arXiv:2404.16325v1 [cs.CV] 25 Apr 2024

Semantic Segmentation Refiner for Ultrasound Applications with Zero-Shot Foundation Models

Hedda Cohen Indelman[1,†]
AI/ML Research, GE Healthcare

Elay Dahan[1,†]
AI/ML Research, GE Healthcare

Angeles M. Perez-Agosto[2]
Clinical Insights & Development, GE Healthcare

Carmit Shiran[3]
Clinical Insights & Development, GE Healthcare

Doron Shaked[1]
AI/ML Research, GE Healthcare

Nati Daniel[1,*]
AI/ML Research, GE Healthcare

Abstract

Despite the remarkable success of deep learning in medical imaging analysis, medical image segmentation remains challenging due to the scarcity of high-quality labeled images for supervision. Further, the significant domain gap between natural and medical images in general and ultrasound images in particular hinders fine-tuning models trained on natural images to the task at hand. In this work, we address the performance degradation of segmentation models in low-data regimes and propose a prompt-less segmentation method harnessing the ability of segmentation foundation models to segment abstract shapes. We do that via our novel prompt point generation algorithm which uses coarse semantic segmentation masks as input and a zero-shot prompt-able foundation model as an optimization target. We demonstrate our method on a segmentation findings task (pathologic anomalies) in ultrasound images. Our method's advantages are brought to light in varying degrees of low-data regime experiments on a small-scale musculoskeletal ultrasound images dataset, yielding a larger performance gain as the training set size decreases.

Keywords Zero-Shot Learning, Foundation Model, Semantic Segmentation, Prompt Engineering, Musculoskeletal Ultrasound, Pathology Finding.

1 Introduction

Ultrasound is a popular medical imaging modality used to image a large variety of organs and tissues. Ultrasound is often the preferred choice due to its non-radiative and non-invasive nature, relatively easy and fast imaging procedure, and lower costs. Automating the diagnosis or highlighting relevant areas in the image will contribute to faster workflows and potentially more consistent and accurate diagnoses.

Artificial Intelligence (AI) has demonstrated remarkable success in automatic medical imaging analysis. Compared to classical methods, previous work based on convolutional neural networks on various medical imaging tasks, such as classification and segmentation, have shown state-of-the-art results [1, 2, 3, 4]. However, effective deep learning segmentation algorithms for medical images is an especially challenging task due to the scarcity of high-quality labeled images for supervision. Moreover, in medical imaging it is often the case that identification of *findings* regions, namely regions of potentially pathological visual anomalies, having neither a clear boundary nor a typical geometry or position is much more challenging than the identification of an anatomy in its context. Findings are also typically rare, which brings to light the challenge of training such models in limited data regimes.

[*] Corresponding author, e-mail: nati.daniel@gehealthcare.com. †These authors have contributed equally to this work. [1] Dept. of AI/ML Research, GE Healthcare, Haifa, Israel. [2] Dept. of Clinical Applications, Point of Care Ultrasound & Handheld, Texas, USA. [3] Dept. of Clinical Applications, Point of Care Ultrasound & Handheld, Wisconsin, USA.

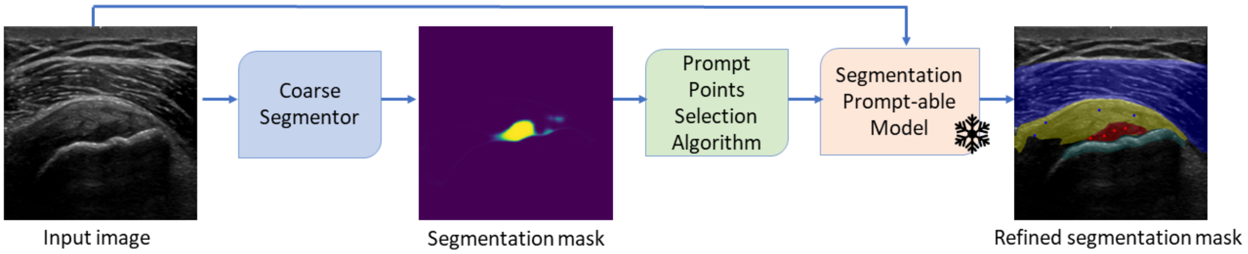

Input image
Coarse Segmentor
Segmentation mask
Prompt Points Selection Algorithm
Segmentation Prompt-able Model
Refined segmentation mask

Figure 1: A high-level illustration of our semantic segmentation refinement method with zero-shot foundation models. A pre-trained segmentation model predicts a semantic segmentation for each class of an input image. In this example, classes comprise anatomies and pathologies in an ultrasound image, and the coarse segmentor output depicts the predicted semantic segmentation of a pathology. A prompt selection model selects positive and negative points. Consequently, a zero-shot semantic segmentation mask of the pathology is predicted by a foundation segmentation model, prompted by the selected points for the input image. Positive prompt points are depicted in red, and negative prompt points are depicted in blue. The pathology semantic segmentation prediction is highlighted in red. For illustration purposes, the muscle is highlighted in purple, the tendon in yellow, and the bone in green. The freeze symbol indicates preventing gradients from being propagated to the model weights.

