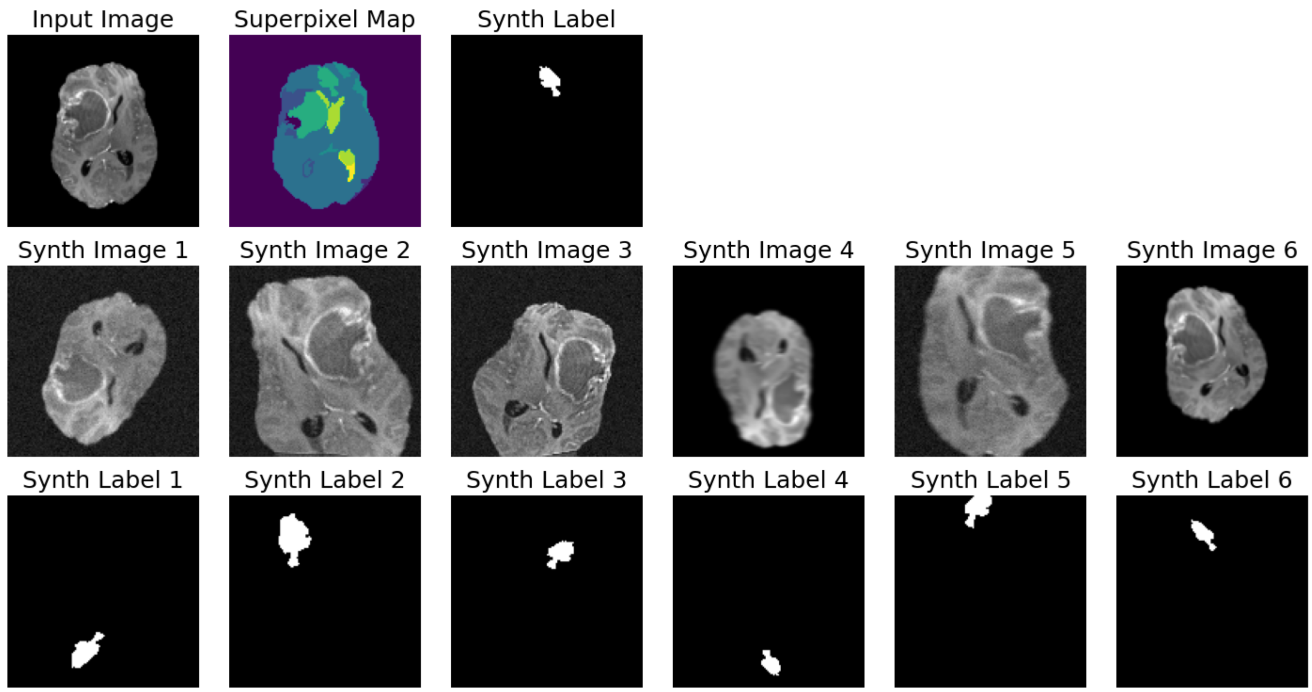


Figure 1. **Synthetic task generation example**. Given an input image, we apply a superpixel algorithm to generate a superpixel map of potential synthetic labels. We randomly sample one of the superpixels to serve as a synthetic label. Next, we duplicate the input image and synthetic label $m + 1$ times and apply data augmentations (Tab. 4) to vary the examples within the synthetic task. We use the first synthetic example as the target and the remaining m synthetic examples as the context set during training.

Augmentations	p	Parameters
Random Affine	0.8	degrees $\in [-25, 25]$ translation $\in [0, 0.2]$ scale $\in [0.9, 1.5]$
Brightness Contrast	0.5	brightness $\in [-0.1, 0.1]$ contrast $\in [0.5, 1.5]$
Elastic Transform	0.8	$\alpha \in [1, 10]$ $\sigma \in [8, 15]$
Sharpness	0.5	sharpness $= 5$
Gaussian Blur	0.5	$\sigma \in [0.1, 1.5]$ $k = 5$
Gaussian Noise	0.5	$\mu \in [0, 0.05]$ $\sigma \in [0, 0.05]$
Horizontal Flip	0.5	None
Vertical Flip	0.5	None

Table 4. **Data augmentations for generating synthetic tasks**. Given a set of $m + 1$ copies of the same example, we randomly sampled data augmentations for each instance to increase the diversity of examples within the task. Each augmentation is sampled with probability p.

them using each method. We use the same random seeds for each method, so the sampled examples are the same across methods for each simulation round.

Tasks. We exclude tasks with fewer than 18 test examples, leaving 161 tasks from 8 evaluation datasets [1–3, 7, 29, 80, 82, 89]. We selected this cutoff based on the distribution of task sizes in our validation data (Fig. 2) to focus on scenarios where a user wants to segment many similar images.

Data. We conducted our evaluation on 256^2 sized images. For each method, we resized the inputs to match the size of the model's training data before performing the forward pass, and then resized the prediction back to 256^2 before calculating the Dice Score. In Appendix E.5 we conduct a sensitivity analysis, performing the evaluation with 128^2 sized images

E.2. Interactions per Image as a Function of Dataset Size

Results by dataset. As more examples are segmented and the context set grows, the number of clicks and scribbles required to get to 90% Dice on the n^{th} example using MultiverSeg decreases substantially. Fig. 4 and Fig. 5 show results averaged by dataset. MultiverSeg and SP+UVS are less effective at reducing the number of clicks for tasks from BUID, a breast ultrasound lesion segmentation dataset, per-

Augmentations	p	Parameters
Random Affine	0.25	degrees $\in [-25, 25]$ translation $\in [0, 0.1]$ scale $\in [0.9, 1.1]$
Brightness Contrast	0.25	brightness $\in [-0.1, 0.1]$ contrast $\in [0.5, 1.5]$
Elastic Transform	0.8	$\alpha \in [1, 2.5]$ $\sigma \in [7, 9]$
Sharpness	0.25	sharpness $= 5$
Gaussian Blur	0.25	$\sigma \in [0.1, 1.0]$ $k = 5$
Gaussian Noise	0.25	$\mu \in [0, 0.05]$ $\sigma \in [0, 0.05]$

(a) **Within-Task Augmentations**

Augmentations	p	Parameters
Random Affine	0.5	degrees $\in [0, 360]$ translates $\in [0, 0.2]$ scale $\in [0.8, 1.1]$
Brightness Contrast	0.5	brightness $\in [-0.1, 0.1]$ contrast $\in [0.8, 1.2]$
Gaussian Blur	0.5	$\sigma \in [0.1, 1.1]$ $k = 5$
Gaussian Noise	0.5	$\mu \in [0, 0.05]$ $\sigma \in [0, 0.05]$
Elastic Transform	0.5	$\alpha \in [1, 2]$ $\sigma \in [6, 8]$
Sharpness	0.5	sharpness $= 5$
Horizontal Flip	0.5	None
Vertical Flip	0.5	None
Sobel Edges Label	0.5	None
Flip Intensities	0.5	None

(b) **Task Augmentations**

Table 5. **Augmentations used to train MultiverSeg.** Within-task data augmentations (**top**) are randomly sampled for each example within a task to increase the diversity within a task. Task augmentations (**bottom**) are randomly sampled for each task and then applied to all examples in a task to increase the diversity of tasks. Each augmentation is randomly sampled with probability p. We apply augmentations after (optional) synthetic task generation and before simulating user interactions.

haps due to the heterogeneity of examples in that dataset.

Tasks with more examples. We show results by task for three datasets with more than 18 test examples per task (Fig. 6, Fig. 7, and Fig. 8). For larger sets of images, using MultiverSeg results in even greater reductions in the total and average number of user interactions.

Context Set Quality. For MultiverSeg and SP+UVS, thresholding the predictions before adding them to the context set improved performance (Fig. 9). We use the validation split of our validation data (at 128^2 resolution) to select the best approach (soft or binary predictions in the context set) for each method.

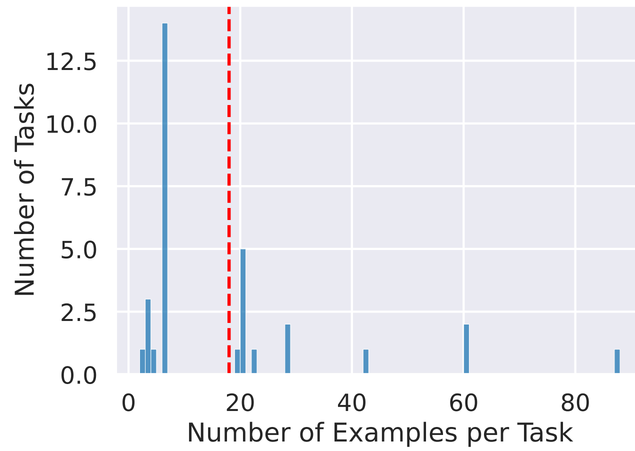


Figure 2. **Examples per task**. We visualize the distribution of examples per task in our validation data. We only consider tasks with at least 18 examples in Experiment 1.

MultiverSeg does not perform well when the context set contains *soft* predictions from previous examples, likely because it was trained with ground truth context labels. The number of interactions to 90% Dice is lowest when the context set contains ground truth labels, however this is not realistic in practice.

SP+UVS. Consistent with the original published results, we find that UniverSeg has poor performance for small context sets and initializing ScribblePrompt using the UniverSeg prediction hurts performance when the context set is small. In our final evaluation of SP+UVS, we set the minimum context set size to be 5 examples: when the context sets contains fewer than 5 examples, we ignore the context and only use ScribblePrompt to make predictions. Fig. 10 shows variations of SP+UVS with different minimum context set sizes on validation data at 128^2 resolution.

Total Interactions. Fig. 11 shows the total number of interactions, average Dice score, and average 95th percentile Hausdorff distance across all tasks.

