Self-Sampling Meta SAM: Enhancing Few-shot Medical Image Segmentation with Meta-Learning

Tianang Leng*
Huazhong University of Science and Technology
Wuhan, China
tianangl@hust.edu.cn

Yiming Zhang*
Tokyo Institute of Technology
Tokyo, Japan
zhang.y.bl@m.titech.ac.jp

Kun Han
University of California, Irvine
Irvine, California, United States
khan7@uci.edu

Xiaohui Xie
University of California, Irvine
Irvine, California, United States
xhx@ics.uci.edu

Abstract

While the Segment Anything Model (SAM) excels in semantic segmentation for general-purpose images, its performance significantly deteriorates when applied to medical images, primarily attributable to insufficient representation of medical images in its training dataset. Nonetheless, gathering comprehensive datasets and training models that are universally applicable is particularly challenging due to the long-tail problem common in medical images.

To address this gap, here we present a Self-Sampling Meta SAM (SSM-SAM) framework for few-shot medical image segmentation. Our innovation lies in the design of three key modules: 1) An online fast gradient descent optimizer, further optimized by a meta-learner, which ensures swift and robust adaptation to new tasks. 2) A Self-Sampling module designed to provide well-aligned visual prompts for improved attention allocation; and 3) A robust attention-based decoder specifically designed for few-shot medical image segmentation to capture relationship between different slices. Extensive experiments on a popular abdominal CT dataset and an MRI dataset demonstrate that the proposed method achieves significant improvements over state-of-the-art methods in few-shot segmentation, with an average improvements of 10.21% and 1.80% in terms of DSC, respectively. In conclusion, we present a novel approach for rapid online adaptation in interactive image segmentation, adapting to a new organ in just 0.83 minutes. Code is available at https://github.com/DragonDescentZerotsu/SSM-SAM

1. Introduction

Medical image segmentation plays a pivotal role in clinical applications, including disease diagnosis, abnormality detection and treatment planning. Historically, the segmentation of anatomical structures was performed manually by experienced physicians, a laborious and time-consuming task.

Recent advancements in deep learning offer automated segmentation tools that deliver near-human accuracy swiftly. Nevertheless, these tools need extensive training on large annotated datasets for optimal performance, which are costly and time-intensive since they require inputs from experts with extensive clinical experience. Thus, in the field of medical imaging, few-shot learning [29, 49, 60] has gained significant interest from researchers because it is able to segment accurately without extensive labeled data. In fact, to achieve optimal performance on unseen classes, few-shot learning models must excel in extracting representative features from limited data. However, current few-shot learning frameworks for medical image segmentation [15, 50, 54] predominantly pretrain their models using data from a single medical domain, often with restricted datasets, leading to limited feature extraction ability. If we can leverage the strong feature extraction abilities obtained from extensive training data of diverse domains, the model can capture more distinctive features, enhancing its adaptability to new tasks.

Recently, Segment anything model (SAM) [28] has attracted significant attention. Trained on a large segmentation dataset of over 1 billion masks across various domains of natural images, SAM has strong feature extraction and generalization abilities and can segment any object on a certain image. Given SAM’s excellent zero-shot
transferability, a natural idea is to directly apply SAM for medical image segmentation to address the issue of relatively scarce medical image data [20, 46]. However, recent studies [40, 47] have shown that the performance of SAM is overall moderate and varies significantly across different datasets and different cases, demonstrating the potential promise of SAM within the context of medical images but also shows that the model cannot be applied directly with high confidence. Such phenomenon is attributed to the vast differences between natural and medical images, as SAM was primarily trained on natural images. To better adapt SAM to medical images, most prior studies focused on integrating lightweight Adapters [10, 56] or on freezing the heavy image and prompt encoders, opting to fine-tune solely the mask decoder [23, 31, 38]. However, the limited data available for unseen classes will hinder the few-shot segmentation performance of these approaches.

