

Article

Self-Prompting Segment Anything Model for Esophageal OCT Images

Cong Wang ^{1,*}  and Meng Gan ² ¹ School of Electronics and Information Engineering, Suzhou Polytechnic University, Suzhou 215104, China² School of Computer Engineering, Suzhou Polytechnic University, Suzhou 215104, China; 92452@jssvc.edu.cn

* Correspondence: 92453@jssvc.edu.cn

Abstract

Optical coherence tomography (OCT) is an established imaging modality for esophageal disease assessment, enabling high-resolution visualization of tissue layers. However, automated segmentation of esophageal OCT images remains challenging due to low contrast, complex tissue structures, in vivo confounders, and scarce labeled clinical data. To address these issues, we propose a self-prompting segment anything model (SAM)-based one-shot segmentation framework tailored for esophageal OCT images. The framework integrates a pre-trained SAM encoder for general feature extraction and a self-pretrained prompt encoder to capture domain-specific features from unlabeled OCT data. Experiments on a self-collected mouse dataset and a public human dataset yielded Dice similarity coefficient (DSC) values exceeding 85%, with the highest mean DSC among the evaluated methods under the one-shot protocol, although greater performance dispersion was observed in the heterogeneous human cohort. These findings demonstrate the potential of combining unlabeled OCT pretraining with one labeled B-scan for annotation-efficient esophageal OCT segmentation.

Keywords: optical coherence tomography; esophageal disease; image segmentation; segment anything model; few-shot learning

1. Introduction

Optical coherence tomography (OCT), first proposed in 1991, is a non-invasive, high-resolution imaging technique that employs low-coherence interferometry to generate cross-sectional images of biological tissues [1,2]. Characterized by real-time, non-destructive tissue visualization, its technological advancement has expanded clinical applicability, especially in the early diagnosis of retinal diseases [3,4]. In the context of esophageal disease analysis, OCT has proven highly effective for evaluating subsurface tissue architectures [5,6]. Specifically, morphological alterations in esophageal tissue layers, such as epithelial thickening, structural irregularities, or the complete loss of layered architecture, serve as critical clinical biomarkers for early-stage pre-malignant and malignant pathologies, including Barrett's esophagus and esophageal adenocarcinoma. However, OCT devices produce enormous volumes of image data, which are often contaminated by specific confounders such as artifacts and mucus [7]. Consequently, manual analysis demands highly experienced gastroenterologists, rendering the process time-consuming and prone to inter-observer subjectivity [8]. Automatic segmentation is therefore essential for rapidly and objectively quantifying key tissue characteristics (e.g., layer thickness and morphology). This automa-



Received: 6 June 2026

Revised: 7 July 2026

Accepted: 8 July 2026

Published: 11 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

tion may support targeted endoscopic biopsies, real-time optical diagnosis, and wider clinical adoption of OCT [9–11].

Beyond esophageal applications, deep learning has advanced automated OCT segmentation across various biological tissues. For instance, architectures incorporating soft-attention mechanisms and deep ensemble learning have been applied to structural segmentation in kidney OCT images [12] and retinal layer segmentation for age-related macular degeneration (AMD) classification [13]. Inspired by these broader developments in medical image analysis, existing research on esophageal OCT image segmentation has transitioned from traditional image processing methods to deep learning-based approaches. Early studies relied on traditional algorithms such as graph search, Canny edge detection, and sparse Bayesian classification, which achieved preliminary layer segmentation but suffered from poor generalization due to their dependence on manually designed features and high sensitivity to speckle noise [7,14]. Addressing these limitations, UNet and its improved variants have become mainstream in esophageal analysis [15]. Subsequent studies integrated techniques such as generative adversarial networks, attention mechanisms, and connectivity modeling to reduce topological errors and improve segmentation accuracy in *in vivo* human and animal esophageal OCT images [9,16]. Despite this progress, tissue morphological irregularities and limited training data continue to constrain the clinical translation of existing algorithms.

Recently, several foundation models have been developed using large-scale pretraining datasets [17,18]. These models learn transferable visual representations for downstream tasks. The segment anything model (SAM) is a representative foundation model designed for promptable image segmentation [19]. Trained on 11 million images paired with 1 billion masks, SAM has demonstrated strong zero-shot generalization on natural images. In the field of medical image analysis, SAM and its medical-adapted variants (e.g., AutoSAM [20], MedSAM [21], SAM-Med3D [22]) have been widely explored to address the limitations of domain-specific segmentation models, such as heavy reliance on annotated medical data and poor adaptability to diverse imaging modalities. These variants have been evaluated for segmenting lesions, anatomical structures, and tissue layers across CT, MRI, ultrasound, and OCT.

Although Meta released SAM2 [23], which introduces advancements in video and volumetric sequence processing, the original SAM remains an extensively validated, stable baseline for 2D medical image segmentation. Because esophageal OCT B-scans in this study are currently processed as independent 2D cross-sections to establish a fundamental framework, we utilize the original SAM architecture as our base model.

More specifically, SAM2's temporal memory is designed to condition the prediction of a current frame on information retained from other frames. Because no cross-frame state is constructed when each B-scan is processed independently, this mechanism offers limited additional benefit under the present protocol. Retaining the original SAM also keeps the foundation encoder consistent with our single-image self-prompting architecture, allowing the contribution of domain-specific prompt learning to be evaluated without introducing a sequence-memory component that is not used by the current task. However, the potential of SAM for esophageal OCT image segmentation remains underexplored and requires further investigation. Esophageal OCT images are typically acquired *in vivo* using endoscopic probes, which allows for the visualization of the layered architecture of the esophageal wall, including the epithelium, lamina propria, muscularis mucosae and submucosa. This 'optical biopsy' is clinically vital for the early detection of Barrett's esophagus (BE), a condition where the normal squamous lining is replaced by columnar epithelium, significantly increasing the risk of esophageal adenocarcinoma. However, the *in vivo* environment introduces significant clinical confounders. For example, tissue

folding occurs frequently due to the collapsible nature of the esophageal tube, which distorts layer morphology and complicates thickness measurements. Mucus coverage and luminal debris cause severe signal attenuation and scattering, often obscuring the interface between the epithelium and submucosa where early cancerous changes typically begin. Furthermore, signal loss due to probe-to-tissue distance or shadowing can mimic pathological tissue degradation. These factors not only hinder existing automated segmentation methods but also increase the subjectivity and time required for gastroenterologists to perform a reliable manual assessment.

To address this gap, we propose a self-prompting SAM for 2D esophageal OCT image segmentation. By combining SAM's general feature representations with the self-pretraining strategy, the network is designed to identify target esophageal tissues using a single labeled sample. In this study, 'one-shot' refers to the annotation burden (one expert-labeled mask for task-specific fine-tuning), while unlabeled clinical data are used to learn domain-specific priors. Our contributions are summarized as follows:

- We developed a one-shot segmentation framework specifically optimized for esophageal OCT images.
- We designed a self-prompting strategy to effectively adapt SAM for esophageal OCT image segmentation tasks.
- We validated the effectiveness of the proposed method on both a self-collected dataset and a public OCT dataset.