E.3. Bootstrapping In-Context Segmentation

Setup. For UniverSeg [11], a non-interactive in-context segmentation method, we segment the dataset by bootstrapping from a single context example with ground truth segmentation. For each image in the dataset, we make an in-context prediction and then add the prediction to the context set for the next image until all images in the dataset have been segmented. As an upper bound on performance, we also evaluated using ground truth labels in the context set

Method	Interactive	In-Context	Interactive In-Context
SAM [40]	✓		
MedSAM [56]	✓		
SAM-Med2D [14]	✓		
SegNext [50]	✓		
ScribblePrompt [83]	✓		
UniverSeg [11]		✓	
LabelAnything [18]		✓	
OnePrompt [84]	✓	✓ (context size = 1)	
SP+UVS	✓	✓	✓
MultiverSeg (ours)	✓	✓	✓

Table 6. **Summary of segmentation methods**.

instead of previously predicted segmentations ("UniverSeg (oracle)").

Results. This approach did not produce accurate results, likely because UniverSeg has poor performance for small context sets and/or context sets with imperfect labels (Fig. 12a). Because UniverSeg does not have a mechanism to incorporate corrections, it was not possible to achieve 90% Dice for most images (Fig. 12b). Fig. 13 shows results by individual dataset.

Context Set Quality. As with other methods (MultiverSeg and SP+UVS), we experimented with thresholding the predictions at 0.5 before adding them to the context set. For UniverSeg, thresholding the predictions did not improve Dice scores compared to using the soft predictions in the context set.

E.4. Comparison to Few-Shot Fine-Tuning

One approach to segmenting a new dataset is to (interactively) segment a few images using a pre-trained foundation model, and then use those examples to train a task-specific interactive segmentation model by fine-tuning the foundation model. In this experiment, we simulated this process using ScribblePrompt.

Setup. For each task and random seed, we sampled 5 random test examples, and used ScribblePrompt to segment those images using simulated random center clicks. For each image of the 5 images, random center clicks were used to prompt ScribblePrompt until a maximum of 20 clicks was reached or the prediction surpassed 90% Dice. Then we used those newly labeled images to fine-tune ScribblePrompt from pre-trained weights. We randomly split the 5 images into 4 training examples and 1 validation example.

We fine-tuned ScribblePrompt using the same training interaction protocol, loss function, and data augmentations (Appendix C.3) as MultiverSeg minus synthetic task augmentations. Each task-specific model was fine-tuned for 300 epochs using the Adam optimizer with a learning rate of $1e^{-6}$ and batch size of 4. These hyperparameters were selected based on experiments with learning rate $\in \{1e^{-4}, 1e^{-5}, 1e^{-6}\}$ and batch size $\in \{4, 8\}$ using the cytoplasm segmentation task from the WBC [89] dataset. For each training run the best checkpoint was selected based on the validation example and then used to interactively segment 13 more test images (to complete the set of 18).

We repreated this procedure of labelling images and training tasks-specific models for 5 random seeds for each task. Due to the large number of tasks-specific models trained for this experiment, we trained and evaluated on images at 128^2 to reduce training time.

Runtime. Fine-tuning ScribblePrompt to produce *each* task-specific interactive segmentation model took on average 20 minutes on a NVIDIA A100 GPU. In contrast, MultiverSeg's inference time is < 150 milliseconds, even with a context set size of 64 examples (Appendix F.3).

Results. Fig. 14 shows MultiverSeg required fewer interactions than fine-tuning ScribblePrompt in 13 out of 16 scenarios. On average, the fine-tuning approach required 5.90 ± 0.10 clicks or 2.63 ± 0.13 scribble steps per image. MultiverSeg required fewer interactions: 4.64 ± 0.10 clicks or 4.64 ± 0.10 scribble steps per image.

E.5. Resolution Sensitivity Analysis

We conduct a sensitivity analysis, evaluating MultiverSeg and the baseline methods at 128^2 resolution.

Results. MultiverSeg outperforms the baselines with greater margins when evaluated at 128^2 resolution compared to 256^2 resolution. As more examples are segmented and the context set grows, the number of interactions required to get to 90% Dice (NoI90) on the n^{th} example using MultiverSeg decreases substantially (Fig. 15).

MultiverSeg required the fewest number of interactions per image on all datasets (Fig. 16). On average, using Mul-

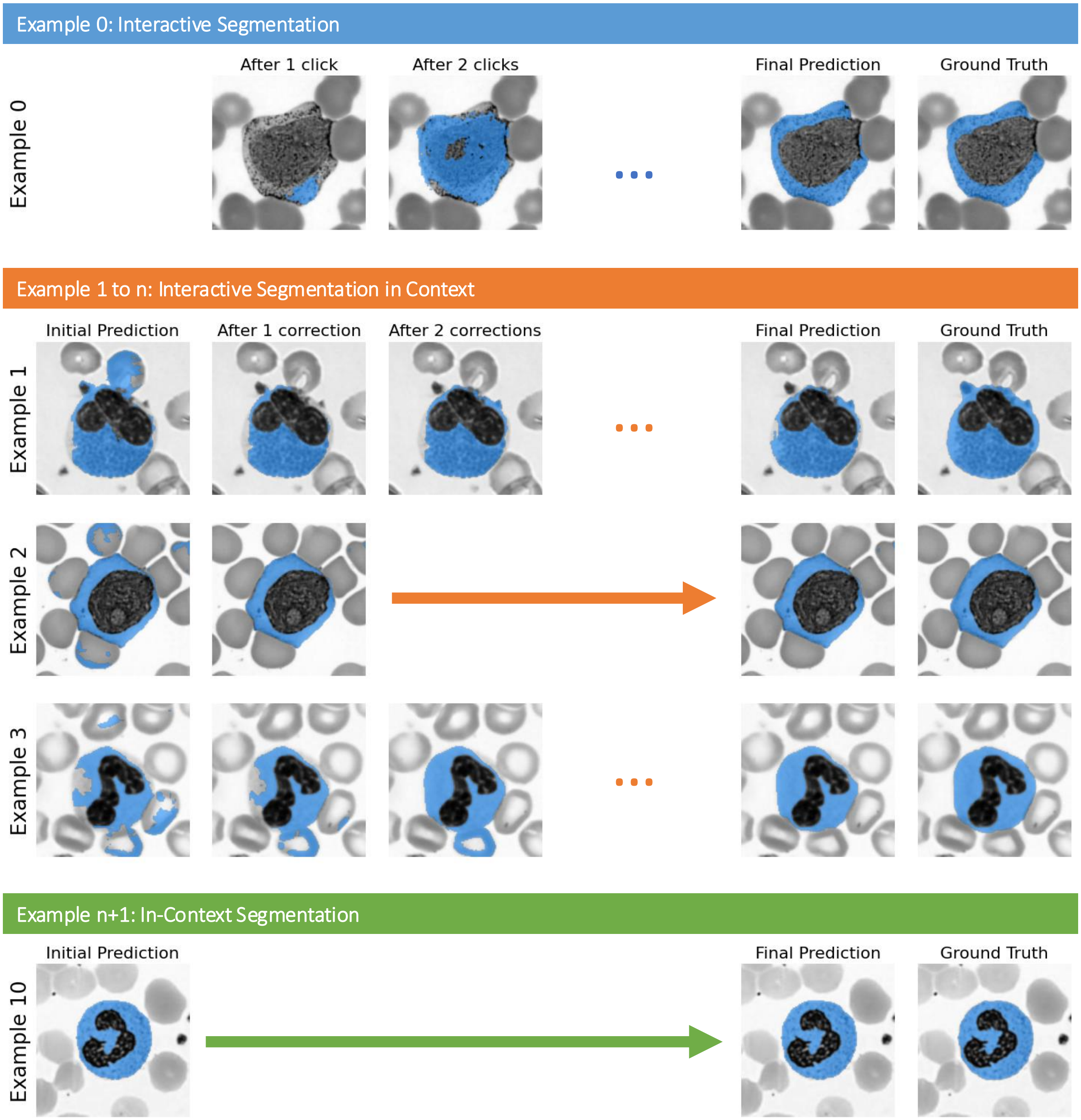


Figure 3. **Example segmentation process with MultiverSeg**. We begin by interactively segmenting a seed image (**Example 0**) to 90% Dice. The Example 0 image and final prediction are added to the context set for subsequent examples. For each subsequent example, we first make an initial in-context segmentation prediction using a context set containing all the previous examples and previously predicted segmentations. Then, we simulate center correction clicks until the predicted segmentation achieves $\geq 90\%$ Dice or we have accrued 20 clicks. For **Example 2**, we only simulated 1 correction because the prediction reached 90% Dice after 1 correction click. For **Example 1** and **Example 3**, additional correction clicks were needed. When the context set is large enough ($>n$), the in-context prediction from MultiverSeg may be accurate enough that no corrections are needed. For **Example 10**, the Dice score of the predicted in-context segmentation is greater than 90% so we do not need to simulate any corrections. In practice, n varies by task.