Recently, new segmentation models have emerged. Trained on data at huge scales, these foundation models aim to be more generic rather than tailored to specific datasets. The Segment Anything Model (SAM) [5] is a foundational model demonstrating zero-shot generalization in segmenting natural images using a prompt-driven approach. The SonoSAM [6] foundational model adapts SAM to ultrasound images by fine-tuning the prompt and mask decoder [6]. Although fine-tuning methods often improve the results on target datasets [7] they essentially downgrade the generalization capabilities of the foundation model. Further, a significant domain gap between natural and medical images, ultrasound images in particular[8], hinders fine-tuning models trained on natural images to the task at hand [7].

In this work, we address the performance degradation of segmentation models in low-data regimes and derive a novel method for harnessing segmentation foundation models' ability to segment arbitrary regions. Our semantic segmentation refinement method comprises two stages: First, a coarse segmentation is predicted by a model trained on a small subset of the training data. In the second stage, our novel points generation from a coarse pathology segmentation algorithm is used to prompt a segmentation foundation model. Positive prompt points are selected using a partition around medoids method as the most representative pathology points. Negative prompt points are selected by a prompt selection optimization algorithm that identify the context anatomy. Importantly, we do not fine-tune the foundation model to our dataset, i.e., it produces a zero-shot segmentation. The end-to-end pipeline is illustrated in Fig. 1.

The method's advantages are brought to light on varying degrees of low-data regimes experiments on a small-scale images dataset, yielding a larger performance gain compared to a state-of-the-art segmentation model [9] as the training set size decreases. Further, ablation studies validate the effectiveness of our semantic segmentation refinement model. Our approach applies to other ultrasound-based medical diagnostics tasks.

The paper is organized as follows: Section 2 presents the semantic segmentation task and leading approaches. Our method is presented in Section 3, and the experimental setup is presented in Section 4. Section 5 presents the results and ablation studies on a discontinuity in tendon fiber (DITF) pathology finding task in a musculoskeletal ultrasound (MSK) dataset, and the conclusions are presented in Section 6.

2 Related Work

2.1 Semantic Segmentation Models

Semantic segmentation aims to assign a label or a class to each pixel in an image. Unlike image classification, which assigns a single label to the entire image, semantic segmentation provides a more detailed understanding of the visual scene by segmenting it into distinct regions corresponding to objects or classes. This is an essential technique for applications, such as autonomous vehicles, medical image analysis, and scene understanding in robotics.

Like other computer vision tasks, deep learning has demonstrated state-of-the-art results in the semantic segmentation of medical images. The semantic segmentation problem can be formulated as follows: Given an image $\mathbf{I} \in \mathbf{R}^{C \times H \times W}$, our goal is to train a deep neural network to predict the pixel-wise probability map $\mathbf{S}^{N \times H \times W}$ of the classes in the dataset, where $\mathbf{N}$ is the number of classes in the dataset.

DeepLabV3 [9] represents a distinctive approach in semantic image segmentation. Utilizing dilated convolutions, the model strategically enlarges the receptive field and manages the balance between global and local features through padding rates. Notably, the spatial pyramid pooling module proposed by the authors aggregates features from dilated convolutions at various scales, enhancing contextual information. Distinctive from encoder-decoder architectures such as the U-Net [10], it is built upon a robust pre-trained encoder, contributing to its success in generating accurate and detailed segmentation masks across diverse applications.

Since DeepLabV3 remains a staple choice for a performant semantic segmentation model, we adopt it as our method's coarse segmentor.