Driven by the desire to maximize the powerful extracted features of SAM through prompts [61], and to leverage its potential zero-shot capabilities without vast training data, we introduce the Self-Sampling Meta SAM (SSM-SAM) framework, which utilizes a Meta-learning method MAML++. Our framework consists of two main parts. The first part is a redesigned backbone that removes the original SAM’s prompt encoder and mask decoder; instead, we employ a self-sampling prompt encoder and Flexible Mask Attention Decoder (FMAD). Also, we integrate adapters into the image encoder as previous works do, allowing SAM to learn features of unfamiliar tasks. This enables SAM to have better transferability for few-shot learning and is termed SS-SAM (Self-Sampling SAM without MAML++). The second part, which we refer to as SSM-SAM (with MAML++), serves as our final framework for few-shot learning. It is a meta-learning based optimizer layered on top of this backbone to further boost SAM’s few-shot performance on medical images.

We performed experiments for few-shot learning on an abdomen CT dataset and an MRI dataset. We also utilized a fully supervised medical image segmentation task to evaluate the performance of our SS-SAM backbone (without MAML++) in comparison to prior methods on CT dataset.

Our contributions are summarized as follows:

- An effective online parameter adaptation technique optimized by a MAML++ meta-learner to enhance SAM’s generalization and few-shot learning capacities.

- A positive-negative self-sampling module that can generate aligned visual prompts to better extract the contextual relationship.

- A novel Flexible Mask Attention Decoder specifically designed for medical image few-shot segmentation.

- Our method outperformed the SOTA framework for few-shot medical image segmentation by an average of 10.21% on MICCAI15 Multi-Atlas Abdomen Labeling challenge dataset [4] and 1.80% on ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge [27] in terms of DSC.

Each module we use is plug-and-play, aiming to facilitate the deployment of fast and robust online image segmentation systems across various industries and making our model a baseline for improving performance where foundational models like SAM have struggled in the past.

2. Related Work

Foundation Models. A foundation model is essentially a large pre-trained model, typically developed using self-supervised learning across diverse datasets. This allows it to be fast adapted to specific tasks through mechanisms by fine-tuning or in-context learning. At present, foundation models have reached a level of maturity in the NLP domain, with models like BERT [13], GPT-3 [7]. In the realm of computer vision, the Segment Anything Model (SAM) [28], pre-trained on 11M images with 1B masks, first introduced a foundation model for image segmentation. This model can interpret a variety of prompts given by users, such as points, boxes, masks, and texts, to generate user-specified segmentations and is adept at segmenting any object within an image. However, despite its zero-shot generalization capabilities, SAM has demonstrated sub-optimal performance across multiple downstream tasks [26]. This observation has stimulated ongoing research efforts to enhance SAM’s performance in these tasks [18, 38, 56, 63].

Adapters. Adapters have emerged as a powerful tool in the realm of transfer learning for NLP and computer vision. Introduced as a lightweight alternative to full model fine-tuning, adapters allow for model customization while preserving the pre-trained parameters [22]. Instead of updating all the parameters in the model, adapter methods insert small bottleneck layers in the model architecture, which are then fine-tuned while the original model parameters remain unchanged. Recently, the ViT-Adapter [12] has been developed to equip a standard ViT [16] with the capability to perform a variety of downstream tasks. In a further development, an Explicit Visual Prompting (EVP) [56] technique has facilitated the integration of explicit visual cues into the Adapters. In this work, we integrate such Adapters [10] into the image encoder of SAM to avoid training a large amount of parameters.

Visual Prompts. The application of visual prompts in computer vision tasks has its roots in interactive segmentation, a methodology that necessitates user input, such as clicks [11, 25, 34, 52, 58], bounding boxes [43, 57], or scribbles [5, 6, 33], or convert spatial queries to masks and feed them
3. Methodology

3.1. Overview of SSM-SAM architecture

Fig. 1 depicts the architecture of our few-shot learning framework. We keep the image encoder of SAM frozen and add a learnable Adapter layer [36] to each of the transformer layers for parameter-efficient training. After passing through the image encoder, we perform a self-sampling operation on the first image embedding to update it. Afterward, we feed the four output image embeddings into our Flexible Mask Attention Decoder (FMAD) to refine these embeddings and generate the predicted mask. On top of this backbone, we implement a MAML++ based meta-learner to search for optimal initialization parameters for rapid adaptation to different organs.