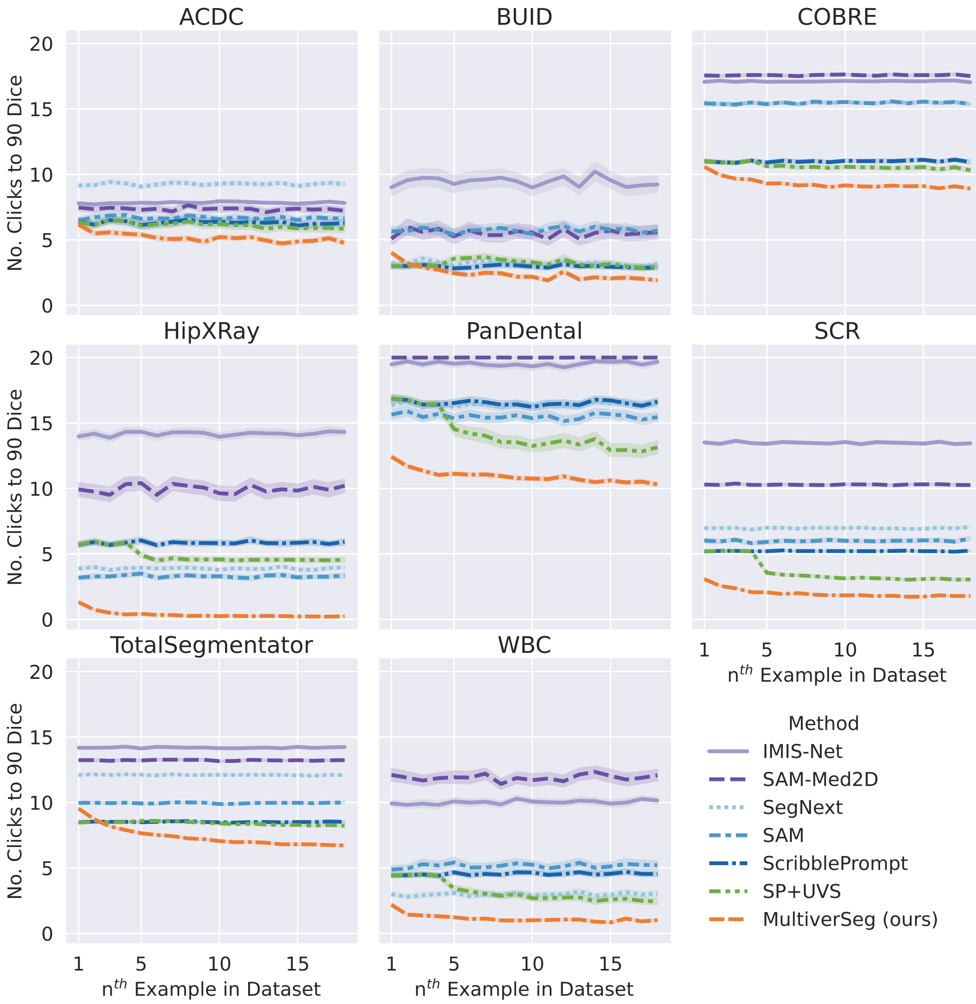


Figure 4. **Clicks to target Dice on unseen datasets**. Number of interactions needed to reach 90% Dice as a function of the example number being segmented. For the n^{th} image being segmented, the context set has n examples. MultiverSeg requires substantially fewer interactions to achieve 90 Dice than the baselines, and as more images are segmented, the average number of interactions required decreases dramatically. Shaded regions show 95% CI from bootstrapping.

tiverSeg reduced the number of clicks required to segment each dataset by $(36.93 \pm 1.53)\%$ and the number of scribble steps required by $(36.93 \pm 1.53)\%$ compared to ScribblePrompt.

F. Experiment 2: Analysis

F.1. In-Context Segmentation

Results. Fig. 17 show results by dataset with different context set sizes.

F.2. Interactive Segmentation In Context

Results. Fig. 18 and Fig. 19 show results by dataset using center clicks and centerline scribbles, respectively.

F.3. Inference Runtime and Memory Usage

MultiverSeg's inference runtime scales linearly with the context set size (Tab. 7). However, even with a context set of 64 examples, the runtime is under 150ms. Prior work on interactive interfaces indicates < 500ms latency is sufficient for cognitive tasks [51]. Since the interactions are stored in masks, inference runtime (per prediction) is not affected by the number of user interaction inputs.

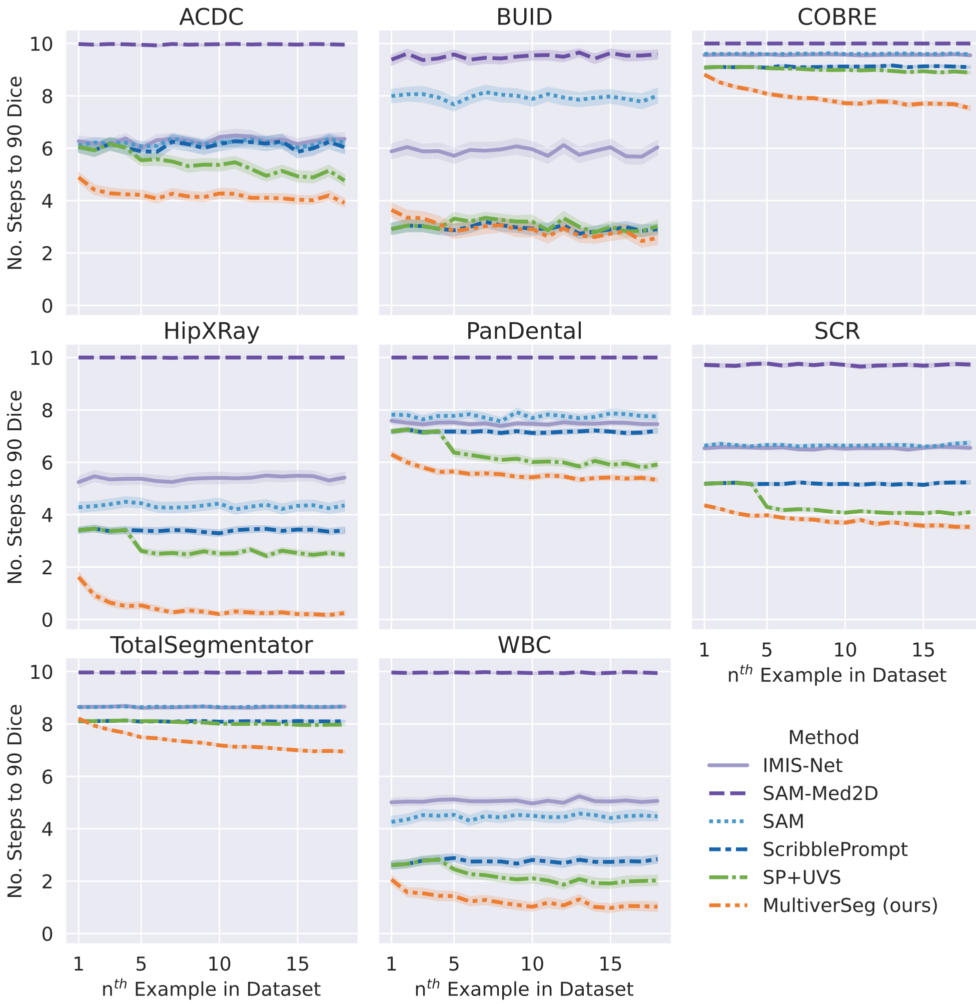


Figure 5. **Scribbles to target Dice on unseen datasets**. Number of interactions needed to reach 90% Dice as a function of the example number being segmented. For the n^{th} image being segmented, the context set has n examples. MultiverSeg requires substantially fewer interactions to achieve 90 Dice than the baselines, and as more images are segmented, the average number of interactions required decreases dramatically. Shaded regions show 95% CI from bootstrapping.

Context Size	Inference Time (ms)	GPU Memory
1	25.28 ± 0.16	28 MB
16	57.05 ± 0.20	1.89 GB
32	86.57 ± 0.06	3.64 GB
64	146.04 ± 0.16	7.15 GB
128	267.42 ± 0.24	12.16 GB
256	604.15 ± 0.36	24.17 GB

Table 7. **Inference runtime and GPU memory usage with different context set (CS) sizes**. We report mean $\pm$ standard deviation runtime in milliseconds across 1,000 predictions at 128^2 resolution with 1 click on an NVIDIA A100 GPU. GPU memory usage is reported as peak allocated memory during inference.

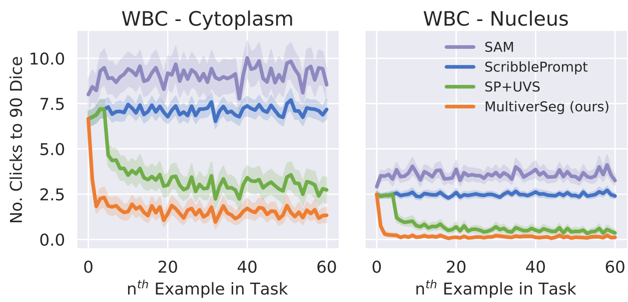


Figure 6. **Scribble steps to target Dice by task for WBC**. Number of interactions needed to reach a 90% Dice as a function of the example number being segmented. For the n^{th} image being segmented, the context set has n examples. Shading shows 95% CI from bootstrapping. WBC [89] is a microscopy dataset containing segmentation tasks for cytoplasm and nuclei of white blood cells. After segmenting a few images from the femur task with MultiverSeg, the rest of the images in the task can be segmented (to $\geq$ 90% Dice) with minimal (or no) additional interactions.

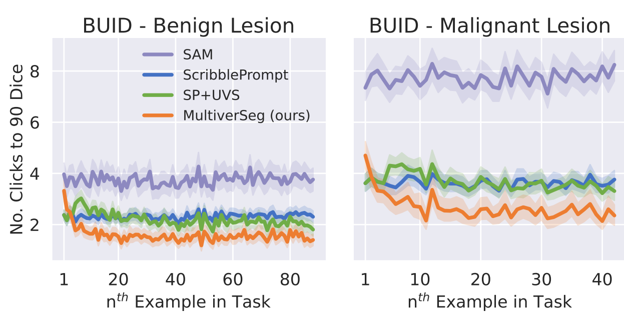


Figure 7. **Scribble steps to target Dice by task for BUID**. Number of interactions needed to reach a 90% Dice as a function of the example number being segmented. For the n^{th} image being segmented, the context set has n examples. Shading shows 95% CI from bootstrapping. BUID [3] is a breast ultrasound dataset containing segmentation tasks for benign and malignant lesions. As the context set of completed segmentations grows, the number of interactions required to segment each additional image with MultiverSeg gradually declines.

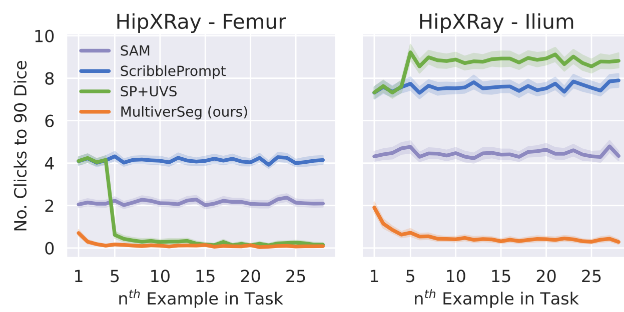


Figure 8. **Center clicks to target Dice by task for HipXRay**. Number of interactions needed to reach 90% Dice as a function of the example number being segmented. For the n^{th} image being segmented, the context set has n examples. Shading shows 95% CI from bootstrapping. HipXRay [29] is an X-Ray dataset with segmentation tasks for the femur and ilium bones. After segmenting a few images from the femur task with MultiverSeg, the rest of the images in the task can be segmented (to $\geq$ 90% Dice) with minimal additional interactions.