2.2 Semantic Segmentation Foundation Models

Foundation models are trained on broad data at a huge scale and are adaptable to a wide range of downstream tasks [11, 12, 13]. The Segment Anything Model (SAM) [5] emerged as a versatile foundation model for natural image segmentation. Trained on a dataset of over 11 million images and 1B masks, it demonstrates impressive zero-shot generalization in segmenting natural images using an interactive and prompt-driven approach. Prompt types include foreground/background points, bounding boxes, masks, and text prompts. However, SAM achieves subpar generalization on medical images due to substantial domain gaps between natural and medical images [14, 15, 16, 17, 18]. Moreover, SAM obtains the poorest results on ultrasound compared to other medical imaging modalities [15]. These results are attributed to the ultrasound characteristics, e.g., the scan cone, poor image quality, and unique speckled texture. A common methodology to overcome this generalization difficulty is to fine-tune a foundation model on a target dataset [19]. An efficient fine-tuning strategy is Low-Rank Adaptation (LoRA) [20], which has been adopted in fine-tuning SAM to relatively small medical imaging datasets [21, 22, 23]. SonoSAM [6] demonstrates state-of-the-art generalization in segmenting ultrasound images. Fine-tuned on a rich and diverse set of ultrasound image-mask pairs, it has emerged as a prompt-able foundational model for ultrasound image segmentation.

Notably, adapting prompt-based models to medical image segmentation is difficult due to the conundrum of crafting high-quality prompts [15]. Manually selecting prompts is time-consuming and requires domain expertise. Methods of extracting prompts from ground-truth masks [23] cannot be applied during inference as they rely on full supervision. Auto-prompting techniques rely on the strong Vision Transformer (ViT-H) image encoder [24] semantic representation capabilities, and suggest generating a segmentation prompt based on SAM's image encoder embedding [18, 25]. Other strategies suggest replacing the mask decoder with a prediction head requiring no prompts [16]. Nevertheless, SAM's zero-shot prediction accuracy is typically lower than that of the segmentation models trained with fully supervised methods [26].

Motivated by the generalization abilities of segmentation foundation models, we devise a points selection algorithm from coarse segmentation masks that allows harnessing prompt-based models to ultrasound segmentation in a zero-shot setting.

3 Method

In this section, we present our method for refining a coarse pathology segmentation mask with zero-shot foundation models. This method can be adapted to natural images, as well as to the medical imaging domain. Herein, we validate it based on a specific challenging task of segmenting a discontinuity of the tendon fiber finding (Sec. 4.1), which is the main ultrasound finding of a tendon partial tear pathology.

Our key intuition is that although the performance of segmentation models decreases significantly in low-data regimes, even such coarse segmentation masks can be utilized for extracting high-quality prompts that harness segmentation foundation models' capabilities. Importantly, we use the publicly available pre-trained foundation models without further modification. The flexibility of our method allows for incorporating either SonoSAM or SAM. Though the above-mentioned foundation models allow several types of prompts, we focus on foreground (positive) and background (negative) prompt points.

Our method makes use of the ground-truth tendon segmentation, denoted T^{gt}. Since the tendon in the context of the DIFT pathology is usually easy to segment due to its typical geometry and position and relatively simple data acquisition and labeling, we assume that strong segmentation models exist for this task and that their output can be used in lieu of the ground-truth segmentation. With that, we introduce our two-stage method, summarized in Algorithm 1.

First, a segmentation model [9] is trained on a random subset of the training data. A coarse semantic segmentation is then predicted for a given test image. Then, k positive and k negative prompt points are selected to prompt a segmentation foundation model. We next describe our prompt points selection algorithm in greater detail.

Algorithm 1 The Semantic Segmentation Refiner Method

Input:

- Input image I
- Ground-truth tendon mask T^{gt}
- Frozen $SonoSAM$ model
- Pre-trained segmentation model S

Output:

- Refined pathology segmentation mask O

```
Coarse segmentation mask $\tilde{O} \leftarrow S(I)$
Positive points selection $pts^{pos} \leftarrow k$-medoids($\tilde{O}$)
Modified ground-truth tendon mask $T^{\tilde{gt}} \leftarrow T^{gt} \setminus \tilde{O}$
Initialize complementary problem
$p\bar{t}s^{neg} \leftarrow pts^{pos}$, $\bar{p}ts^{pos} \leftarrow$ random from $T^{\tilde{gt}}$
for $t$ in range$(1, T)$ do
    Optimize $\bar{p}ts^{pos}$ as parameters:
    $\ell_{ce}(p\bar{t}s, T^{\tilde{gt}}) = -T^{\tilde{gt}} \log\left(SonoSAM(I, p\bar{t}s)\right)$
    Update $\bar{p}ts^{pos} \leftarrow \bar{p}ts^{pos}$
end for
Flip: $pts^{neg} \leftarrow \bar{p}ts^{pos}$
Output $O \leftarrow SonoSAM(I, pts)$
```

3.1 Positive Points Selection

We aim to select points that are the most representative of the coarse pathology segmentation mask as the positive prompt points. This selection objective translates to the partitioning around the medoids method's approach. This approach is preferable compared to a selection based on a minimization of the sum of squared distance (i.e., the k-means) in the case of multiple pathology blobs since the latter might select centroids in between pathology blobs. Thus, k mass centers of the coarse pathology segmentation masks are selected as positive points using the k- medoids clustering algorithm [27].