3.2. Few-shot Online Optimizer

Segmentation models’ clinical deployment has historically been hindered by a distribution mismatch between training and testing datasets. Given the impossibility of collecting comprehensive representative data, we have innovated in our training approach. Rather than creating a universal offline model, we designed our model to recognize and adapt to new data types, ensuring it remains relevant for unseen images. In our few-shot medical image segmentation scenario, we consider a distribution over organs, denoted as \( p(O) \), to which we aim our model to adapt. We represent our model as \( f_\theta \) with parameter \( \theta \), which is trained on images of different organs, \( I_{\alpha_i} \), following the distribution \( p(I_{\alpha_i}) \), where \( \alpha_i \) signifies the \( i^{th} \) organ. The online optimizer optimizes the parameters via back-propagating steps as follows:

\[
\theta_i' \leftarrow \theta_i' - \alpha \nabla_{\theta_i'} L_{\alpha_i}(f_{\theta_i'})
\] (1)

where \( \alpha \) is the learning rate and \( \theta_i' \) represents the model parameters adapted to fit the \( i^{th} \) organ. We employ balanced cross-entropy and IoU loss to supervise our network so \( L_{\alpha_i} \) can be expressed as:

\[
L_{\alpha_i} = L_{\text{bce}} + L_{\text{iou}}
\] (2)

The overview of the online optimizer is outlined in Algorithm 1.

Figure 1. Overview of our proposed framework for SSM-SAM. SSM-SAM backbone consists of three main components: a) Injecting Adapters into the encoder of SAM to quickly acquire task-related information. b) Using the Self-Sampling Module to replace the position encoding based prompt encoder to strengthen the relationship between the feature and the prompt. c) Flexible Mask Attention Decoder (FMAD) to enhance boundary information and refine the final generated mask map. On top of the backbone, we employ a meta-learning based online optimizer.
3.3. Meta-learning

We employ meta-learning for our model training to guarantee that the initial parameters $\theta$ exhibit robust generalization capabilities, facilitating rapid online optimization. MAML++ is an ideal strategy for our purpose, as it is designed to learn suitable initial model parameters optimized for swift adaptation to new tasks. The complete algorithm is detailed in Algorithm 2. Note that the outermost for-loop is the meta-learner, which is defined as follows:

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \frac{1}{T} \sum_{i} L_{o_{i}}(f_{\theta_{i}}^{t})$$  \hspace{1cm} (3)

where $i$ is the $i^{th}$ organ. $T$ is the number of organs in a task batch for optimizing the meta-learner. We incorporate the cosine annealing learning rate as recommended in [1] to enhance our model’s efficiency in fitting the training set.

$$\text{CosineAnnealingLR}(\beta, \epsilon)$$

The image encoder of SAM uses a Vision Transformer (ViT) [16], which is pre-trained with MAE [19]. When given an image of any size, it is essential to first resize it to a resolution of 1024 x 1024. The ViT then processes this resized image to generate an image embedding with dimensions $C \times H \times W$. For this research, we chose the ViT-B variant (SAM-b), (where $C = 256$, $H = 64$, $W = 64$). To enable efficient learning with faster updates and address the issue of excessive GPU memory usage, we keep the image encoder frozen and inject a trainable Adapter to each of the 12 transformer layers as mentioned before. We only train the parameters within Adapters, which are tasked with learning task-specific knowledge and low-level structural information from the features extracted from image embeddings. For the $k$-th Adapter layer, we take patch embedding $F^{k}$ as input and obtain updated parameters $P^{k}$

$$P^{k} = \text{MLP}_{\text{up}}(\text{GELU}(\text{MLP}_{\text{tune}}^{k}(F^{k})))$$  \hspace{1cm} (4)

where $\text{MLP}_{\text{tune}}^{k}$ refers to a linear layer within each Adapter for generating distinct prompts. $\text{MLP}_{\text{up}}$ is an up-projection layer shared across all Adapters designed to match the dimension of transformer features. $P^{k}$ represents the output associated with each transformer layer.

Also, we divided image encoder to four sub-blocks, which is carried out with the intent of deriving multi-level information [32]. After passing through the image encoder, these four image embeddings of size $B \times 256 \times 64 \times 64$ are fed into the subsequent module.

3.4. Adapted Image Encoder

Figure 2. Structure of queries and prompt interaction during training and evaluation.