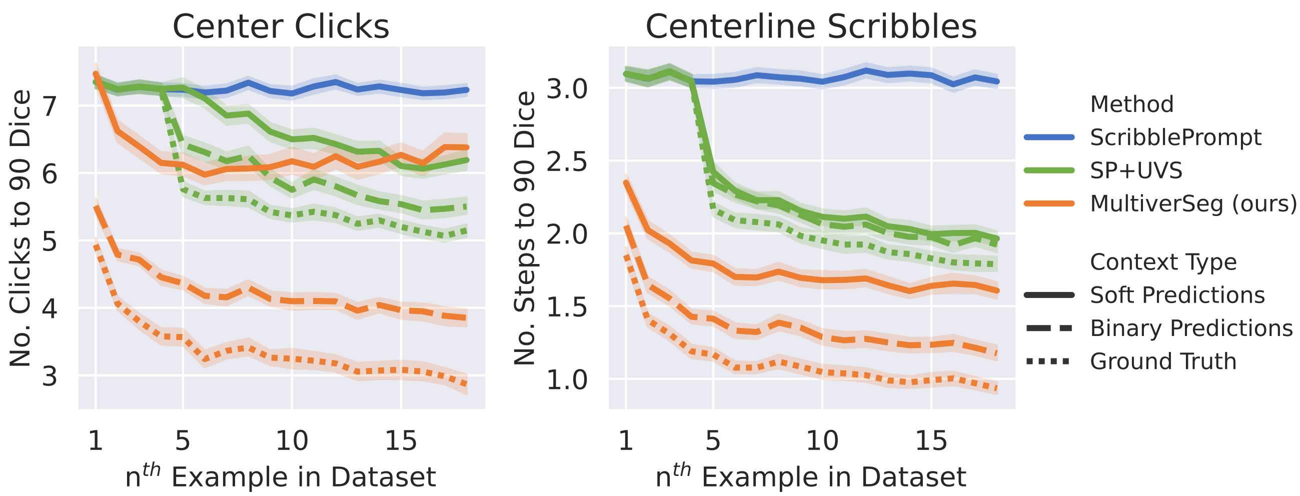


Figure 9. **Interactions to target dice on unseen datasets with different types of context sets**. Number of interactions needed to reach a 90% Dice as a function of the example number being segmented. For the n^{th} image being segmented, the context set has n examples. We show results with and without thresholding the predictions ("Binary Predictions" vs. "Soft Predictions") . We expect the number of interactions with "Ground Truth" context to be a lower bound on the number of interactions to reach 90% Dice. We show results averaged across validation tasks.

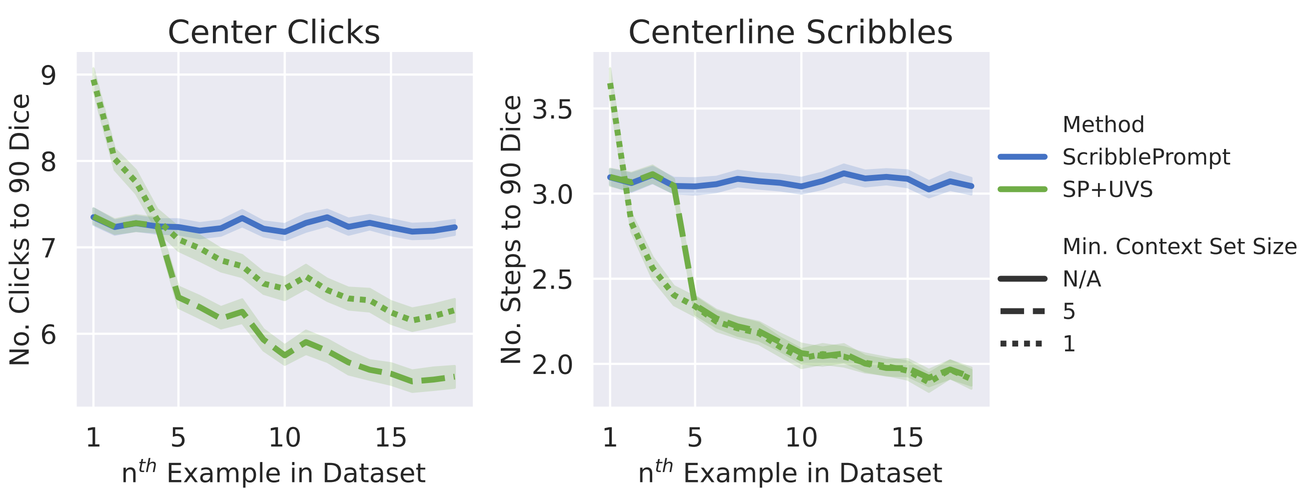
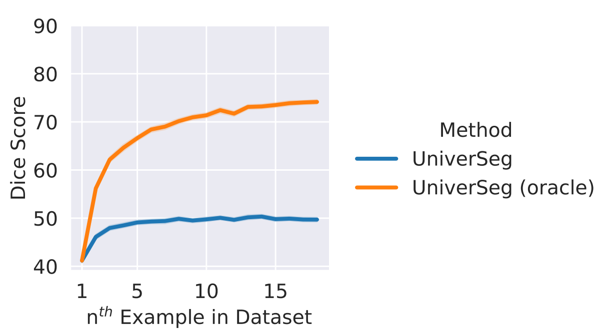


Figure 10. **Variations of SP+UVS**. Number of interactions needed to reach a 90% Dice as a function of the example number being segmented. For the n^{th} image being segmented, the context set has n examples. We show results for SP+UVS with different minimum context set size cutoffs, along with ScribblePrompt for reference. SP+UVS with a minimum context set size of k, means that when the context set has fewer than k examples, we perform interactive segmentation with ScribblePrompt (ignoring the context examples). When the context set is larger than the minimum size, we first make an in-context segmentation prediction using UniverSeg and then correct that prediction with ScribblePrompt. For small context set sizes, UniverSeg does not make accurate predictions, and initializing ScribblePrompt with UniverSeg's prediction increases the number of interactions required to reach 90% Dice. We show results averaged across validation tasks.

Interaction Protocol	Method	Dice Score ↑	HD95 ↓	Total Steps ↓
Center Clicks	SAM-Med2D	85.88 ± 0.14	3.76 ± 0.22	215.58 ± 2.22
	IMIS-Net	81.38 ± 0.30	13.05 ± 0.79	255.47 ± 2.53
	SAM	90.40 ± 0.06	1.40 ± 0.03	152.55 ± 1.76
	SegNext	90.50 ± 0.05	1.84 ± 0.06	158.16 ± 0.95
	ScribblePrompt	90.80 ± 0.08	1.48 ± 0.04	137.10 ± 1.21
	SP+UVS	90.70 ± 0.09	1.49 ± 0.06	122.01 ± 1.93
	MultiverSeg (ours)	$\mathbf{91.40 \pm 0.14}$	$\mathbf{1.26 \pm 0.11}$	$\mathbf{87.18 \pm 1.92}$
Centerline Scribbles	SAM-Med2D	29.58 ± 3.92	26.42 ± 3.36	178.00 ± 1.19
	IMIS-Net	80.93 ± 0.40	3.43 ± 0.32	123.46 ± 2.85
	SAM	80.19 ± 0.74	19.79 ± 1.78	125.14 ± 2.56
	ScribblePrompt	88.19 ± 0.24	$\mathbf{1.44 \pm 0.06}$	100.70 ± 2.67
	SP+UVS	88.57 ± 0.23	1.44 ± 0.07	92.50 ± 1.95
	MultiverSeg (ours)	$\mathbf{88.65 \pm 0.22}$	1.49 ± 0.13	$\mathbf{75.23 \pm 1.50}$

Figure 11. **Average segmentation quality and total interactions per unseen task**. We measure average segmentation quality across a set of 18 test images using Dice score and 95th percentile Hausdorff distance (HD95). For each metric, we show mean and standard deviation from bootstrapping. Dice and HD95 are similar across methods because we simulate interactions until the predicted segmentation has $\geq 90\%$ Dice or the maximum number of interaction steps is reached. MultiverSeg requires the fewest interaction steps per task on average. We report results on images at 256^2 resolution from 200 simulations.



(a) **Dice score by example number**. We show average Dice Score across unseen test data by example number. We exclude the initial seed example, such that for the n^{th} image being segmented, the context set has n examples.

Method	Dice Score ↑	No. Failures ↓
UniverSeg	48.89 ± 1.87	16.76 ± 0.40
UniverSeg (oracle)	68.15 ± 1.00	13.58 ± 0.24

(b) **Average performance on unseen tasks**. We report average Dice score per task of 18 images and the average number of examples where the Dice score was less than 90%. We report standard deviation across 200 simulations.

Figure 12. **Bootstrapping UniverSeg**. We use UniverSeg to sequentially segment images starting from a single example with a ground truth segmentation. After segmenting each image, the image and predicted segmentation are added to the context set for the next example. For the "oracle" version, we use ground truth labels in the context set instead of previously predicted segmentations. Even when using ground truth labels in the context set, which we expect to be an upper bound on performance, it was not possible to achieve 90% Dice for most images.