To reduce the probability of selecting false positive points, a threshold is applied to the coarse pathology segmentation masks before selection. We denote the selected positive points as $pts^{pos} = \{pts_i^{pos}\}_{i=1}^{k}$. This process is illustrated in Fig. 2.

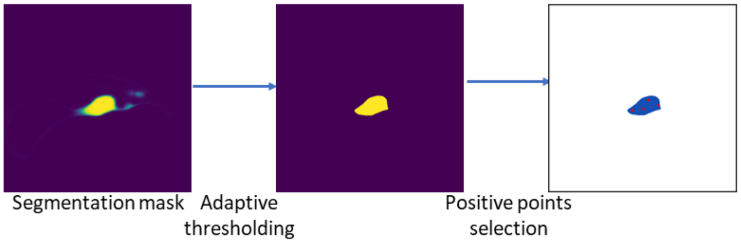


Figure 2: An illustration of our positive (foreground) points selection module, depicted in red. A threshold is applied to the coarse segmentation prediction. A k- medoids clustering algorithm is applied to select k positive pathology points.

3.2 Negative Points Refinement

We take inspiration from hard negative selection literature [28, 29, 30], and aim to select the most informative negative points w.r.t. the foreground object. To that end, we formulate a complementary prompt points selection problem w.r.t. the background given the k selected foreground points (3.1), $p\bar{t}s = \{p\bar{t}s^{pos}, p\bar{t}s^{neg}\}$. When the foreground is the pathology, the background is the context anatomy, herein the background is a tendon anatomy.

The complementary prompt points selection is optimized to decrease the binary cross-entropy (BCE) loss between the foundation model's zero-shot tendon segmentation mask prompted on these points and a modified ground-truth tendon mask, denoted $T^{\tilde{gt}}$. To avoid predicting tendon points within foreground pathology, the values of the ground-truth tendon mask overlapping with the coarse pathology detection are modified to zero. As points initialization for this complementary problem, we flip the labels of pts^{pos} such that they correspond to negative points, $\bar{pts}^{neg} \leftarrow pts^{pos}$. Further, k points are selected at random from $T^{\tilde{gt}}$, denoted $\bar{pts}^{pos}$. While freezing the foundation model, the point prompt optimization is performed for a maximum of 100 steps or until convergence. The optimization is performed such that the selected points are optimal w.r.t. the complementary problem of the tendon segmentation given the foreground pathology predicted by the coarse segmentor.

Denote an input image as I, SonoSAM's zero-shot tendon segmentation given input I and its corresponding optimized prompt points $\bar{pts}$ as $SonoSAM(I, \bar{pts})$. Then, the BCE loss of the complementary problem is:

$$\ell_{ce}(\bar{pts}, T^{\tilde{gt}}) = -T^{\tilde{gt}} \log\left(SonoSAM(I, \bar{pts})\right). \tag{1}$$

We used the AdamW [31] optimizer, with learning rate of $4e^{-3}$, and standard betas to optimize the positive points $\bar{pts}^{pos}$. The optimized positive tendon points selected by this model serve as k negative prompt points, $pts^{neg} \leftarrow \bar{pts}^{pos}$, towards the foreground pathology segmentation. This process is illustrated in Fig. 3.

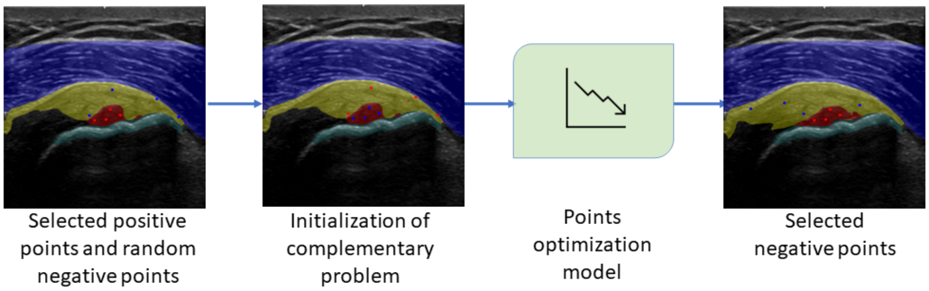

Selected positive points and random negative points
Initialization of complementary problem
Points optimization model
Selected negative points

Figure 3: An illustration of our negative (background) points selection module. In addition to the positive selected points (Sec. 3.1), negative points are selected randomly from the modified ground-truth tendon mask. The points are flipped to initialize the settings of the complementary tendon segmentation problem. Our points optimization model optimizes prompt points selection w.r.t. the complementary tendon zero-shot segmentation problem (Sec. 3.2). Finally, prompt points are again flipped to account for positive and negative prompt points towards the pathology segmentation.