2. Related Works

2.1. Esophageal OCT Image Segmentation Using Deep Learning

Deep learning has significantly advanced esophageal OCT image segmentation, overcoming traditional methods' reliance on hand-crafted features and poor generalization by automatically learning discriminative features from data. UNet and its variants are widely used in this field. Li et al. [15] proposed a UNet-based network optimized for thin esophageal tissues and reported a DSC above 0.85 for epithelial/submucosal segmentation. Yang et al. [9] incorporated channel attention and connectivity modeling into CE-Net and reported a DSC of 0.91 and an mIoU of 0.84. Adversarial learning has also been applied: Wang et al. [16] used GANs for guinea pig esophageal OCT segmentation, reducing errors in low-contrast regions via generator-discriminator adversarial training. Other architectures based on attention mechanism or self-pretraining strategy were also proposed to further improve the segmentation results [10,24,25].

2.2. SAM for Medical Image Segmentation

The SAM, introduced by Kirillov et al. [19], marked a paradigm shift in image segmentation by demonstrating that a model trained on a massive dataset (SA-1B) could achieve robust zero-shot generalization. More recently, Ravi et al. [23] released SAM2, which extends the foundation model architecture to support both images and videos through a streaming memory mechanism. While SAM2 excels at processing temporal sequences, its zero-shot performance in medical imaging still faces challenges due to the substantial domain shift between natural scenes and specialized medical modalities like OCT [26].

Recent medical SAM adaptations differ substantially in how they introduce domain knowledge and prompts. MedSAM uses supervised fine-tuning on more than 1.5 million medical image-mask pairs and retains bounding boxes to identify the segmentation target [21], whereas AutoSAM freezes the SAM image encoder, removes prompt tokens from the mask decoder, and learns a prompt-free task-specific prediction head from labeled medical volumes [20]. Wu et al. [27] instead train a linear pixel-wise classifier on SAM embeddings from a few labeled support images; its coarse prediction is converted into posi-

tional prompts that are returned to SAM. CPC-SAM uses unlabeled images differently: two segmentation branches generate cross-prompts and pseudo-supervision for one another under a semi-supervised segmentation objective [28].

A second branch emphasizes parameter-efficient adaptation. The SAM-Adapter and related medical adapters insert small trainable modules into a largely frozen SAM backbone rather than updating the complete model [29,30]. The Medical SAM Adapter further introduces space–depth transformation and prompt-conditioned adapters for volumetric data [31], while SAMedOCT adapts SAM specifically for retinal OCT segmentation [32]. A systematic study by Gu et al. [33] showed that both parameter-efficient tuning and additional self-supervised training can improve medical SAM adaptation. In that formulation, however, masked reconstruction further pretrains the SAM backbone before task-specific supervised fine-tuning.

The proposed framework therefore addresses a different adaptation setting. It neither generates point or box prompts nor inserts adapters into the SAM backbone. Instead, a separate prompt encoder is pretrained by masked reconstruction using only unlabeled esophageal OCT B-scans; its dense domain-specific feature map is then fused with the general embedding from the SAM image encoder, and only one labeled B-scan is introduced during segmentation training. Compared with supervised automatic-prompt methods, this design decouples OCT representation learning from pixel-level annotation. Compared with semi-supervised methods, the unlabeled images are used in a separate label-free pretraining stage rather than through segmentation pseudo-labels. The resulting inference is fully automatic, while the principal tradeoffs are the additional encoder and dependence on access to target-domain unlabeled images.

3. Methods

3.1. Overview of the Self-Prompting SAM for Esophageal OCT Images

The overall framework of the proposed method is illustrated in Figure 1. It comprises three core components: a SAM encoder, a prompt encoder, and a segmentation decoder. The SAM encoder maps the input esophageal OCT image into a high-dimensional image embedding space, capturing discriminative tissue features [34]. The prompt encoder, implemented as a vision transformer encoder, transforms the input image into task-specific feature representations. Finally, the segmentation decoder fuses the image embeddings from the SAM encoder and the features from the prompt encoder to generate accurate esophageal tissue segmentation results.

The process consists of two stages. In the self-pretraining stage, The prompt encoder is pretrained using unlabeled esophageal OCT images following the masked autoencoder (MAE) strategy [35], enabling it to learn robust, domain-specific feature representations without relying on labeled data. In the segmentation stage, a segmentation network is constructed using the pretrained SAM encoder and the self-pretrained prompt encoder. This network is fine-tuned using only one labeled esophageal OCT image to accomplish the target tissue segmentation task. Detailed implementations of each component are provided in subsequent sections.

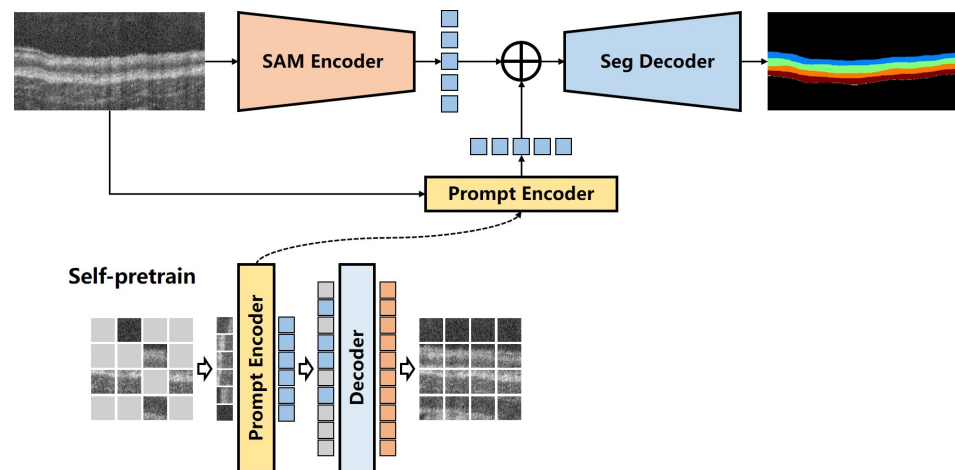


Figure 1. Overall workflow of the proposed self-prompting SAM framework for esophageal OCT images. The framework operates in two stages. Self-pretraining stage: The prompt encoder is pretrained on unlabeled OCT B-scans using a masked autoencoder (MAE) strategy to learn domain-specific tissue textures. Segmentation stage: For a given input image, the pre-trained SAM encoder extracts robust general structural embeddings, while the self-pretrained prompt encoder extracts domain-specific OCT features. These features are fused and passed through the segmentation decoder, which is fine-tuned using a single labeled image (one-shot) to output the final tissue segmentation mask. Different colours are used only to distinguish functional modules, feature embeddings, masked/reconstructed image patches, and segmented tissue classes.

3.2. The Self-Pretrained Prompt Encoder

The prompt encoder is pretrained using the MAE strategy, as illustrated in Figure 1. Specifically, the input esophageal OCT image is first divided into non-overlapping patches. These patches are then randomly categorized into visible and masked groups following a uniform distribution, with a large proportion of patches masked out. This design ensures the pretraining task cannot be easily solved by extrapolating information from adjacent visible patches [35].