We employ a positive-negative attention self-sampling module inspired by [65] to convert all kinds of non-textual queries to visual prompts that lie in the same visual embedding space, as opposed to the conventional position encoding in SAM [28]. Our algorithm is illustrated in Fig. 3 and
can be summarized as follows:

\[ P = \text{concatenate}([P_{\text{positive}}, P_{\text{negative}}]) \]  \hspace{1cm} (5)

\[ P_v = \text{Self-Sampler}(P, Z_h) \]  \hspace{1cm} (6)

where \(P_{\text{positive}}, P_{\text{negative}}\) stand for positive points (region we aim to segment) and negative points (background region) sampled from the image, respectively. \(Z_h\) is the feature maps extracted from the image and \(P\) denotes the point prompts.

Afterward, we concatenated these positive points and negative points to form the point prompts. This approach ensures that the model does not only focus on the positive parts, which could potentially lead to a higher false positive rate, but also adequately attends to the negative parts. With this method, the model can better discerns the organ boundaries, enhancing segmentation performance when we only give one point to prompt the model during inference.

On receiving the point prompts, we perform direct interpolation from the feature map to get the corresponding embeddings as visual prompts. This ensures that the prompts and the image embeddings share identical semantic features. We initialize a few tokens as learnable global queries and concatenate them with these visual prompts. After that, we apply self-attention \[51\] between a set of learnable queries and the visual prompts. Then, cross-attention is applied only between the image embeddings and these updated queries.

### 3.6. Flexible Mask Attention Decoder

Since we divided each 3D volume data into 12 chunks as in section 4.1, there is a certain pattern to how each organ changes across different slices. This observation inspired us to treat the segmentation task like a tracking task. Consequently, we introduced the Flexible Mask Attention Decoder (FMAD). The core of this decoder is the Flexible Mask Attention Module (FMAM), as illustrated in Fig 3. By leveraging cross-attention, we can expand or contract the blurred support mask to predict the mask for the query image, mitigating the challenge of direct mask prediction. The Gaussian blurred support masks provide the model with ample room to explore potential mask positions while efficiently suppressing distractions from other parts of the image.

In decoding, the attention block bridges the Gaussian blurred support mask and queries images to update the mask. We first apply a Gaussian filter on original support mask \(M \in \mathbb{R}^{H \times W \times 1}\), repeat it for \(C\) times and reshape to \(N \times C\) to get \(M' \in \mathbb{R}^{N \times C}\). Then we can easily cast it on query feature \(F \in \mathbb{R}^{N \times C}\) extracted from image encoder to get \(K \in \mathbb{R}^{N \times C}\) and \(Q \in \mathbb{R}^{N \times C}\) as follows:

\[ K = (M' \odot F) W^K \]  \hspace{1cm} (7)

where \(\odot\) means element-wise production, \(W^K \in \mathbb{R}^{C \times C}\) means key projection matrix. Following \[51\], we also adopt the dot-product to compute the similarity matrix \(A_{K \rightarrow Q} \in \mathbb{R}^{N \times C}\) between the query and key as follows:

\[ A_{K \rightarrow Q} = \text{Atten}(Q, K) = \text{Softmax}_{\text{col}}(\bar{Q} \bar{K}^T / \tau) \]  \hspace{1cm} (8)

where \(\bar{Q}\) and \(\bar{K}\) are \(l_2\)-normalized features of \(Q\) and \(K\) across the channel dimension, and \(\tau\) is a temperature parameter controlling the Softmax distribution, which is the same as in \[55\]. Then we convert \(M'\) to \(V \in \mathbb{R}^{N \times C}\) though:

\[ V = (M') W^V \]  \hspace{1cm} (9)
where $W^V \in \mathbb{R}^{C \times C}$ denotes value projection matrix. With the attention matrix $A_{K \rightarrow Q}$ from key to query, we can transform the value via $A_{K \rightarrow Q}V \in \mathbb{R}^{N \times C}$.

Unlike typical attention layers, at the first residual connection, we use multiply & Norm instead of Add & Norm to propagate query feature $F$ as follows:

$$F' = \text{InsNorm}(A_{K \rightarrow Q}V \odot F)$$  \hspace{1cm} (10)

By using the updated Gaussian mask from the attention layer, we can further refine the predicted mask. After processing each of the four image embeddings, we upsample, concatenate, and put them through a convolutional layer. This approach facilitates improved information integration across layers to predict the final mask.

4. Experiments

4.1. Setup

Datasets To verify the generality and robustness of our approach, we perform experiments on two datasets, ABD-30 and ABD-MRI. We evaluate the complete few-shot segmentation framework (SS-SAM) on both datasets, while the backbone (SS-SAM) is only assessed on ABD-30 to validate its efficiency.