4 Experiments

4.1 Dataset

The data used for this study is ultrasound images of tendons around the shoulder joint. Specifically, we acquired images of the supraspinatus tendon, infraspinatus tendon, and subscapularis. The images are acquired from both the short-axis and the long-axis views. The main parameters of our data are summarized in Table 1.

In this work, we aim to segment the partial tear pathology within the tendon, thus our data consists of images paired with the corresponding segmentation mask of anatomies and pathologies. Our data includes semantic labeling of the following classes: DITF, bone, tendon, and muscle. Table 2 summarizes the semantic labeling statistics.

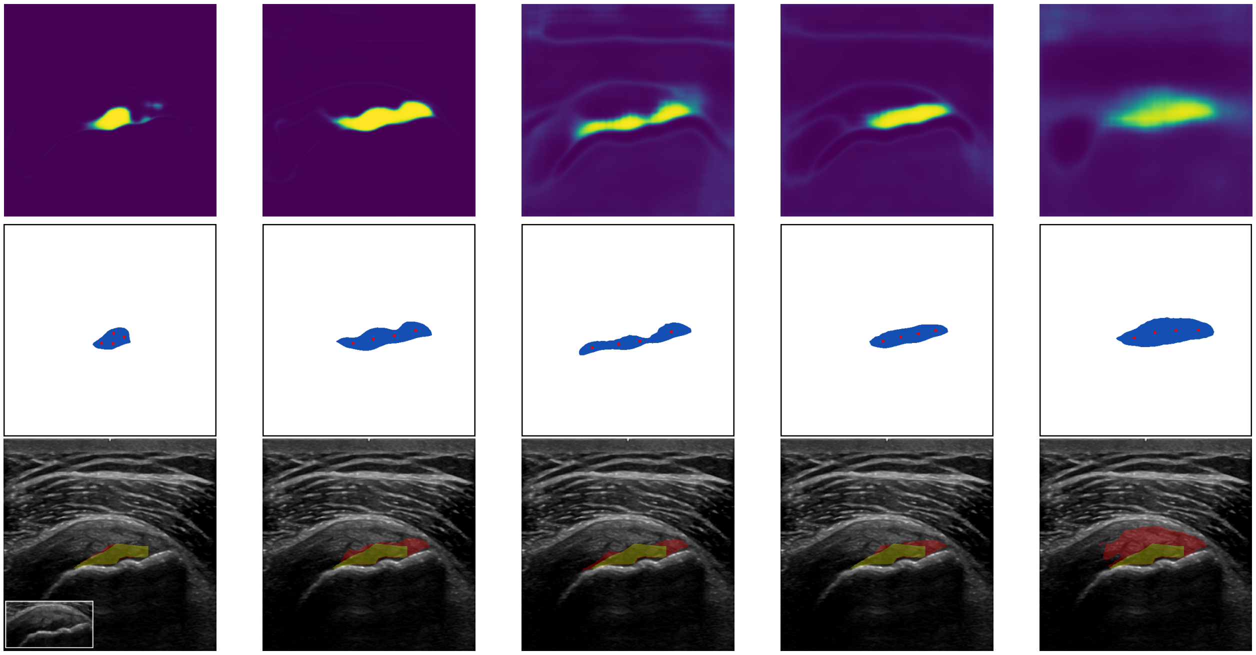
In total, our dataset includes 388 images from 124 subjects, 80% of which are used for training, and the remaining 20% are used for validation. The test set comprises 40 images. To prevent data leakage, the test set images are collected from subjects that do not appear in the train data. All images are resized to a constant resolution of 512x512 pixels. All data comply with the Institutional Review Board (IRB) data sharing agreement.

4.2 Evaluation Metric

We use the Dice similarity coefficient [32] evaluation metric, commonly used in medical image segmentation research to measure the overlapping pixels between prediction and ground truth masks. The Dice similarity coefficient is defined as $\frac{2|A \cap B|}{|A|+|B|}$, where A and B are the pixels of the prediction and the ground truth respectively.

4.3 A Segmentation Model In Low-Data Regimes

In this experiment, we investigate the performance and properties of a state-of-the-art semantic segmentation model with a limited training set size of MSK ultrasound images. Our goal is two-fold: (i) to validate our conjecture that high-quality



(a) 100% of train set. (b) 35% of train set. (c) 15% of train set. (d) 8% of train set. (e) 5% of train set.