During pretraining, the encoder is responsible for modeling latent feature representations of the masked patches, which are subsequently used to predict the original signals in the masked regions. In contrast, the decoder is tasked with reconstructing images from these latent embeddings; it is designed with shallower transformer blocks to reduce the computational cost of the pretraining process [35].

The prompt encoder is trained using mean squared error as the reconstruction loss. This loss is calculated between the reconstructed and original images in the pixel space of the masked patches, consistent with established MAE training protocols [35,36].

3.3. The Segmentation Decoder

The segmentation decoder consists of a series of segmentation blocks, as illustrated in Figure 2. It incorporates a squeeze-and-excited (SE) structure composed of global average pooling (GAP) and two fully connected (FC) layers, which serves as the channel-attention mechanism [37]. Specifically, this SE structure utilizes GAP and FC layers to calculate importance weights, which are then assigned to different feature channels.

This channel-attention design enables the network to focus on feature maps that are critical for the final segmentation task, while suppressing irrelevant or redundant feature channels. Additionally, the “Up” block in the decoder denotes the upsampling operation, which is implemented using transpose convolutional layers with a kernel size of 2×2 .

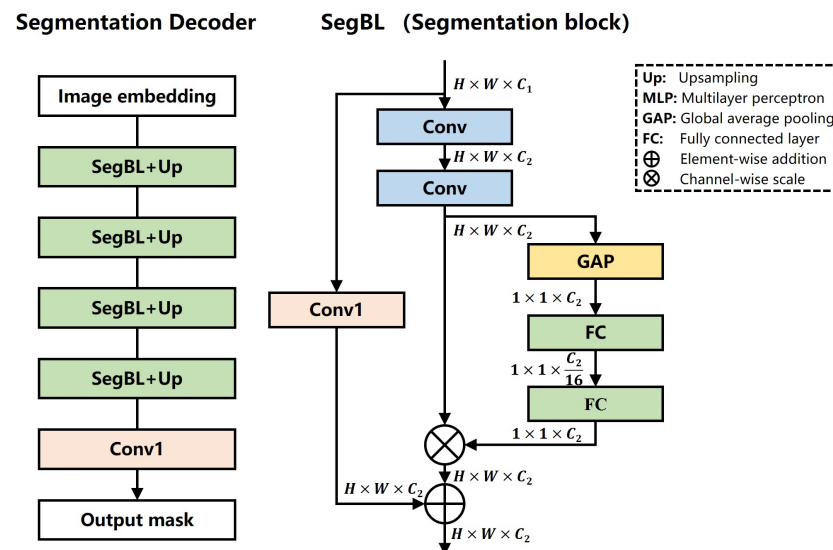


Figure 2. Architecture of the segmentation decoder. The decoder utilizes a SE channel-attention mechanism (comprising Global Average Pooling and Fully Connected layers) to selectively emphasize feature channels relevant to target tissues. The “Up” blocks denote transpose convolutional layers used to progressively upsample the fused embeddings back to the original image resolution. Colours are used to visually distinguish convolutional layers, attention-related operations, upsampling blocks, and output layers.

The overall loss function of the segmentation network is defined as Equation (1), which combines cross-entropy loss and Dice loss to balance classification accuracy and spatial overlap performance.

$$L = L_{CE} + \lambda L_{Dice} \quad (1)$$

In this equation, L_{CE} denotes the cross entropy loss which quantifies the classification accuracy of each pixel. It is formally defined as Equation (2), where N is the pixel number, $g_l(x)$ is the target probability that pixel x belongs to class l with one for the true label and zero entries for the others. $p_l(x)$ is the estimated probability of pixel x belongs to class l .

$$L_{CE} = - \sum_{i=1}^N g_l(x) \log p_l(x) \quad (2)$$

L_{Dice} represents the Dice loss, which evaluates the spatial overlap between the predicted segmentation mask and the ground truth. Its mathematical definition is given in Equation (3):

$$L_{Dice} = 1 - \frac{2 \sum_{i=1}^N p_l(x) g_l(x)}{\sum_{i=1}^N p_l^2(x) + \sum_{i=1}^N g_l^2(x)} \quad (3)$$

In Equation (3), all parameters are defined consistently with those in Equation (2). The weight coefficient λ that balances the contributions of L_{CE} and L_{Dice} is set to 5 in this study.

While esophageal OCT images suffer from low contrast and speckle noise, which often necessitate complex boundary-aware or topology-aware losses, we opted for the CE + Dice combination with a heavily weighted Dice component for two primary reasons. First, the high weight assigned to the Dice loss mitigates the class imbalance caused by thin esophageal tissue layers compared with the background. Second, the pre-trained SAM encoder provides structural and boundary priors that reduce the burden on the loss function to enforce topological consistency. The CE + Dice formulation therefore provides stable convergence in the present one-shot training regime.

4. Experiments

4.1. Dataset

To verify the effectiveness of the proposed algorithm, two datasets were used in the experiments: a self-collected mouse esophageal OCT dataset and a public human esophageal OCT dataset. An overview of these two datasets is presented in Table 1.

Table 1. Dataset characteristics and evaluation protocols.

Characteristic	Dataset #1	Dataset #2
Data source	Self-collected mouse esophageal OCT data	Public human esophageal OCT data [9]
Cohort size	5 mice; 100 B-scans	30 subjects; 784 B-scans
Clinical composition	All healthy	23 healthy subjects and 7 subjects with Barrett's esophagus
Segmentation target	Four tissue classes	Epithelium
Image dimensions (pixels)	512 × 1000	Width: 506–1020; height: 310–648
Cross-validation	5-fold subject-level cross-validation	10-fold subject-level cross-validation
Fold composition	4 subjects for training and 1 subject for testing	27 subjects for training and 3 subjects for testing
One-shot protocol	One randomly selected labeled B-scan in each training fold	One randomly selected labeled B-scan in each training fold

Dataset #1: This dataset is self-collected and consists of 100 esophageal OCT images obtained from five C57BL mice. All images were acquired in vivo using an 800 nm endoscopic OCT system. The animal experiment was approved by the Animal Science Center of Suzhou Institute of Biomedical Engineering and Technology. Annotated labels for the dataset were generated by an experienced grader using ITK-SNAP software (v3.8.0).

Dataset #2: This dataset is a public human esophageal OCT dataset [9], comprising 30 subjects with a total of 784 B-scans. Among these 30 subjects, 6 had non-dysplastic BE, and 1 had BE with low-grade dysplasia. The dataset was independently labeled by three graders. The annotations from the most experienced grader (Grader #1) were adopted as the gold standard to establish a consistent baseline [9]. It is important to note that due to the inherent subjective difficulty of esophageal OCT interpretation, graders were instructed to segment the target tissue only in regions where they were highly confident. Consequently, the manual segmentations often do not span the entire B-scan, and notable inter-observer variability exists among the graders [9].