- ABD-30 (from MICCAI15 Multi-Atlas Abdomen Labeling challenge [4]) contains 30 cases with 3779 axial abdominal clinical CT images.

- ABD-MRI (from Combined Healthy Abdominal Organ Segmentation (CHAOS) challenge [27] held in IEEE International Symposium on Biomedical Imaging (ISBI) 2019) contains 20 3D abdominal MRI scans with total four different labels.

Liver, spleen and left and right kidney are used as semantic classes following previous settings [14, 42, 50]. Within each experiment, one organ is considered as unseen semantic class for testing while the rest are used for training.

Evaluation Metrics We use the Dice Similarity Coefficient (DSC) to evaluate the prediction mask $m$ against the ground truth mask $g$:

$$\text{DSC}(m, g) = \frac{2 |m \cap g|}{|m| + |g|}$$  \hspace{1cm} (11)

Implementation Details All the images extracted from 3D volume data are reshaped to $1024 \times 1024$ to fit into the SAM model. We follow the same protocol used in [42, 45, 50] to do 1-way 1-shot learning by dividing the 3D CT scans into 12 chunks and segmenting all the query slices in one chunk by using the center slice in the chunk as the support image, see Fig 4 for visual representation. We use ViT-B [16] version of SAM and supervise our network using balanced cross entropy loss and IoU loss as Eq. (2) between the predicted mask and ground truth mask. AdamW optimizer [37] is used for all the experiments with an initial learning rate of 2e-4. Cosine decay [1] is applied to the learning rate. Few-shot segmentation task is trained for 50 epochs on a single NVIDIA RTX 3090 GPU and fully supervised segmentation task is trained for 50 epochs using PyTorch on a single NVIDIA RTX 3090 GPU.

4.2. Main Results

Results on Few-shot Medical Image Segmentation Table 1 shows the few-shot segmentation performance of SSM-SAM with previous work on ABD-30 and ABD-MRI respectively. SE-Net [45] represents the first architecture designed explicitly for few-shot medical image segmentation. PANet [54] is an extended version of the widely-used prototypical network [49], tailored for natural image segmentation. SSL-ALPNet [42] integrates self-supervised learning with prototypical networks. Affine denotes the accuracy result after aligning the support and query images globally using an affine transformation. RP-Net [50] stands as the state-of-the-art framework for few-shot medical image segmentation, leveraging both a context relation encoder and a recurrent module. [50] reported performance for all the methods above, so these numbers are directly quoted.

First, compared to state-of-the-art method RP-Net, SSM-SAM outperforms RP-Net by 10.21% and 1.80% on ABD-30 and ABD-MRI respectively. Second, the use of FMAD results in improvements of 2.17% ~ 3.41% and 1.84% ~ 2.62% respectively. Additionally, employing the MAML++ based optimizer leads to enhancements of 5.98% ~ 7.22% and 4.47% ~ 5.25% respectively. This indicates that both our meta-optimizer and FMAD are highly effective.