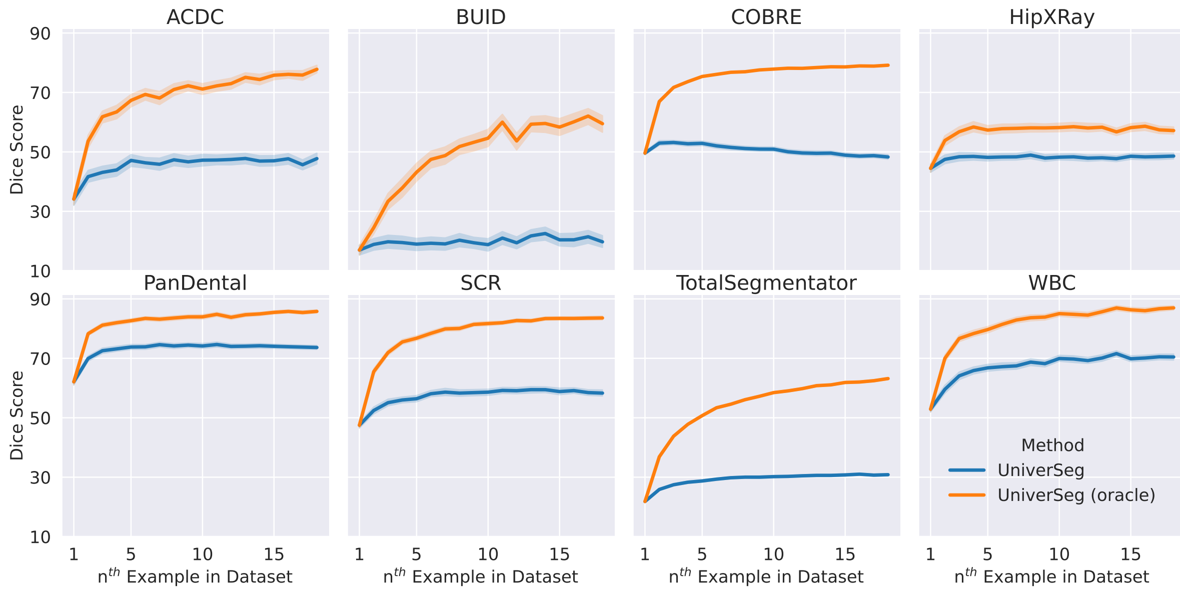


Figure 13. **Bootstrapping UniverSeg results by dataset**. We show Dice score vs. example number for unseen tasks averaged by dataset. After segmenting each image, the image and predicted segmentation are added to the context set for the next example. For the "oracle" version, we use ground truth labels in the context set instead of previously predicted segmentations. We exclude the initial seed example, such that for the n^{th} image being segmented, the context set has n examples. Shaded regions show 95% CI from bootstrapping.

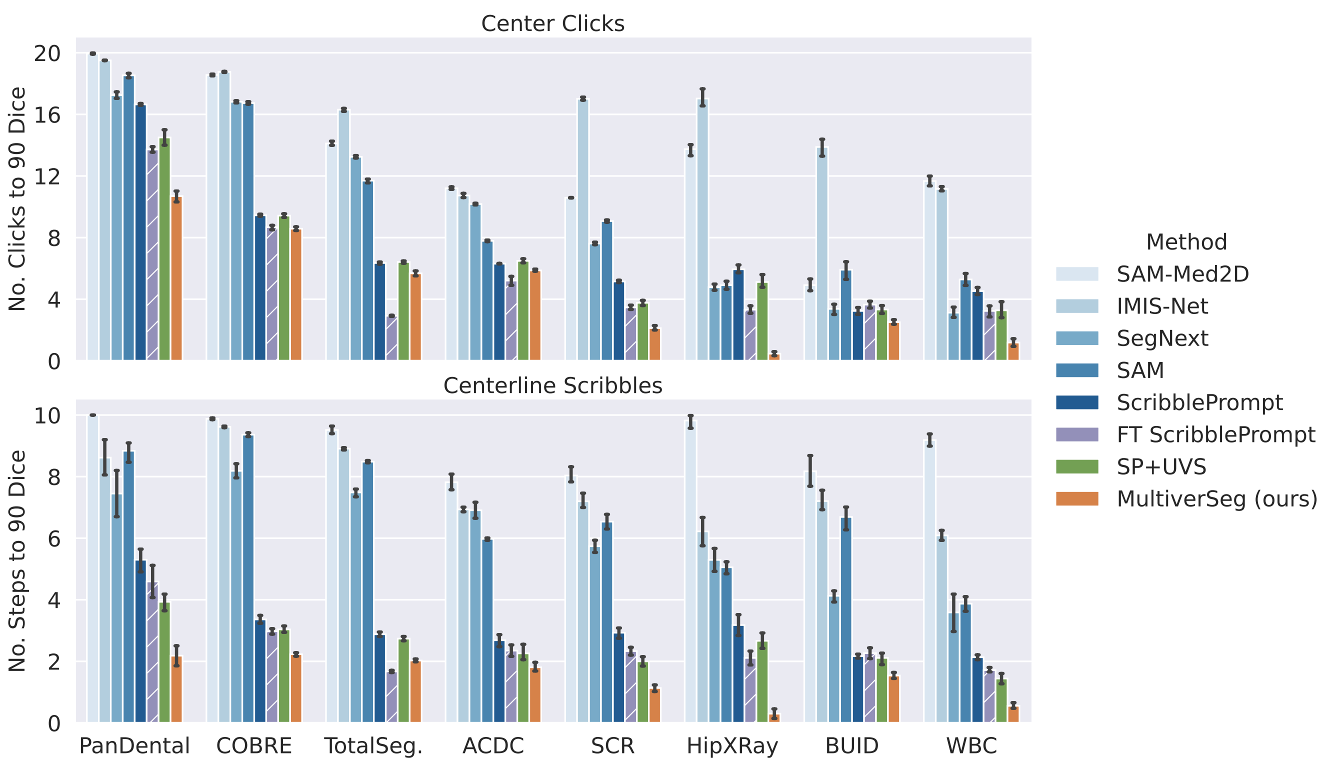


Figure 14. **MultiverSeg outperforms task-specific fine-tuning on most datasets**. We show average number of clicks and scribble steps per image to segment 18 images to $\geq 90\%$ Dice for each method. For *FT ScribblePrompt* (shaded), we used ScribblePrompt to interactively segment 5 images and then used those examples to fine-tune ScribblePrompt before interactively segmenting the rest. MultiverSeg required fewer interactions thant fine-tuned ScribblePrompt in 13 out of 14 scenarios. Error bars show 95% CI accross 5 random seeds.

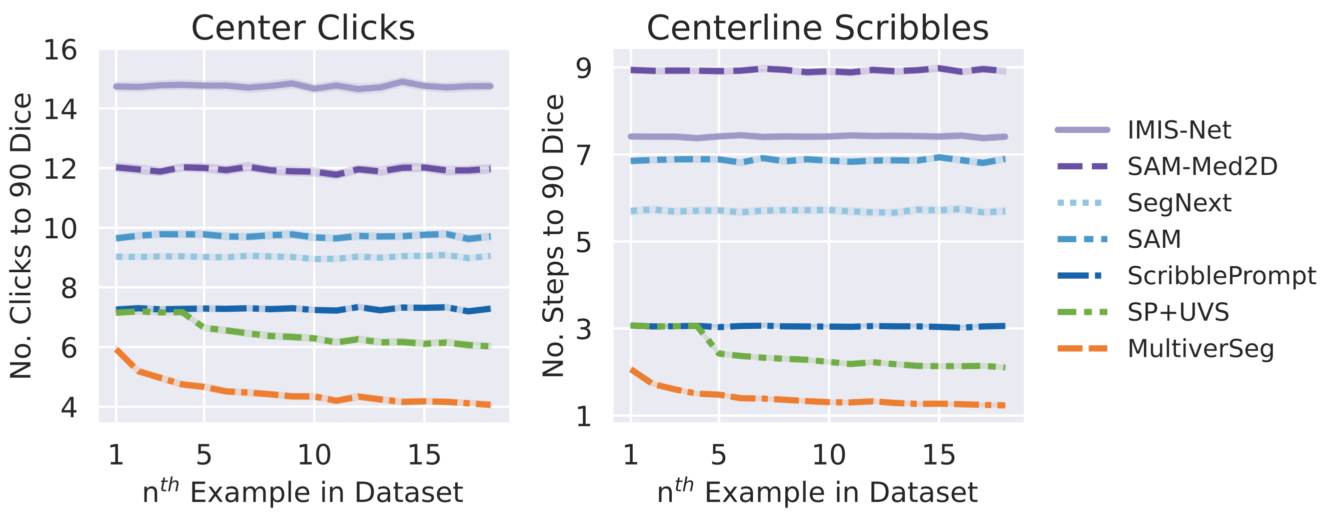


Figure 15. **Interactions to target Dice on unseen tasks at 128^2 resolution**. Number of interactions needed to reach a 90% Dice as a function of the example number being segmented. For the n^{th} image being segmented, the context set has n examples. MultiverSeg requires substantially fewer number of interactions to achieve 90% Dice than the baselines, and as more images are segmented, the average number of interactions required decreases dramatically. Shaded regions show 95% CI accross 200 random seeds.