Figure 4: Positive pathology points retainment in increasingly coarse segmentation mask prediction and our method's results. Top row: Pathology segmentation mask predicted with a DeepLabV3 model trained on varying percent of the training set. Middle row: Positive points selected on binary pathology mask by our positive points selection module. Bottom row: An illustration of our method's pathology segmentation output, highlighted in red, compared to the ground-truth segmentation, highlighted in green. The tendon area is shown at the bottom left image for reference. Our method achieves for this test image a Dice similarity coefficient of $0.89, 0.71, 0.73, 0.72, 0.50$ when the coarse segmentor is trained on $100\%, 35\%, 15\%, 8\%, 5\%$ of the train set, respectively.

Table 1: Summary of MSK pathology segmentation dataset main parameters.

Parameters/Dataset	**MSK Ultrasound Images**
Total frames	388
Original frame size	1536 X 796 or 1044 X 646 pixels
Subjects	90 (52.82% males, 47.18% females)
Average BMI	24.69 ± 8.92
Vendor	GE Healthcare™
Ultrasound system	Logiq S8™, Eagle™, LogiqE10™
Data collection	Linear
Collection Sites	USA, Israel

prompts can be extracted even from a coarse semantic segmentation prediction, and (ii) to measure the performance degradation in increasingly low-data regimes. These properties are the basis of our two-stage method for exploiting the advantages of a prompt-able foundation segmentation model. Concretely, for an input image $I \in \mathbb{R}^{512\times512}$ the segmentation model prediction $S \in \mathbb{R}^{7\times512\times512}$ corresponds to a semantic segmentation for each class as detailed in Table 2.

4.4 Segmentation Refinement With Zero-Shot Foundation Models

Positive Points Selection

A combination of a constant and an adaptive threshold is applied to the coarse segmentation prediction prior to positive point selection. Denote by c_0 the coarse segmentation mask prediction at the foreground channel (DITF in our case).

Table 2: Semantic labeling statistics at the 512X512 patches level. M: Million.

Class	**MSK Type**	**Number of images (% of total)**	**Total Area (pixels)**	**Mean fraction out of total patch area**
Discontinuity in tendon fiber	Pathology	179 (46.13%)	1.11M	1.09%
Bone	Anatomy	288 (74.22%)	2.75M	2.7%
Tendon		388 (100%)	10.64M	10.46%
Muscle		388 (100%)	28.13M	27.65%

We apply a double thresholding mechanism to disregard the noise in the prediction.

$$\tilde{c} = c_0 > t_{min} \tag{2}$$

$$c = \tilde{c} > 0.4 * \max(\tilde{c}) \tag{3}$$

The initial threshold screens predictions that lack sufficient global (cross-classes) certainty, when the minimum threshold is set to $t_{min} = 0.15$. The second thresholding term adaptively screens all predictions that lack sufficient local (class-wise) certainty. Further, we set the k-medoids++ medoid initialization method [33] which selects more separated initial medoids than those selected by the other methods. The hyper-parameter k is adaptively set such that the sum of distances of samples to their closest cluster center (inertia) is minimized, $k \in [4, 6]$.

Negative Points Refinement

We deploy in our experiments the SonoSAM semantic segmentation foundation model since it is expected to better generalize to zero-shot segmentation of ultrasound images than SAM.

Due to the randomness in the initialization of the complementary positive points $\bar{pts}^{pos}$ selection problem, evaluation is performed over 10 random initialization.

4.5 Training Procedure

Our coarse segmentor is DeepLabV3 [9], a state-of-the-art convolutional approach to handle objects in images of varying scales, with a ResNet-50 backbone [34]. As our complete dataset consists of only 275 training images, the model is pre-trained on the ImageNet dataset [35]. To evaluate our method across different data regimes we trained our coarse segmentor on varying n percentage of the training data, $n \in [5, 8, 12, 20, 35, 60, 100]$, sub-sampled at random. The model is trained with equally weighted BCE loss and a Dice similarity coefficient loss between the predicted and ground-truth segmentation for each class. Each such experiment is trained for 100 epochs, where the weights of the maximal validation loss have been selected for testing. We used the robust AdamW [31] optimizer, with no learning rate scheduler and parameters of $\beta_1 = 0.9, \beta_2 = 0.999$ and learning rate of $4e^{-3}$. The test set remains constant across the different training experiments. The model training and evaluation code is implemented with the PyTorch [36] framework.

5 Results

5.1 Semantic Segmentation Model In Low-Data Regimes

The results of this experiment validate our conjecture that positive pathology points are consistently selected in increasingly coarse segmentation mask predictions.