4.2. Implementation Details

For the one-shot segmentation protocol, the single labeled training image in each fold was selected strictly at random from the available training set. No manual curation based on image quality, structural clarity, or representative layer visibility was performed. To ensure statistical robustness and mitigate the bias of any single random selection, the overall performance was evaluated using cross-validation (5-fold for Dataset #1 and 10-fold for Dataset #2). The final quantitative metrics represent the average across all folds, effectively serving as multiple independent runs with different randomly selected one-shot training samples.

All experiments were implemented using the PyTorch framework (v2.8.0) with computational support from a 40 GB Nvidia A100 GPU. The software environment was configured with CUDA 11.3 and cuDNN v8 to accelerate model training and inference.

It is worth noting that to evaluate the pure feature extraction and segmentation capabilities of the proposed SAM-based framework, no specialized OCT preprocessing (e.g., speckle noise reduction, shadow compensation, or contrast enhancement like CLAHE) or postprocessing (e.g., morphological operations or connected component analysis) steps were applied. The raw OCT B-scans were simply normalized to a range of [0, 1] and fed directly into the network.

For pretraining, simple data augmentation techniques (random flip and random crop) were employed to enhance the generalization ability of the prompt encoder. The AdamW optimizer implemented in PyTorch was used for training, with a learning rate of 0.001, weight decay of 0.05, batch size of 64, and total pretraining epochs set to 400.

For the segmentation network, stochastic gradient descent (SGD) implemented in PyTorch was adopted as the optimizer, with a momentum of 0.9 and weight decay of 0.001 to optimize model parameters. The poly learning rate adjustment strategy [38] was employed to dynamically adjust the learning rate during training. The initial learning rate was set to 0.05, and the power parameter was set to 0.9. Additionally, the mixed-precision training strategy was adopted to reduce memory consumption [39] and improve training efficiency, with the total number of training epochs set to 500.

4.3. Self-Pretraining of the Prompt Encoder

During the self-pretraining process, the mask ratio is set to 75% [35]. Reconstruction results for images selected from the test set are presented in Figure 3, including the original images, masked images, and reconstructed images. The reconstruction results in Figure 3 indicate that the prompt encoder learns structural information from OCT images despite the limited visible context. These pretrained representations are subsequently used to support one-shot segmentation.

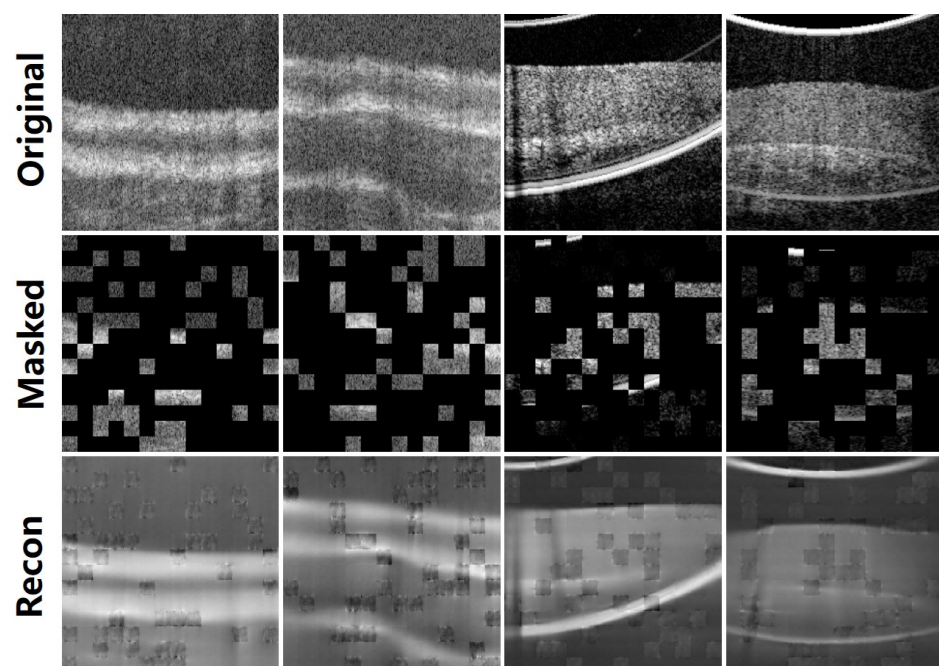


Figure 3. Visualization of the MAE self-pretraining reconstruction on unlabeled esophageal OCT images. From top to bottom: the original input B-scan, the artificially masked image (with a 75% mask ratio), and the reconstructed image generated by the prompt encoder. The high-fidelity reconstruction demonstrates the encoder’s ability to learn and restore underlying OCT tissue structures from extremely sparse visible information.

4.4. Segmentation Results and Comparisons

To evaluate the proposed method, we compared it with several representative networks, including UNet [40], CPFNet [41], UNETR [42], and AutoSAM [20]. UNet adopted a ResNet-50 encoder [43], and UNETR adopted a ViT-B encoder [34]. In the present experiments, both encoders were initialized with their standard pretrained weights, while AutoSAM retained the pretrained SAM image encoder. To comprehensively evaluate the segmentation performance of all compared methods, several metrics were calculated,

including precision, recall, DSC, and accuracy [44]. The detailed quantitative results are presented in Tables 2 and 3.

The quantitative results reveal three observations under the present one-shot protocol: (1) UNet, CPFNet, and UNETR obtained mean DSC values below 70%; (2) AutoSAM achieved higher mean DSC values than these task-specific networks; and (3) the proposed self-prompting SAM achieved the highest mean DSC among the evaluated methods on both datasets, exceeding 85%.

It is also noted that, in Table 3, relatively large standard deviations are observed across all methods on Dataset #2, which can be attributed to two main factors. First, the strictly random selection of the single labeled training image means that in some folds, the network may learn from an image with less representative features, introducing variance into the training process. Second, as previously noted, the omission of standard pre- and post-processing steps leaves the networks highly sensitive to extreme artifacts and severe signal attenuation inherent in raw human clinical data.

Table 2. Quantitative metrics of different networks for dataset #1.

Method	Precision (%)	Recall (%)	DSC (%)	Accuracy (%)
UNet [40]	68.78 ± 14.67	62.20 ± 16.90	67.13 ± 16.40	92.88 ± 2.32
CPFNet [41]	82.01 ± 6.77	73.25 ± 6.02	76.59 ± 7.21	96.58 ± 1.78
UNETR [42]	82.43 ± 5.66	61.18 ± 14.84	69.06 ± 12.45	95.72 ± 2.01
AutoSAM [20]	85.36 ± 5.21	76.90 ± 6.96	80.11 ± 5.48	97.48 ± 1.33
Self-prompting SAM (Ours)	85.96 ± 4.08	85.63 ± 2.82	85.67 ± 3.25	98.07 ± 0.68
Ours (w/o SAM)	82.12 ± 5.90	77.25 ± 6.85	79.42 ± 5.87	97.34 ± 1.35
Ours (w/o self-prompting)	81.85 ± 6.04	76.42 ± 7.12	78.34 ± 6.45	97.05 ± 1.50

The best performance is in bold.

Table 3. Quantitative metrics of different networks for dataset #2.