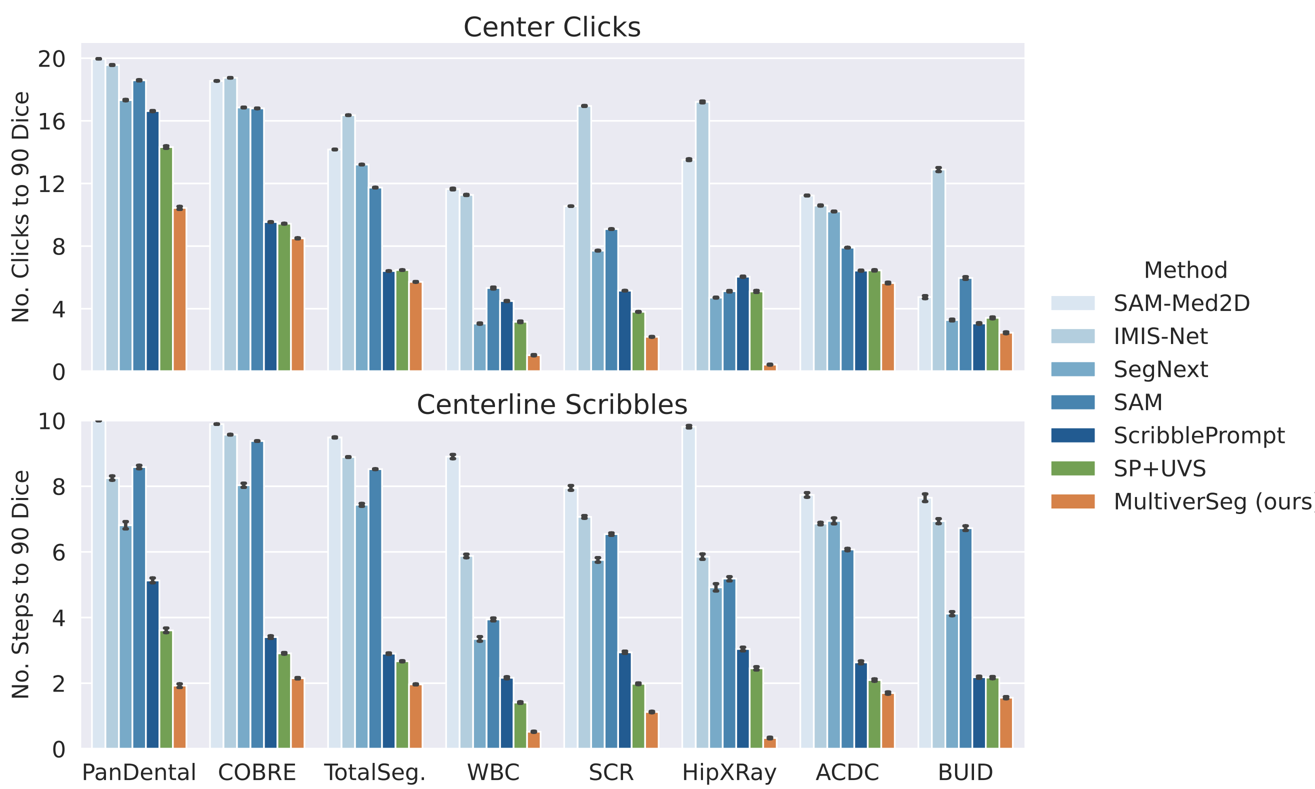


Figure 16. **Interactions per image by unseen dataset at 128^2 resolution**. We show average number of clicks and scribble steps per image to segment 18 images to $\geq$ 90% Dice for each method. In all scenarios, MultiverSeg required fewer or the same number of interactions than the best baseline. Error bars show 95% CI accross 200 random seed.

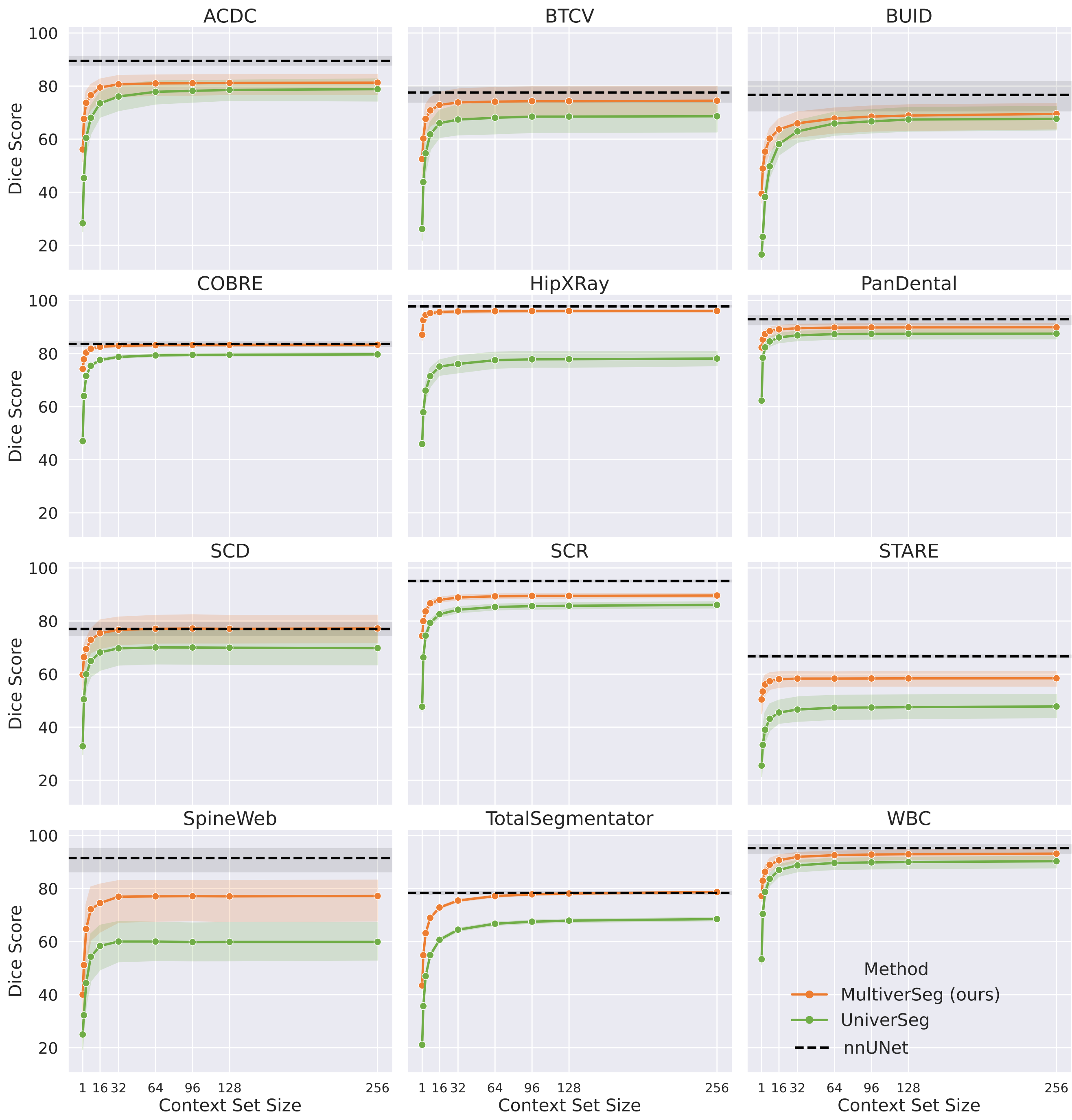


Figure 17. **In-context segmentation performance across context set sizes on unseen datasets**. We compare MultiverSeg to UniverSeg, an in-context segmentation method, given ground truth context labels. Points show results for context set sizes 1, 2, 4, 8, 16, 32, 64, 96, 128 and 256. Shading shows 95% CI from bootstrapping.

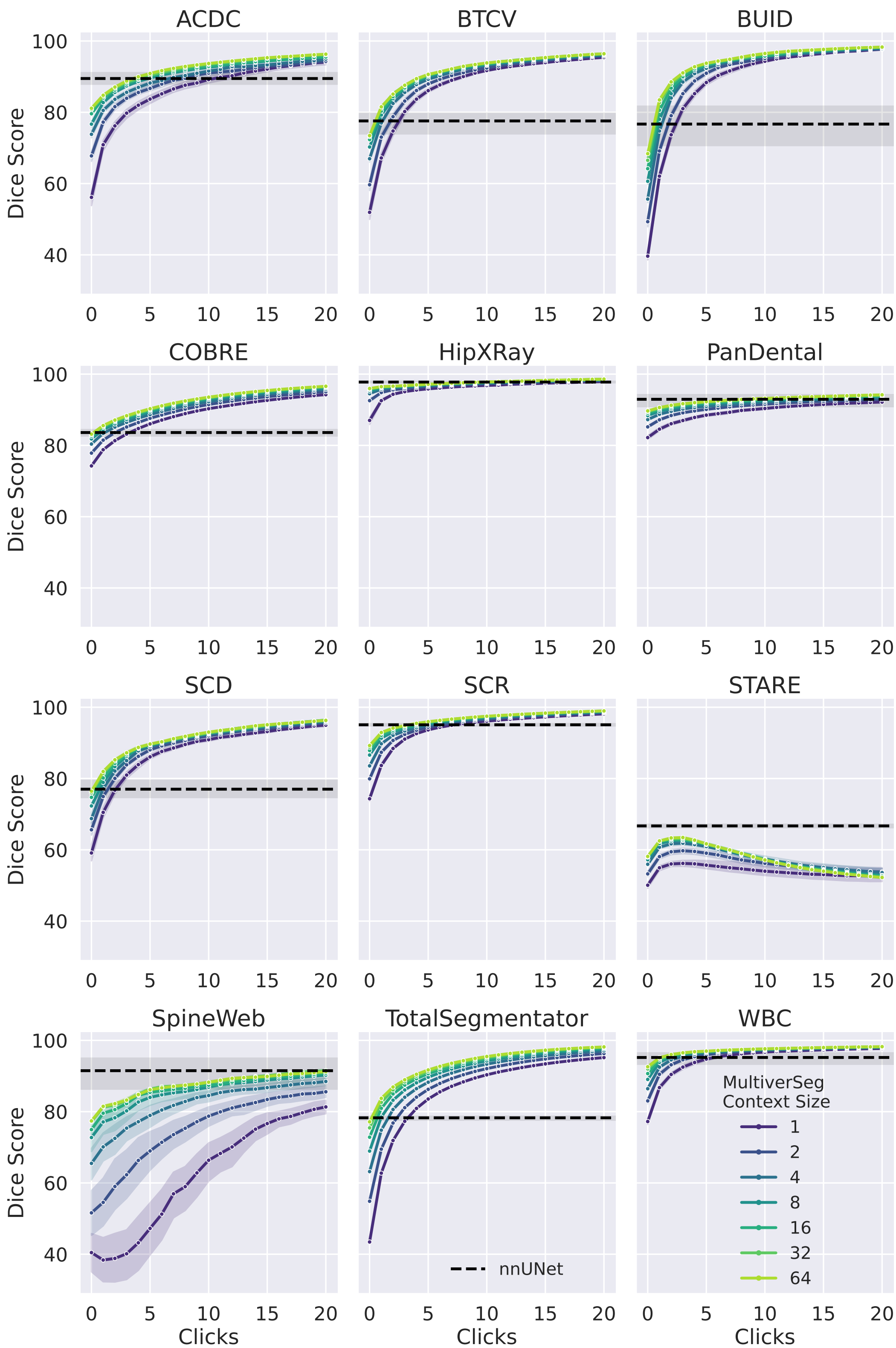


Figure 18. **Interactive segmentation in context with center clicks on unseen datasets**. MultiverSeg's interactive segmentation performance with the same number of interactions improves as the context set size grows. We first make an initial prediction based on the context set (step 0), and then simulate corrections with one center click at a time. Shading shows 95% CI from bootstrapping.