As the segmentation model is trained on an increasingly smaller training set, the segmentation mask prediction becomes coarse: the pathology segmentation boundaries become less defined and its prediction probability decreases (Fig. 4, top row). Nevertheless, the positive pathology points selected by our method remain generally consistent (Fig. 4, middle row). Consistent with these results, we find that the average Dice similarity coefficient of the segmentation model decreases rapidly when the model is trained on increasingly smaller training set sizes (Fig. 5, 'Segmentation Model'). These results validate our method's motivation and approach.

5.2 Semantic Segmentation Refinement With Zero-Shot Foundation Model

Fig. 5 summarizes the results of our method in comparison with those of the baseline segmentation model in various training set sizes. Our method's average Dice is higher than the baseline's in every training set size. Moreover, our method's performance gain is larger as the training set size decreases ($\sim 10\%$ average Dice increase in 5% and 8% training set sizes), substantiating the advantage of our method in low-data regimes. Our method's pathology segmentation output in varying training set sizes compared to the ground-truth segmentation is illustrated in Fig. 4, bottom row.

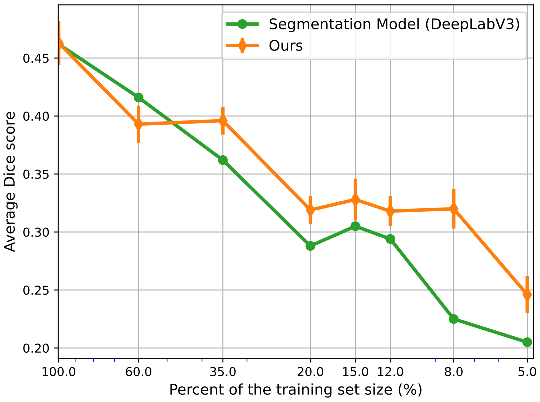


Figure 5: A summary of the average DITF Dice similarity coefficient of methods in various training set sizes. Depicted are the results of the baseline segmentation model[9] and our segmentation refinement with zero-shot SonoSAM foundation model. Error bars depict the standard deviation of our method's statistics.

To analyze the stochasticity effect of our method's negative points random initialization (Sec. 3.2), we compare our method's DITF Dice score statistics over ten random initialization and the baseline segmentation model's average DITF Dice similarity coefficient. Results show that our method's performance is robust, exhibiting relatively low standard deviation in all train set sizes (Fig. 5). Additionally, our method's mean DITF Dice surpasses that of the baseline's in all but one train set size, and is higher by 4% on average than the baseline.

5.3 Ablation Studies

In this section, we present ablation studies substantiating the effectiveness of our negative prompt points refinement (NPPR) model, as well as examining our method's performance when replacing the SonoSAM foundation model with SAM.

5.3.1 SAM vs. SonoSAM as a segmentation foundation model

In this study, we investigate the impact of replacing SonoSAM with SAM as the zero-shot semantic segmentation foundation model in our method. Table 3 shows that harnessing SonoSAM's generalizability for MSK ultrasound images is preferable to SAM in low-data regimes and on par with SAM otherwise.

5.3.2 Random negative prompt points section

In this experiment, we investigate the effectiveness of our negative prompt points refinement model by comparing it to a random negative prompt points selection algorithm. Concretely, k negative prompt points are randomly selected from the modified ground-truth tendon mask, T^{gt}. Our positive points selection approach remains unchanged. Results in Table 3 demonstrate that this naive selection algorithm achieves subpar average Dice scores across almost all train set sizes, especially in low-data regimes. These results establish the advantage of our negative points optimization algorithm.

6 Conclusions

In this paper, we address the performance degradation of a state-of-the-art semantic segmentation model in low-data regimes. A novel prompt points selection algorithm optimized on a zero-shot segmentation foundation model was presented, as a means of refining a coarse pathology segmentation.

Our method's advantages are brought to light in varying degrees of low-data regimes experiments, demonstrating a larger performance gain compared to the baseline segmentation model as the training set size decreases (Fig. 5).

Table 3: Ablation studies: quantitative segmentation test results of the mean DITF Dice similarity coefficient (DSC) for different approaches over 10 run cycles. Our method is using zero-shot SonoSAM [6] foundation model. A higher DSC is better, with the best scores marked in bold. NPPR: Negative Prompt Points Refinement.

Methods	Percent of the training set							
	100%	60%	35%	20%	15%	12%	8%	5%
Ours without NPPR	44.6%	40.0%	34.2%	27.8%	30.3%	27.5%	20.7%	16.6%
Ours with SAM	45.5%	**41.6%**	**39.7%**	29.3%	**32.9%**	28.3%	27.6%	23.0%
Ours	**46.3%**	39.3%	39.6%	**31.9%**	32.8%	**31.8%**	**32.0%**	**24.6%**

Further, we validate our method's robustness to negative point initialization stochasticity and study the effectiveness of our prompt points refinement model (Section 5.3.2). Results demonstrate that the generalization of SonoSAM in extremely low data regimes is better than SAM's (Section 5.3.1). Our approach can be used for other ultrasound-based medical diagnostics tasks. An inherent limitation of our two-stage method is that its latency is higher than that of a core segmentation model.