Method	Precision (%)	Recall (%)	DSC (%)	Accuracy (%)
UNet [40]	60.31 ± 23.02	58.56 ± 20.52	59.41 ± 22.21	76.33 ± 12.61
CPFNet [41]	68.26 ± 20.57	64.33 ± 16.15	64.98 ± 17.68	81.61 ± 9.64
UNETR [42]	70.88 ± 20.32	57.35 ± 10.89	58.87 ± 12.52	80.16 ± 5.58
AutoSAM [20]	84.12 ± 14.88	86.91 ± 9.72	84.09 ± 13.42	90.28 ± 7.79
Self-prompting SAM (Ours)	84.81 ± 14.25	89.31 ± 9.17	85.56 ± 13.08	90.68 ± 7.45
Ours (w/o SAM)	80.50 ± 15.30	82.60 ± 11.20	81.25 ± 14.80	88.50 ± 8.20
Ours (w/o self-prompting)	79.10 ± 16.10	80.40 ± 12.50	79.40 ± 15.50	87.30 ± 8.80

The best performance is in bold.

As shown in Table 4, the proposed framework has 187.6 million parameters and requires 190 ms per B-scan, with a peak inference memory consumption of 8.5 GB. Its computational demand is higher than that of the comparison methods because both the SAM image encoder and the domain-specific prompt encoder are executed during inference. The measured latency corresponds to approximately 5.3 B-scans per second, providing sub-second feedback and demonstrating preliminary feasibility for near-real-time, frame-selected clinical analysis. Further acceleration through lightweight prompt encoders or feature reuse remains necessary for full-rate streaming OCT applications.

Table 4. Computational efficiency of the evaluated segmentation models.

Method	Parameters (M)	Inference Time (ms/Image)	Peak GPU Memory (GB)
UNet (ResNet-50)	32.4	41	2.4
CPFNet	44.3	49	2.8
UNETR (ViT-B)	92.6	128	6.2
AutoSAM (ViT-B)	96.8	93	4.7
Self-prompting SAM (Ours)	187.6	190	8.5

All images were resized to 1024 × 1024 pixels. Measurements were obtained with a batch size of 1 under FP16 inference on an NVIDIA A100 40 GB GPU. Inference time was averaged over 200 runs after 50 warm-up runs.

To assess sensitivity to the particular one-shot sample, we repeated the subject-level cross-validation after independently reselecting the single labeled B-scan in every training fold, while keeping the data partitions, pretrained encoders, network configuration, and optimization hyperparameters unchanged. The repeated selection yielded a mean DSC of $85.26 \pm 4.16\%$ on Dataset #1 and $84.96 \pm 14.11\%$ on Dataset #2, compared with $85.67 \pm 3.25\%$ and $85.56 \pm 13.08\%$, respectively, in the primary evaluation. The closely aligned mean DSC values indicate a limited effect of one-shot sample selection on aggregate performance. The larger dispersion on Dataset #2 remained evident, consistent with the greater heterogeneity of the human cohort. These additional results are summarized in Table 5.

Table 5. Additional quantitative comparisons.

Analysis	Dataset	Setting/Model	DSC (%)
Sample sensitivity	Dataset #1	Primary one-shot selection	85.67 ± 3.25
	Dataset #1	Reselected one-shot sample	85.26 ± 4.16
	Dataset #2	Primary one-shot selection	85.56 ± 13.08
	Dataset #2	Reselected one-shot sample	84.96 ± 14.11
Model comparison	Dataset #2	SAM with identical box prompts	82.14 ± 11.78
	Dataset #2	SAM2 with identical box prompts	81.36 ± 12.19

Dataset #1 denotes the self-collected mouse esophageal OCT dataset, and Dataset #2 denotes the public human esophageal OCT dataset. For the SAM/SAM2 comparison, a paired Wilcoxon signed-rank test based on subject-level mean DSC values showed no statistically significant difference ($p = 0.56$).

An intuitive demonstration of the segmentation results for different networks is shown in Figures 4 and 5. For Figure 4, which presents images from Dataset #1, the segmentation task involves identifying four target tissue types. While the compared networks can recognize major tissue structures, they exhibit various topological errors, such as discontinuous layers (UNet), misclassified tissues (UNETR), and fuzzy boundaries (AutoSAM). We also tested the SAM model with manual prompts; it can identify rough target tissue regions but struggles to achieve detailed tissue recognition. In comparison, the proposed network produces more continuous tissue layers with clearer boundaries in the examples shown. This visual observation is consistent with the quantitative results in Table 2, where the proposed method achieved the highest mean DSC on Dataset #1.

Figure 5 illustrates the segmentation results from Dataset #2, where segmentation masks are overlaid on the original images for clear visualization. Trained with only one labeled image, the compared networks are easily affected by artifacts and adjacent tissues. In contrast, the proposed network generates segmentation results that are close to manual annotations, with more continuous tissue regions. The SAM model with manual prompts is also included in this figure; with a properly positioned boundary box, it can recognize major tissue structures but still suffers from topological errors, such as small holes and blobs. These qualitative differences correspond to the quantitative comparison in Table 3, in which the proposed method achieved the highest mean DSC on Dataset #2 under the present protocol.

Furthermore, to validate our specific choice of the base model in light of recent foundational advancements, we conducted an additional evaluation using the newer SAM2 [23] (in a zero-shot manual prompt setup) on a subset of the data. While SAM2 excels inherently in video and volumetric sequence tracking, its direct application to our 2D B-scans yielded performance fundamentally similar to the original SAM with manual prompts as shown in Figures 4 and 5. This reinforced our hypothesis that without the proposed domain-specific adaptation and self-prompting mechanism, upgrading the base foundation model alone is insufficient to address the unique morphological challenges (e.g., speckle noise, tissue loss) found in esophageal OCT.

To supplement this qualitative observation with a controlled quantitative comparison, we evaluated zero-shot SAM and SAM2 on Dataset #2 using identical anatomy-informed box prompts. Each box covered the full B-scan width, while its upper and lower boundaries were independently expanded outward by a random distance of 0–20 pixels from the annotated epithelium extent at the network input resolution. Under this common prompt protocol, SAM achieved a DSC of $82.14 \pm 11.78\%$, whereas SAM2 achieved $81.36 \pm 12.19\%$. A paired Wilcoxon signed-rank test based on subject-level mean DSC values showed no statistically significant difference between the models ($p = 0.56$). These results indicate that SAM2 did not provide a clear quantitative advantage over the original SAM for this independent 2D single-layer task. This comparison evaluates prompt-conditioned image segmentation and does not assess SAM2's sequence-memory capability. The quantitative SAM/SAM2 comparison is also included in Table 5.