Results on Fully Supervised Medical Image Segmentation To evaluate the performance of our backbone (without MAML++), we compare SS-SAM with recent SOTA methods on ABD-30 dataset without Flexible Mask Attention Module, including U-Net [44], Att-Unet [41], TransUnet [9], Swin-UNet [8], MissFormer [24], TransDeepLab [3], HiFormer [21], DAE-Former [2] and SAMed [62] following [62]. For a fair comparison, we integrate the self-sampling module into some of the state-of-the-art models because the original SAM and our method utilize visual prompts, which incorporate ground truth information into the model while previous state-of-the-art models did...
Figure 5. Testing process of MAML++ based model (in deep blue) and Fine-tuning of a model pretrained on the same distribution of tasks without MAML++ (in light blue)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Spleen ↑</th>
<th>Kidney L ↑</th>
<th>Kidney R ↑</th>
<th>Liver ↑</th>
<th>mean ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABD-30</td>
<td>SE-Net [45]</td>
<td>0.23</td>
<td>32.83</td>
<td>14.34</td>
<td>0.27</td>
<td>11.91</td>
</tr>
<tr>
<td></td>
<td>PANet [54]</td>
<td>25.59</td>
<td>32.34</td>
<td>17.37</td>
<td>38.42</td>
<td>29.42</td>
</tr>
<tr>
<td></td>
<td>SSL-ALPNet [42]</td>
<td>60.25</td>
<td>63.34</td>
<td>54.82</td>
<td>73.65</td>
<td>63.02</td>
</tr>
<tr>
<td></td>
<td>Affine</td>
<td>48.99</td>
<td>43.44</td>
<td>45.67</td>
<td>68.93</td>
<td>51.75</td>
</tr>
<tr>
<td></td>
<td>RP-Net [50]</td>
<td>72.19</td>
<td>75.03</td>
<td>70.89</td>
<td>80.53</td>
<td>74.66</td>
</tr>
<tr>
<td></td>
<td>SS-SAM (w/o MAML++, w/o FMAD)</td>
<td>77.24</td>
<td>72.89</td>
<td>79.78</td>
<td>72.00</td>
<td>75.48</td>
</tr>
<tr>
<td></td>
<td>SS-SAM (w/o MAML++, w/ FMAD)</td>
<td>78.99</td>
<td>73.33</td>
<td>83.49</td>
<td>74.81</td>
<td>77.65</td>
</tr>
<tr>
<td></td>
<td>SSM-SAM (w/ MAML++, w/o FMAD)</td>
<td>80.55</td>
<td>77.83</td>
<td>82.79</td>
<td>82.70</td>
<td>81.46</td>
</tr>
<tr>
<td></td>
<td>SSM-SAM (w/ MAML++, w/ FMAD)</td>
<td><strong>86.95</strong></td>
<td><strong>80.96</strong></td>
<td><strong>84.47</strong></td>
<td><strong>87.12</strong></td>
<td><strong>84.87</strong></td>
</tr>
<tr>
<td>ABD-MRI</td>
<td>SE-Net [45]</td>
<td>51.80</td>
<td>62.11</td>
<td>61.32</td>
<td>27.43</td>
<td>50.66</td>
</tr>
<tr>
<td></td>
<td>PANet [54]</td>
<td>50.90</td>
<td>53.45</td>
<td>38.64</td>
<td>42.26</td>
<td>46.33</td>
</tr>
<tr>
<td></td>
<td>SSL-ALPNet [42]</td>
<td>67.02</td>
<td>73.63</td>
<td>78.39</td>
<td>73.05</td>
<td>73.02</td>
</tr>
<tr>
<td></td>
<td>Affine</td>
<td>62.87</td>
<td>64.70</td>
<td>69.10</td>
<td>65.00</td>
<td>65.41</td>
</tr>
<tr>
<td></td>
<td>RP-Net [50]</td>
<td>75.69</td>
<td>79.30</td>
<td><strong>84.66</strong></td>
<td>71.51</td>
<td>77.79</td>
</tr>
<tr>
<td></td>
<td>SS-SAM (w/o MAML++, w/o FMAD)</td>
<td>71.99</td>
<td>76.52</td>
<td>72.13</td>
<td>69.36</td>
<td>72.50</td>
</tr>
<tr>
<td></td>
<td>SS-SAM (w/o MAML++, w/ FMAD)</td>
<td>73.64</td>
<td>78.04</td>
<td>76.69</td>
<td>72.13</td>
<td>75.12</td>
</tr>
<tr>
<td></td>
<td>SSM-SAM (w/ MAML++, w/o FMAD)</td>
<td>76.77</td>
<td>80.47</td>
<td>77.44</td>
<td>76.32</td>
<td>77.75</td>
</tr>
<tr>
<td></td>
<td>SSM-SAM (w/ MAML++, w/ FMAD)</td>
<td><strong>78.81</strong></td>
<td><strong>81.70</strong></td>
<td><strong>80.38</strong></td>
<td><strong>77.50</strong></td>
<td><strong>79.59</strong></td>
</tr>
</tbody>
</table>

Table 1. DSC comparison with other methods on ABD-30 and ABD-MRI for few-shot learning (unit:%).