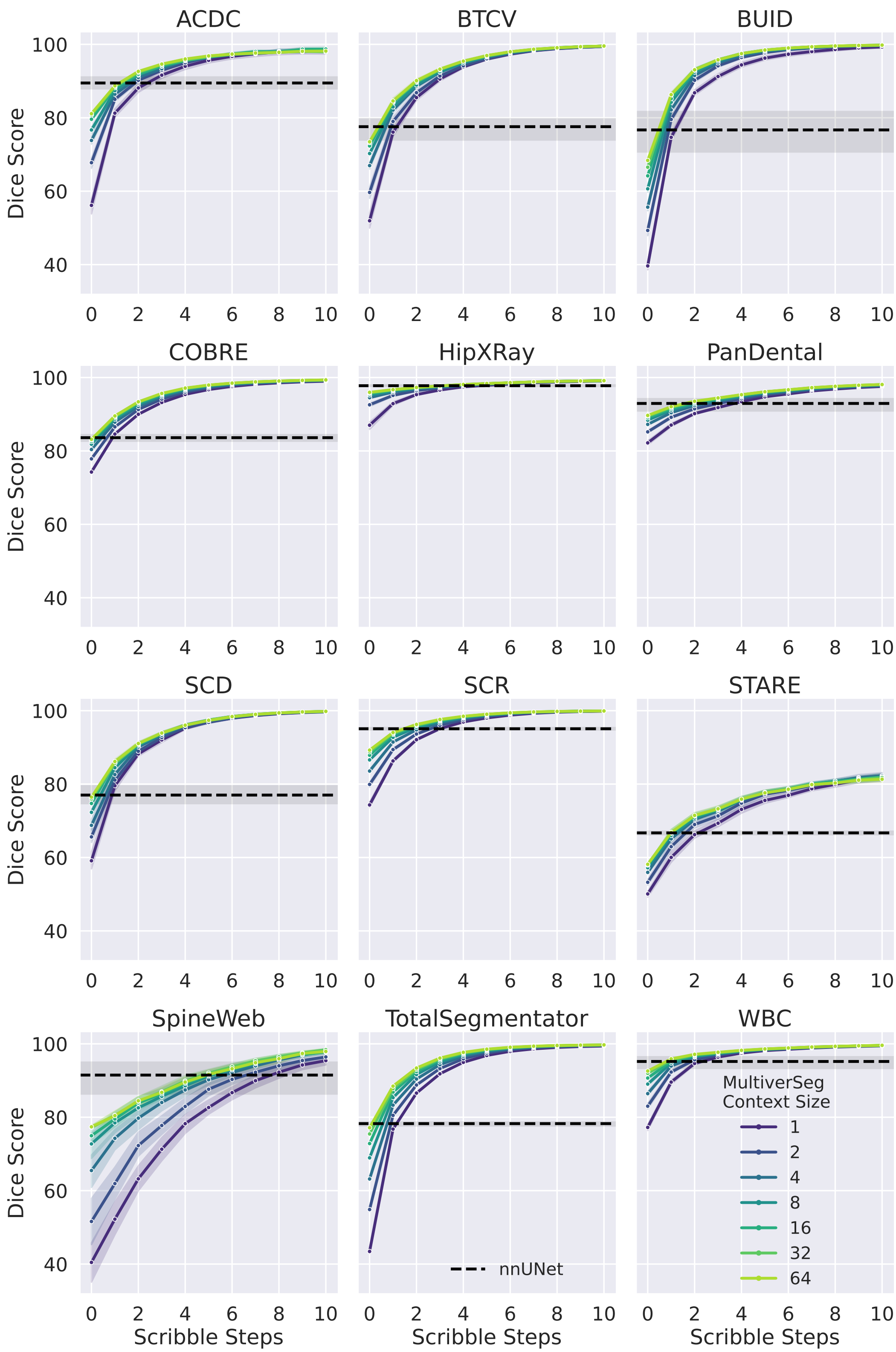


Figure 19. **Interactive segmentation in context with centerline scribbles on unseen datasets**. MultiverSeg's interactive segmentation performance with the same number of interactions improves as the context set size grows. We first make an initial prediction based on the context set (step 0), and then simulate centerline scribble corrections. Shading shows 95% CI from bootstrapping.

References

[1] Amir Hossein Abdi, Shohreh Kasaei, and Mojdeh Mehdizadeh. Automatic segmentation of mandible in panoramic x-ray. *Journal of Medical Imaging*, 2(4):044003, 2015. 3, 5

[2] C. J. Aine, H. J. Bockholt, J. R. Bustillo, J. M. Cañive, A. Caprihan, C. Gasparovic, F. M. Hanlon, J. M. Houck, R. E. Jung, J. Lauriello, J. Liu, A. R. Mayer, N. I. Perrone-Bizzozero, S. Posse, J. M. Stephen, J. A. Turner, V. P. Clark, and Vince D. Calhoun. Multimodal Neuroimaging in Schizophrenia: Description and Dissemination. *Neuroinformatics*, 15(4):343–364, 2017. 2, 3

[3] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. 3, 5, 11

[4] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 4

[5] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1): 1–13, 2017. 4

[6] Sophia Bano, Francisco Vasconcelos, Luke M Shepherd, Emmanuel Vander Poorten, Tom Vercauteren, Sebastien Ourselin, Anna L David, Jan Deprest, and Danail Stoyanov. Deep placental vessel segmentation for fetoscopic mosaicking. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 763–773. Springer, 2020. 4

[7] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37 (11):2514–2525, 2018. 2, 3, 5

[8] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019. 4

[9] Nicholas Bloch, Anant Madabhushi, Henkjan Huisman, John Freymann, Justin Kirby, Michael Grauer, Andinet Enquobahrie, Carl Jaffe, Larry Clarke, and Keyvan Farahani. Nci-isbi 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370(6):5, 2015. 4

[10] Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in biology and medicine*, 109:218–225, 2019. 4

[11] Victor Ion Butoi*, Jose Javier Gonzalez Ortiz*, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation. In *ICCV*, 2023. 1, 2, 6, 7

[12] Juan C. Caicedo, Allen Goodman, Kyle W. Karhohs, Beth A. Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, Mohammad Rohban, Shantanu Singh, and Anne E. Carpenter. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature Methods*, 16(12):1247–1253, 2019. 4

[13] Albert Cardona, Stephan Saalfeld, Stephan Preibisch, Benjamin Schmid, Anchi Cheng, Jim Pulokas, Pavel Tomancak, and Volker Hartenstein. An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. *PLoS biology*, 8(10):e1000502, 2010. 4

[14] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, Hui Sun, Junjun He, Shaoting Zhang, Min Zhu, and Yu Qiao. SAM-Med2D, 2023. arXiv:2308.16184 [cs]. 7

[15] Noel C. F. Codella, David A. Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC). *CoRR*, abs/1710.05006, 2017. 4

[16] Adrian V Dalca, John Guttag, and Mert R Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9290–9299, 2018. 4

[17] Adrian V Dalca, John Guttag, and Mert R Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9290–9299, 2018. 3

[18] Pasquale De Marinis, Nicola Fanelli, Raffaele Scaringi, Emanuele Colonna, Giuseppe Fiameni, Gennaro Vessio, and Giovanna Castellano. Label anything: Multi-class few-shot semantic segmentation with visual prompts. *arXiv preprint arXiv:2407.02075*, 2024. 2, 7

[19] Etienne Decenciere, Guy Cazuguel, Xiwei Zhang, Guillaume Thibault, J-C Klein, Fernand Meyer, Beatriz Marcotegui, Gwénolé Quellec, Mathieu Lamard, Ronan Danno, et al. Teleophta: Machine learning and image processing methods for teleophthalmology. *Irbm*, 34(2):196–203, 2013. 4

[20] Aysen Degerli, Morteza Zabihi, Serkan Kiranyaz, Tahir Hamid, Rashid Mazhar, Ridha Hamila, and Moncef Gabbouj. Early detection of myocardial infarction in low-quality echocardiography. *IEEE Access*, 9:34442–34453, 2021. 4