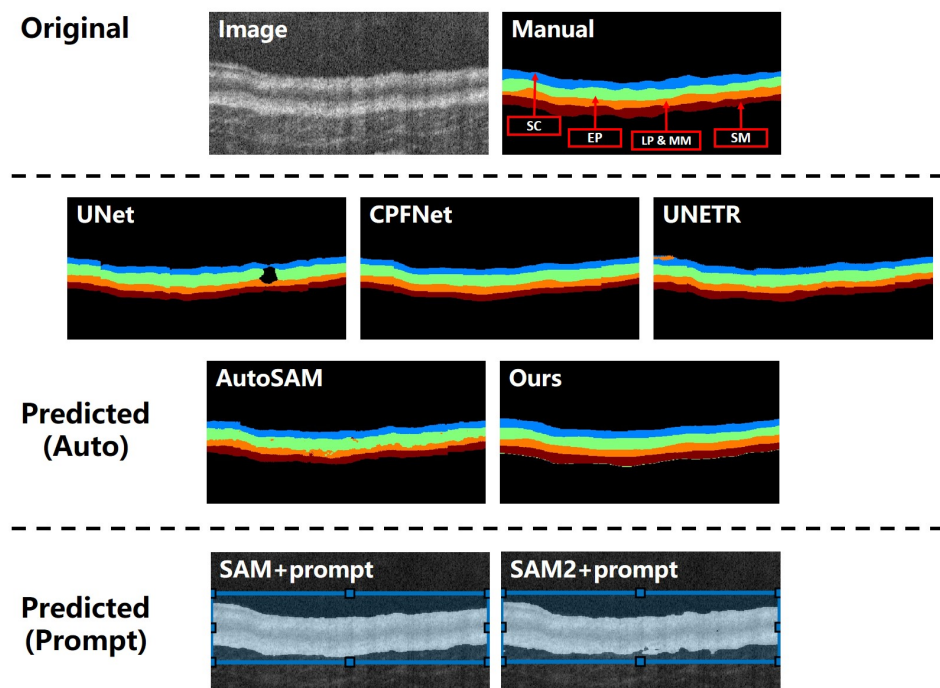


Figure 4. Visualization of segmentation results of different methods on the OCT image from Dataset #1. Target tissues include epithelium stratum corneum (SC), epithelium (EP), lamina propria and muscularis mucosae (LP & MM) and submucosa (SM). The upper row shows the original OCT image and manual annotation; the middle row shows automatic segmentation results; and the lower row shows prompt-based SAM and SAM2 results. Different mask colours correspond to different tissue classes.

The conventional baselines were therefore not trained entirely from random initialization: UNet and UNETR used pretrained ResNet-50 and ViT-B weights, respectively, while AutoSAM used the pretrained SAM encoder. All methods were then adapted to the target segmentation task using the same one-shot annotation budget. Nevertheless, the proposed framework additionally benefits from target-domain unlabeled MAE pretraining, and its total pretraining exposure is not matched by the conventional baselines. Their results should consequently be interpreted as references under a shared one-shot annotation constraint rather than as a fully data-matched comparison. A more balanced assessment is provided by the SAM-based comparisons and the subsequent ablations that separately remove the SAM encoder and self-pretrained prompt encoder.

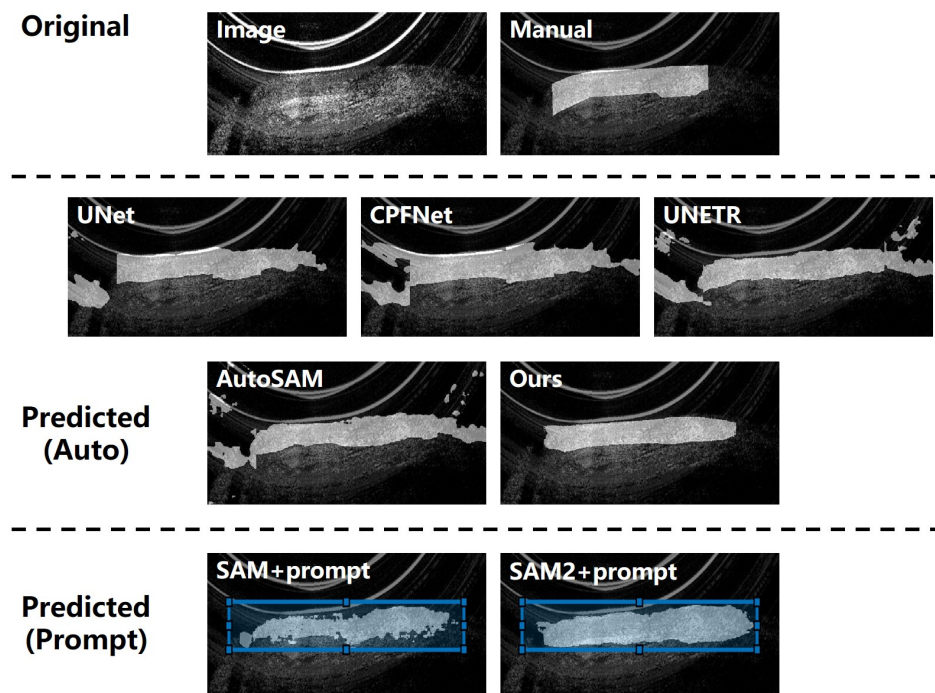


Figure 5. Visualization of segmentation results of different methods on OCT images from Dataset #2. The target tissue is the epithelium, and predicted masks are overlaid on the original B-scans to show boundary continuity and topology. The upper row shows the original OCT image and manual annotation; the middle row shows automatic segmentation results; and the lower row shows prompt-based SAM and SAM2 results. The bright overlay indicates the segmented epithelium region.

4.5. Ablation Study

To systematically evaluate the individual contributions of the core components in the proposed framework, an ablation study was conducted with results listed in Tables 2 and 3. We evaluated two structural variations: (1) removing the pre-trained SAM encoder (w/o SAM); (2) replacing the self-pretrained MAE prompt encoder with a randomly initialized encoder (w/o self-prompting). Removing the SAM encoder reduced the mean DSC on both datasets, indicating its contribution to global structural representation. Replacing the self-pretrained prompt encoder with a randomly initialized encoder reduced the DSC to 78.34% and 79.40%. Without unlabeled domain-specific pretraining, this configuration still retained useful segmentation capacity under the present protocol. However, its comparison with the conventional networks should be interpreted in light of their different architectures and pretraining exposure rather than as a strictly data-matched result. These results indicate that self-pretraining on unlabeled OCT data contributes to the extraction of domain-specific features. The complete network achieved the highest mean DSC among the evaluated configurations under the present protocol, supporting the complementary roles of the SAM encoder and the self-pretrained prompt encoder.

Furthermore, to justify our choice of the CE + Dice loss function, we conducted an additional ablation study comparing our baseline loss against more advanced formulations frequently used in medical imaging, specifically Focal loss + Dice loss ($\gamma = 2.0$, $\alpha = 0.25$) [45], and Focal Tversky loss ($\alpha = 0.3$, $\beta = 0.7$, $\gamma = 4/3$) [46]. In the one-shot setting on Dataset #1, substituting CE + Dice with Focal + Dice yielded a mean DSC of $85.42 \pm 3.41\%$, while the Focal Tversky loss achieved a mean DSC of $85.73 \pm 3.18\%$. Although these advanced losses are theoretically advantageous for handling hard boundaries and extreme class imbalance, their performance was largely comparable to our heavily weighted CE + Dice baseline (DSC of $85.67 \pm 3.25\%$). We attribute this to the fact that the pre-trained SAM encoder already injects robust topological and boundary priors into the

network, resolving most boundary ambiguities. Consequently, CE + Dice was selected for its simplicity, stability, and highly competitive performance without the need to tune additional complex hyper-parameters.