<table>
<thead>
<tr>
<th>Method</th>
<th>DSC ↑</th>
<th>Aorta</th>
<th>Gallbladder</th>
<th>Kidney(L)</th>
<th>Kidney(R)</th>
<th>Liver</th>
<th>Pancreas</th>
<th>Spleen</th>
<th>Stomach</th>
</tr>
</thead>
<tbody>
<tr>
<td>TransUNet [9]</td>
<td>77.48</td>
<td>87.23</td>
<td>63.13</td>
<td>81.87</td>
<td>77.02</td>
<td>94.08</td>
<td>55.86</td>
<td>85.08</td>
<td>75.62</td>
</tr>
<tr>
<td>SwinUnet [8]</td>
<td>79.13</td>
<td>85.47</td>
<td>66.53</td>
<td>83.28</td>
<td>79.61</td>
<td>94.29</td>
<td>56.58</td>
<td>90.66</td>
<td>76.60</td>
</tr>
<tr>
<td>MissFormer [24]</td>
<td>81.96</td>
<td>86.99</td>
<td>68.65</td>
<td>85.21</td>
<td>82.00</td>
<td>94.41</td>
<td>65.67</td>
<td>91.92</td>
<td>80.81</td>
</tr>
<tr>
<td>HiFormer [21]</td>
<td>80.39</td>
<td>86.21</td>
<td>65.69</td>
<td>85.23</td>
<td>79.77</td>
<td>94.61</td>
<td>59.52</td>
<td>90.99</td>
<td>81.08</td>
</tr>
<tr>
<td>DAE-Former [2]</td>
<td>82.43</td>
<td>88.96</td>
<td>72.30</td>
<td>86.08</td>
<td>80.88</td>
<td>94.98</td>
<td>65.12</td>
<td>91.94</td>
<td>79.19</td>
</tr>
<tr>
<td>PaNN [64]</td>
<td>90.8</td>
<td>92.5</td>
<td>72.9</td>
<td>95.3</td>
<td>92.0</td>
<td>95.3</td>
<td>-</td>
<td><strong>96.8</strong></td>
<td>-</td>
</tr>
<tr>
<td><strong>MissFormer</strong> [24]</td>
<td><strong>78.84</strong></td>
<td><strong>85.89</strong></td>
<td><strong>64.75</strong></td>
<td><strong>84.19</strong></td>
<td><strong>76.98</strong></td>
<td><strong>94.18</strong></td>
<td><strong>55.19</strong></td>
<td><strong>89.07</strong></td>
<td><strong>80.47</strong></td>
</tr>
<tr>
<td><strong>HiFormer</strong> [21]</td>
<td><strong>78.24</strong></td>
<td><strong>85.94</strong></td>
<td><strong>60.88</strong></td>
<td><strong>84.16</strong></td>
<td><strong>76.57</strong></td>
<td><strong>94.54</strong></td>
<td><strong>54.00</strong></td>
<td><strong>89.81</strong></td>
<td><strong>80.02</strong></td>
</tr>
<tr>
<td><strong>DAE-Former</strong> [2]</td>
<td><strong>80.46</strong></td>
<td><strong>87.26</strong></td>
<td><strong>67.46</strong></td>
<td><strong>82.15</strong></td>
<td><strong>77.53</strong></td>
<td><strong>94.36</strong></td>
<td><strong>63.32</strong></td>
<td><strong>90.85</strong></td>
<td><strong>80.77</strong></td>
</tr>
<tr>
<td>SAMed [62]</td>
<td>81.88</td>
<td>87.77</td>
<td>69.11</td>
<td>80.45</td>
<td>79.95</td>
<td>94.80</td>
<td>72.17</td>
<td>88.72</td>
<td>82.06</td>
</tr>
<tr>
<td>SAM Adapter [10]</td>
<td>92.01</td>
<td>95.35</td>
<td>87.88</td>
<td><strong>96.16</strong></td>
<td>96.73</td>
<td>97.23</td>
<td><strong>79.86</strong></td>
<td>90.60</td>
<td>92.26</td>
</tr>
<tr>
<td><strong>SS-SAM</strong></td>
<td><strong>93.09</strong></td>
<td><strong>95.96</strong></td>
<td><strong>89.77</strong></td>
<td><strong>96.09</strong></td>
<td><strong>96.79</strong></td>
<td><strong>97.61</strong></td>
<td>79.15</td>
<td>95.15</td>
<td><strong>94.23</strong></td>
</tr>
</tbody>
</table>

Table 2. Comparison to state-of-the-art models using a fully supervised method on the ABD-30 dataset. Best results are highlighted in bold (unit:%). For a fair comparison, the results in italics are from the corresponding model with the same self-sampling module. Notably, our modified SAM (SS-SAM) outperforms other leading models.

not leverage such additional information. The results are highlighted in blue italics in Tab. 2. Surprisingly, since these models are unaware of how to process this information, their performance slightly declined. We also observe
that SS-SAM achieves state-of-the-art performance. These experiments demonstrate that our models are capable of achieving high performance and also refute the potential claim that introducing prompts is a form of ”cheating”.