References

[1] Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. Deep learning in medical ultrasound analysis: a review. *Engineering*, 5(2):261–275, 2019.

[2] Puyang Wang, Vishal M. Patel, and Ilker Hacihaliloglu. Simultaneous segmentation and classification of bone surfaces from ultrasound using a multi-feature guided cnn. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 134–142, Cham, 2018. Springer International Publishing.

[3] R. Aggarwal, V. Sounderajah, G. Martin, D. S. W. Ting, A. Karthikesalingam, D. King, H. Ashrafian, and A. Darzi. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med*, 4(1):65, Apr 2021.

[4] Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, Jan 2019.

[5] 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, 2023.

[6] Hariharan Ravishankar, Rohan Patil, Vikram Melapudi, Parminder Bhatia, Kass-Hout Taha, and Pavan Annangi. Sonosam – segment anything on ultrasound images, 2023.

[7] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, Sijing Liu, Haozhe Chi, Xindi Hu, Kejuan Yue, Lei Li, Vicente Grau, Deng-Ping Fan, Fajin Dong, and Dong Ni. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, February 2024.

[8] 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.

[9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *ArXiv*, abs/1706.05587, 2017.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[14] Sheng He, Rina Bao, Jingpeng Li, Jeffrey Stout, Atle Bjornerud, P. Ellen Grant, and Yangming Ou. Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets, 2023.

[15] Maciej A. Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis*, 89:102918, October 2023.

[16] Xinrong Hu, Xiaowei Xu, and Yiyu Shi. How to efficiently adapt large segmentation model(sam) to medical images, 2023.

[17] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images, 2023.

[18] Chengyin Li, Prashant Khanduri, Yao Qiang, Rafi Ibn Sultan, Indrin Chetty, and Dongxiao Zhu. Auto-prompting sam for mobile friendly 3d medical image segmentation, 2023.

[19] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Xiaokang Yang, and Wei Shen. Sam-parser: Fine-tuning sam efficiently by parameter space reconstruction, 2023.

[20] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[21] Weijia Feng, Lingting Zhu, and Lequan Yu. Cheap lunch for medical image segmentation by fine-tuning sam on few exemplars, 2023.

[22] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation, 2023.

[23] Xinrun Chen, Chengliang Wang, Haojian Ning, and Shiying Li. Sam-octa: Prompting segment-anything for octa image segmentation, 2023.

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[25] Deepa Anand, Gurunath Reddy M, Vanika Singhal, Dattesh D. Shanbhag, Shriram KS, Uday Patil, Chitresh Bhushan, Kavitha Manickam, Dawei Gui, Rakesh Mullick, Avinash Gopal, Parminder Bhatia, and Taha Kass-Hout. One-shot localization and segmentation of medical images with foundation models, 2023.

[26] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. Sam on medical images: A comprehensive study on three prompt modes, 2023.

[27] Leonard Kaufmann and Peter Rousseeuw. Clustering by means of medoids. *Data Analysis based on the L1-Norm and Related Methods*, pages 405–416, 01 1987.

[28] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *ArXiv*, abs/2010.04592, 2020.

[29] Hakim Hafidi, Mounir Ghogho, Philippe Ciblat, and Ananthram Swami. Negative sampling strategies for contrastive self-supervised learning of graph representations. *Signal Processing*, 190:108310, 09 2021.

[30] Huangjie Zheng, Xu Chen, Jiangchao Yao, Hongxia Yang, Chunyuan Li, Ya Zhang, Hao Zhang, Ivor Tsang, Jingren Zhou, and Mingyuan Zhou. Contrastive attraction and contrastive repulsion for representation learning, 2023.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[32] Aaron Carass, Snehashis Roy, Adrian Gherman, Jacob Reinhold, Andrew Jesson, Tal Arbel, Oskar Maier, Heinz Handels, Mohsen Ghafoorian, Bram Platel, Ariel Birenbaum, Heather Greenspan, Dzung Pham, Ciprian Crainiceanu, Peter Calabresi, Jerry Prince, William Roncal, Russell Shinohara, and Ipek Oguz. Evaluating white matter lesion segmentations with refined sørensen-dice analysis. *Scientific Reports*, 10:8242, 05 2020.

[33] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[35] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.