5. Discussion

Automated segmentation of esophageal OCT is constrained by image artifacts and scarce expert annotation. Within the present one-shot protocol, the proposed framework achieved the highest mean DSC among the evaluated methods on both the mouse and human datasets (Tables 2 and 3), while the visual results showed more continuous layers and clearer boundaries (Figures 4 and 5). The strength of this evidence differs between datasets. On Dataset #1, the mean DSC increased from 80.11% for AutoSAM to 85.67%, and the standard deviation decreased from 5.48% to 3.25%. On Dataset #2, the improvement was smaller (84.09% to 85.56%), with similarly large standard deviations (13.42% and 13.08%). The heterogeneous human cohort includes healthy and Barrett's esophagus cases, variable image dimensions, artifacts, and annotations restricted to confident regions. Consequently, the human results support feasibility but should not be interpreted as definitive evidence of clinical benefit.

Published esophageal OCT studies provide important context but are not directly comparable because their species, segmentation targets, image resolutions, annotations, and supervision differ. Li et al. [15] combined three U-Nets trained with 215 labeled guinea-pig B-scans and reported a boundary error of 1.4 μm . Wang et al. [16] used adversarial learning with 700 labeled training/validation B-scans and an independent 400-image test set to improve layer continuity. For human OCT, Yang et al. [9] used connectivity modeling on 784 B-scans from 30 subjects and reported a mean DSC of 0.903. These studies demonstrate the value of ensemble, adversarial, and connectivity constraints under task-specific supervision. By contrast, the principal objective here is not the highest cross-study DSC but reducing task-specific pixel-level annotation to one B-scan.

The framework also occupies a different position from recent medical SAM adaptations. MedSAM prioritizes broad promptable generalization through large-scale supervised medical training [21]; adapter methods reduce the number of updated parameters but still learn task adaptation from labeled targets [30,31]; and few-shot self-prompting methods derive positional prompts from predictions learned with labeled support images [27]. CPC-SAM jointly learns from labeled and unlabeled images through cross-prompted pseudo-supervision [28]. In contrast, our method performs label-free MAE pretraining only on a separate OCT prompt encoder and subsequently fuses its dense features with SAM embeddings for one-shot segmentation. Its specific contribution is therefore the separation of domain representation learning from pixel-level annotation while retaining automatic inference, rather than a general claim of performance advantage over these methods. This specialization also entails additional computation and reliance on unlabeled target-domain OCT images.

Clinically, continuous tissue-layer masks may support measurements of layer thickness and morphology and help clinicians review suspicious regions. Low-annotation adaptation could be useful in hospitals with limited grading resources, after installation of a new OCT console or probe, or when initializing models for different sites in a multi-center study. In each case, routine unlabeled scans can support domain pretraining and a single densely annotated B-scan can initialize the segmentation task. This capability should nevertheless be regarded as low-annotation initialization rather than validated autonomous deployment; prospective evaluation, quality control, and expert oversight remain necessary.

The ablation study also defines an important limitation. Without MAE pretraining, the mean DSC decreased from 85.67% to 78.34% on Dataset #1 and from 85.56% to 79.40%

on Dataset #2, showing that performance depends on access to unlabeled target-domain images. When only a small local archive is available, the source-domain prompt encoder could be incrementally updated through MAE pretraining as routine scans accumulate. With no target-device images, however, cross-device shift remains unaddressed and performance is uncertain. Future studies should therefore report learning curves against the number of unlabeled target scans and evaluate pooled multi-device self-supervised pretraining.

Other limitations arise from the small cohorts (five mice and 30 human subjects), the absence of multi-center and multi-device external validation, and the use of a single grader as the reference for Dataset #2. In a one-shot setting, an incomplete or grader-specific mask can exert disproportionate influence because its bias cannot be averaged across many annotations. The current model also targets major tissue layers rather than subtle dysplasia, combines two encoders with substantial computational cost, and was evaluated without specialized OCT pre- or post-processing. These factors restrict claims of robustness and real-time clinical readiness.

Future validation should use subject-level partitions on independent multi-center and multi-device cohorts, with leave-one-center-out and leave-one-device-out testing followed by a separate assessment of local one-shot adaptation. Reporting per-site and per-device performance will distinguish intrinsic generalization from adaptation gains. Further work will also examine pathological and boundary-aware objectives, lightweight prompt encoders, and OCT-specific noise processing. Finally, sequential or volumetric acquisitions provide the appropriate setting for SAM2 [23]: domain features from the self-pretrained prompt encoder could be fused with per-frame embeddings, while SAM2's memory propagates tissue information across adjacent B-scans. Such studies should evaluate inter-frame boundary continuity and volumetric consistency in addition to per-frame accuracy.

6. Conclusions

This study presents a self-prompting SAM framework for the one-shot segmentation of esophageal OCT images. The framework combines general image features from a pretrained SAM encoder with domain-specific features learned by a prompt encoder from unlabeled OCT data. On the mouse and human datasets, it achieved DSC values above 85% and the highest mean DSC among the evaluated methods under the present protocol, with greater variability observed in the heterogeneous human cohort. These results establish the feasibility of the proposed low-annotation strategy and support its further development for automated esophageal OCT analysis.

Author Contributions: C.W.: Conceptualization, Methodology, Writing—original draft, Writing—review & editing, Investigation, Software. M.G.: Conceptualization, Writing—review & editing, Supervision, Resources and Supervisions. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Shandong Province of China (Grant No. ZR2021QF068, ZR2021QF105) and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 25KJD520014).

Institutional Review Board Statement: The animal study protocol was approved by the Animal Science Center of Suzhou Institute of Biomedical Engineering and Technology (protocol code 2020-A02).