### 4.3. Ablation Study

**Effect of self-sampling.** As previously explained, the self-sampling module derives visual prompts by interpolating between the provided point coordinates and image embeddings. Subsequently, it conducts cross-attention with the image embeddings. On the other hand, SAM’s point prompt employs positional encoding on the points before performing cross-attention with the image embeddings. These two approaches differ in their treatment of the sampled points. To evaluate the effectiveness of the proposed methods, we performed ablation studies on four distinct organs: the spleen, left kidney, right kidney, and liver, using a fully supervised approach. For each organ, we kept the image encoder and decoder the same, with three different prompt methods: a) no prompt, b) point prompt with position encoding, and c) point prompt with self-sampling. As presented in Tab. 3, our self-sampling method achieved the highest mean support.

<table>
<thead>
<tr>
<th>Prompt</th>
<th>Spleen</th>
<th>Kidney L</th>
<th>Kidney R</th>
<th>Liver</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>no prompt</td>
<td>91.83</td>
<td>93.65</td>
<td>92.55</td>
<td>96.21</td>
<td>93.56</td>
</tr>
<tr>
<td>w/ position encoding</td>
<td>93.05</td>
<td>95.75</td>
<td><strong>96.87</strong></td>
<td>97.27</td>
<td>95.74</td>
</tr>
<tr>
<td>w/ self-sampling</td>
<td><strong>95.15</strong></td>
<td><strong>96.09</strong></td>
<td>96.79</td>
<td><strong>97.61</strong></td>
<td>96.41</td>
</tr>
</tbody>
</table>

Table 3. Ablation Study on no prompt, position encoding (pe) of points and self-sampling

**Figure 6.** Examples of predictions of SAM, SSM-SAM without FMAM and SSM-SAM with FMAM

**Effects of Meta-learner.** To highlight the few-shot learning advantages of our model achieved by integrating a meta-learner, apart from the experiment setting in Tab. 1, we set up a more rigorous experimental setting to compare performances with and without the meta-learner.

Following the configurations specified in [17][30], we selected only 125 images for the four organs. In each experiment, one organ is chosen as the test task (e.g., liver), while the remaining three organs (spleen, left kidney, right kidney) are used as training tasks. Only five images of each training task (spleen, left kidney, right kidney) are allocated during training. During testing, we adopt a 1-way, 5-shot approach, using five images of the test task (liver) as the support set for training. The remaining 120 images serve as the query set to evaluate the model’s performance.

Our findings were illuminating. A visual representation of the online optimization process of each of the four organs is provided in Fig.5. With just a limited number of training images, models equipped with the meta-learner outperformed those without by a significant margin of 12.07% on average. Furthermore, models augmented by the meta-learner not only demonstrated swifter convergence in the initial epochs but also exhibited enhanced stability and performance in the concluding epochs.

### 4.4. Qualitative Results

In Fig.6, we present masks produced by SAM and variants of our algorithm. Notably, SAM struggles to segment smaller organs or those with indistinct boundaries. Conversely, SSM-SAM with FMAM outperforms its counterpart without FMAM, effectively minimizing the impact of similar distracting regions in the CT scans.

### 5. Conclusions

In this paper, we introduce a universal approach, SSM-SAM, designed to optimize and adapt a foundational model such as the Segment Anything Model (SAM), for few-shot medical image segmentation. Our method incorporates a positive-negative Self-Sampling prompt encoder and a Flexible Mask Attention Decoder to enhance the contextual relationship and tiny boundary information essential for mask generation. Moreover, our fast online meta-learning based optimizer facilitates high performance even without extensive training data and can be plugged into other frameworks effortlessly. Experiments demonstrate that SSM-SAM outperforms the previous state-of-the-art approach by as much as 10% in terms of DSC. Furthermore, the proposed SSM-SAM can produce segmentations in just 50 seconds per organ, indicating its potential for real-time applications in medical settings.


[52] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image...