[21] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 2

[22] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012. 3
[23] J Gamper, NA Koohbanani, K Benes, S Graham, M Jahanifar, SA Khurram, A Azam, K Hewitt, and N Rajpoot. Pannuke dataset extension, insights and baselines. arxiv. 2020 doi: 10.48550. *ARXIV*, 2003. 4
[24] Stephan Gerhard, Jan Funke, Julien Martel, Albert Cardona, and Richard Fetter. Segmented anisotropic ssTEM dataset of neural tissue. 2013. 4
[25] Randy L Gollub, Jody M Shoemaker, Margaret D King, Tonya White, Stefan Ehrlich, Scott R Sponheim, Vincent P Clark, Jessica A Turner, Bryon A Mueller, Vince Magnotta, et al. The mcic collection: a shared repository of multimodal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics*, 11:367–388, 2013. 4
[26] Ioannis S Gousias, Daniel Rueckert, Rolf A Heckemann, Leigh E Dyet, James P Boardman, A David Edwards, and Alexander Hammers. Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest. *Neuroimage*, 40 (2):672–684, 2008. 4
[27] Ioannis S Gousias, A David Edwards, Mary A Rutherford, Serena J Counsell, Jo V Hajnal, Daniel Rueckert, and Alexander Hammers. Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants. *Neuroimage*, 62(3):1499–1509, 2012. 4
[28] Endre Grøvik, Darvin Yi, Michael Iv, Elizabeth Tong, Daniel Rubin, and Greg Zaharchuk. Deep learning enables automatic detection and segmentation of brain metastases on multisequence mri. *Journal of Magnetic Resonance Imaging*, 51(1):175–182, 2020. 4
[29] Daniel Gut. X-ray images of the hip joints. 1, 2021. Publisher: Mendeley Data. 3, 5, 11
[30] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, page 101821, 2020. 4
[31] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1): 762, 2022. 4
[32] Andrew Hoopes, Malte Hoffmann, Douglas N. Greve, Bruce Fischl, John Guttag, and Adrian V. Dalca. Learning the effect of registration hyperparameters with hypermorph. *Machine Learning for Biomedical Imaging*, 1:1–30, 2022. 4
[33] AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3):203–210, 2000. 4
[34] Humans in the Loop. Teeth segmentation dataset. 4
[35] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022. 4
[36] Rashed Karim, R James Housden, Mayuragoban Balasubramaniam, Zhong Chen, Daniel Perry, Ayesha Uddin, Yosra Al-Beyatti, Ebrahim Palkhi, Prince Acheampong, Samantha Obom, et al. Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge. *Journal of Cardiovascular Magnetic Resonance*, 15(1):1–17, 2013. 4
[37] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data, 2019. 4
[38] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 4
[39] Serkan Kiranyaz, Aysen Degerli, Tahir Hamid, Rashid Mazhar, Rayyan El Fadil Ahmed, Rayaan Abouhasera, Morteza Zabihi, Junaid Malik, Ridha Hamila, and Moncef Gabbouj. Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection. *IEEE Access*, 8:210301–210317, 2020. 4
[40] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023. 2, 7
[41] Markus Krönke, Christine Eilers, Desislava Dimova, Melanie Köhler, Gabriel Buschner, Lilit Schweiger, Lemonia Konstantinidou, Marcus Makowski, James Nagarajah, Nassir Navab, et al. Tracked 3d ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. *Plos one*, 17(7):e0268550, 2022. 4
[42] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11): 2556–2568, 2019. 4
[43] Maria Kuklisova-Murgasova, Paul Aljabar, Latha Srinivasan, Serena J Counsell, Valentina Doria, Ahmed Serag, Ioannis S Gousias, James P Boardman, Mary A Rutherford, A David Edwards, et al. A dynamic 4d probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763, 2011. 4
[44] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image*

Processing Theory, Tools and Applications (IPTA), pages 1–6. IEEE, 2020. 4

[45] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault Workshop Challenge*, page 12, 2015. 2, 3, 4

[46] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019. 4

[47] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015. 4

[48] Mingchao Li, Yuhan Zhang, Zexuan Ji, Keren Xie, Songtao Yuan, Qinghuai Liu, and Qiang Chen. Ipn-v2 and octa-500: Methodology and dataset for retinal image segmentation. *arXiv preprint arXiv:2012.07261*, 2020. 4

[49] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. 4

[50] Qin Liu, Jaemin Cho, Mohit Bansal, and Marc Niethammer. Rethinking interactive image segmentation with low latency high quality and diverse prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3782, 2024. 7

[51] Zhicheng Liu and Jeffrey Heer. The effects of interactive latency on exploratory visual analysis. *IEEE transactions on visualization and computer graphics*, 20(12):2122–2131, 2014. 9

[52] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012. 4

[53] Maximilian T Löffler, Anjany Sekuboyina, Alina Jacob, Anna-Lena Grau, Andreas Scharr, Malek El Husseini, Mareike Kallweit, Claus Zimmer, Thomas Baum, and Jan S Kirschke. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2(4):e190138, 2020. 4

[54] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Tao Song, Xiaofan Zhang, Kang Li, Guotai Wang, and Shaoting Zhang. Word: Revisiting organs segmentation in the whole abdominal region. *arXiv preprint arXiv:2111.02403*, 2021. 4

[55] Jun Ma, Yao Zhang, Song Gu, Xingle An, Zhihe Wang, Cheng Ge, Congcong Wang, Fan Zhang, Yu Wang, Yinan Xu, et al. Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. *Medical Image Analysis*, 82: 102616, 2022. 4

[56] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:1–9, 2024. 7

[57] Yuhui Ma, Huaying Hao, Jianyang Xie, Huazhu Fu, Jiong Zhang, Jianlong Yang, Zhen Wang, Jiang Liu, Yalin Zheng, and Yitian Zhao. Rose: a retinal oct-angiography vessel segmentation dataset and new model. *IEEE Transactions on Medical Imaging*, 40(3):928–939, 2021. 4

[58] Jacob A. Macdonald, Zhe Zhu, Brandon Konkel, Maciej Mazurowski, Walter Wiggins, and Mustafa Bashir. Duke liver dataset (MRI) v2, 2023. 4

[59] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007. 4

[60] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011. 4

[61] Francesco Marzola, Nens Van Alfen, Jonne Doorduin, and Kristen M. Meiburger. Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment. *Computers in Biology and Medicine*, 135:104623, 2021. 4

[62] Maciej A Mazurowski, Kal Clark, Nicholas M Czarnek, Parisa Shamsesfandabadi, Katherine B Peters, and Ashirbani Saha. Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with the cancer genome atlas data. *Journal of neuro-oncology*, 133:27–35, 2017. 4

[63] Bjoern Menze, Leo Joskowicz, Spyridon Bakas, Andras Jakab, Ender Konukoglu, Anton Becker, Amber Simpson, and Richard D. Quantification of uncertainties in biomedical image quantification 2021. *4th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2021)*, 2021. 4

[64] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 4

[65] Anna Montoya, Hasnin, kaggle446, shirzad, Will Cukierski, and yffud. Ultrasound nerve segmentation, 2016. 4

[66] Kelly Payette, Priscille de Dumast, Hamza Kebiri, Ivan Ezhov, Johannes C Paetzold, Suprosanna Shit, Asim Iqbal, Romesa Khan, Raimund Kottke, Patrice Grehten, et al. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. *Scientific Data*, 8(1):1–14, 2021. 4

[67] Lina Pedraza, Carlos Vargas, Fabián Narváez, Oscar Durán, Emma Muñoz, and Eduardo Romero. An open access thyroid ultrasound image database. In *10th international symposium on medical information processing and analysis*, page 92870W. SPIE / International Society for Optics and Photonics, 2015. 4

[68] Gašper Podobnik, Primož Strojan, Primož Peterlin, Bulat Ibragimov, and Tomaž Vrtovec. HaN-Seg: The head and neck organ-at-risk CT and MR segmentation dataset. *Medical Physics*, 50(3):1917–1927, 2023. tex.eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.16197. 4

[69] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid), 2018. 4

[70] Perry Radau, Yingli Lu, Kim Connelly, Gideon Paul, AJWG Dick, and Graham Wright. Evaluation framework for algorithms segmenting short axis cardiac mri. *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, 49, 2009. 2, 3

[71] Marianne Rakic, Hallee E. Wong, Jose Javier Gonzalez Ortiz, Beth Cimini, John V. Guttag, and Adrian V. Dalca. Tyche: Stochastic in-context learning for medical image segmentation. *Computer Vision and Pattern Reconition (CVPR)*, 2024. 1, 2

[72] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L. Rubin. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020. 4

[73] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, SQ Truong, CD Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, AY Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *MedRxiv*, 2021. 4

[74] Constantin Seibold, Simon Reiß, Saquib Sarfraz, Matthias A. Fink, Victoria Mayer, Jan Sellner, Moon Sung Kim, Klaus H. Maier-Hein, Jens Kleesiek, and Rainer Stiefelhagen. Detailed annotations of chest x-rays via ct projection for report understanding. In *Proceedings of the 33th British Machine Vision Conference (BMVC)*, 2022. 4

[75] Ahmed Serag, Paul Aljabar, Gareth Ball, Serena J Counsell, James P Boardman, Mary A Rutherford, A David Edwards, Joseph V Hajnal, and Daniel Rueckert. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *Neuroimage*, 59(3):2255–2265, 2012. 4

[76] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017. 4

[77] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 4

[78] Yuxin Song, Jing Zheng, Long Lei, Zhipeng Ni, Baoliang Zhao, and Ying Hu. CT2US: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. *Ultrasonics*, 122:106706, 2022. 4

[79] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004. 3

[80] Bram van Ginneken, Mikkel B. Stegmann, and Marco Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical Image Analysis*, 10(1):19–40, 2006. 3, 5

[81] Santiago Vitale, José Ignacio Orlando, Emmanuel Iarussi, and Ignacio Larrabide. Improving realism in patient-specific abdominal ultrasound simulation using cyclegans. *International journal of computer assisted radiology and surgery*, 15(2):183–192, 2020. 4

[82] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023. 2, 3, 5

[83] Hallee E. Wong, Marianne Rakic, John Guttag, and Adrian V. Dalca. Scribbleprompt: Fast and flexible interactive segmentation for any medical image. *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 7

[84] Junde Wu and Min Xu. One-prompt to segment all medical images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11302–11312, 2024. 2, 7

[85] Yingtao Zhang, Min Xian, Heng-Da Cheng, Bryar Shareef, Jianrui Ding, Fei Xu, Kuan Huang, Boyu Zhang, Chunping Ning, and Ying Wang. Busis: A benchmark for breast ultrasound image segmentation. In *Healthcare*, page 729. MDPI, 2022. 4

[86] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V. Guttag, and Adrian V. Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[87] Qi Zhao, Shuchang Lyu, Wenpei Bai, Linghan Cai, Binghao Liu, Meijing Wu, Xiubo Sang, Min Yang, and Lijiang Chen. A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. *CoRR*, abs/2207.06799, 2022. 4

[88] Guoyan Zheng, Chengwen Chu, Daniel L Belavỳ, Bulat Ibragimov, Robert Korez, Tomaž Vrtovec, Hugo Hutt, Richard Everson, Judith Meakin, Isabel Lǒpez Andrade, et al. Evaluation and comparison of 3d intervertebral disc localization and segmentation methods for 3d t2 mr data: A grand challenge. *Medical image analysis*, 35:327–344, 2017. 2, 3

[89] Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018. 3, 5, 7, 11