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset generated during and/or analyzed in the current study is available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Huang, D.; Swanson, E.A.; Lin, C.P.; Schuman, J.S.; Stinson, W.G.; Chang, W.; Hee, M.R.; Flotte, T.; Gregory, K.; Puliafito, C.A.; et al. Optical Coherence Tomography. *Science* **1991**, *254*, 1178–1181. [\[CrossRef\]](#) [\[PubMed\]](#)
- Jiang, H.; Zhang, B.; Xiang, B.; Liu, J.; Ma, Z.; Lv, H.; Yu, Y.; Zhao, Y.; Yang, Y.; Luan, J.; et al. Diopter Measurement of Human Eye Based on Dual-Focus Swept Source Optical Coherence Tomography. *Photonics* **2025**, *12*, 856. [\[CrossRef\]](#)
- Palmer, L.D.; Thompson, A.C.; Asrani, S. Diagnosing glaucoma progression with optical coherence tomography. *Curr. Opin. Ophthalmol.* **2025**, *36*, 130–134. [\[PubMed\]](#)
- Mehrjoo, M.; Khamar, P.; Darzi, S.; Verma, S.; Shetty, R.; Arba Mosquera, S. Automated Characterization of Intrastromal Corneal Cuts Induced by Two Femtosecond Laser Systems Using OCT Imaging. *Photonics* **2024**, *11*, 1123. [\[CrossRef\]](#)
- Wax, A.; Zhang, H.; Jelly, E.T.; Price, H.B.; Sun, T.; Chu, K.K.; Cotton, C.C.; Eluri, S.; Goldblum, J.R.; Shaheen, N.J. Prospective Identification of Dysplasia in Barrett's Esophagus with Combined Optical Coherence Tomography and Light Scattering Measurements. *J. Biophotonics* **2026**, *19*, e202500380. [\[PubMed\]](#)
- Sterkenburg, A.; Dijkhuis, T.; Danskin, T.; Vahrmijer, A.; De Boer, J.; Nagengast, W. Assessing the safety and feasibility of optical coherence tomography and near-infrared fluorescence capsule endoscopy in Barrett's Esophagus patients. *Endoscopy* **2025**, *57*, eP299. [\[CrossRef\]](#)
- Zhang, J.L.; Yuan, W.; Liang, W.X.; Yu, S.Y.; Liang, Y.M.; Xu, Z.Y.; Wei, Y.X.; Li, X.D. Automatic and robust segmentation of endoscopic OCT images and optical staining. *Biomed. Opt. Express* **2017**, *8*, 2697–2708. [\[CrossRef\]](#) [\[PubMed\]](#)
- Qi, X.; Pan, Y.S.; Sivak, M.V.; Willis, J.E.; Isenberg, G.; Rollins, A.M. Image analysis for classification of dysplasia in Barrett's esophagus using endoscopic optical coherence tomography. *Biomed. Opt. Express* **2010**, *1*, 825–847. [\[CrossRef\]](#) [\[PubMed\]](#)
- Yang, Z.Y.; Soltanian-Zadeh, S.; Chu, K.K.; Zhang, H.R.; Moussa, L.; Watts, A.E.; Shaheen, N.J.; Wax, A.; Farsiu, S. Connectivity-based deep learning approach for segmentation of the epithelium in in vivo human esophageal OCT images. *Biomed. Opt. Express* **2021**, *12*, 6326–6340. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wang, C.; Gan, M. Wavelet attention network for the segmentation of layer structures on OCT images. *Biomed. Opt. Express* **2022**, *13*, 6167–6181. [\[CrossRef\]](#) [\[PubMed\]](#)
- Jeiouni, P.; Dehzangi, O.; Amireskandari, A.; Rezai, A.; Nasrabadi, N.M. MultiSDGAN: Translation of OCT Images to Superresolved Segmentation Labels Using Multi-Discriminators in Multi-Stages. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 1614–1627. [\[CrossRef\]](#) [\[PubMed\]](#)
- Moradi, M.; Du, X.; Huan, T.; Chen, Y. Feasibility of the soft attention-based models for automatic segmentation of OCT kidney images. *Biomed. Opt. Express* **2022**, *13*, 2728–2738. [\[CrossRef\]](#) [\[PubMed\]](#)
- Moradi, M.; Chen, Y.; Du, X.; Seddon, J.M. Deep ensemble learning for automated non-advanced AMD classification using optimized retinal layer segmentation and SD-OCT scans. *Comput. Biol. Med.* **2023**, *154*, 106512. [\[CrossRef\]](#) [\[PubMed\]](#)
- Gan, M.; Wang, C.; Yang, T.; Yang, N.; Zhang, M.; Yuan, W.; Li, X.D.; Wang, L.R. Robust layer segmentation of esophageal OCT images based on graph search using edge-enhanced weights. *Biomed. Opt. Express* **2018**, *9*, 4481–4495. [\[CrossRef\]](#) [\[PubMed\]](#)
- Li, D.W.; Wu, J.M.; He, Y.F.; Yao, X.W.; Yuan, W.; Chen, D.F.; Park, H.C.; Yu, S.Y.; Prince, J.L.; Li, X.D. Parallel deep neural networks for endoscopic OCT image segmentation. *Biomed. Opt. Express* **2019**, *10*, 1126–1135. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wang, C.; Gan, M.; Zhang, M.; Li, D.Y. Adversarial convolutional network for esophageal tissue segmentation on OCT images. *Biomed. Opt. Express* **2020**, *11*, 3095–3110. [\[CrossRef\]](#) [\[PubMed\]](#)
- Moor, M.; Banerjee, O.; Abad, Z.S.H.; Krumholz, H.M.; Leskovec, J.; Topol, E.J.; Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature* **2023**, *616*, 259–265. [\[CrossRef\]](#) [\[PubMed\]](#)
- Awais, M.; Naseer, M.; Khan, S.; Anwer, R.M.; Cholakkal, H.; Shah, M.; Yang, M.H.; Khan, F.S. Foundation models defining a new era in vision: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 2245–2264. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 3992–4003.
- Hu, X.; Xu, X.; Shi, Y. How to efficiently adapt large segmentation model (sam) to medical images. *arXiv* **2023**, arXiv:2306.13731.
- Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nat. Commun.* **2024**, *15*, 654. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wang, H.; Guo, S.; Ye, J.; Deng, Z.; Cheng, J.; Li, T.; Chen, J.; Su, Y.; Huang, Z.; Shen, Y.; et al. SAM-Med3D: A vision foundation model for general-purpose segmentation on volumetric medical images. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 1024. [\[CrossRef\]](#)
- Ravi, N.; Gabeur, V.; Hu, Y.T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. SAM 2: Segment Anything in Images and Videos. *arXiv* **2024**, arXiv:2408.00714.
- Li, D.; Cheng, Y.; Guo, Y.; Wang, L. Esophageal tissue segmentation on OCT images with hybrid attention network. *Multimed. Tools Appl.* **2024**, *83*, 42609–42628.

25. Wang, C.; Gan, M. Few-shot segmentation for esophageal OCT images based on self-supervised vision transformer. *Int. J. Imaging Syst. Technol.* **2024**, *34*, e23006.
26. Ali, M.; Wu, T.; Hu, H.; Luo, Q.; Xu, D.; Zheng, W.; Jin, N.; Yang, C.; Yao, J. A review of the segment anything model (sam) for medical image analysis: Accomplishments and perspectives. *Comput. Med Imaging Graph.* **2025**, *119*, 102473. [[CrossRef](#)] [[PubMed](#)]
27. Wu, Q.; Zhang, Y.; Elbatel, M. Self-prompting Large Vision Models for Few-Shot Medical Image Segmentation. In *Proceedings of the Domain Adaptation and Representation Transfer; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2024; Volume 14293*, pp. 156–167. [[CrossRef](#)]
28. Miao, J.; Chen, C.; Zhang, K.; Chuai, J.; Li, Q.; Heng, P.A. Cross Prompting Consistency with Segment Anything Model for Semi-supervised Medical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2024; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2024; Volume 15011*, pp. 167–177. [[CrossRef](#)]
29. Chen, T.; Zhu, L.; Ding, C.; Cao, R.; Wang, Y.; Zhang, S.; Li, Z.; Sun, L.; Zang, Y.; Mao, P. SAM-Adapter: Adapting Segment Anything in Underperformed Scenes. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, 2–6 October 2023*; pp. 3359–3367. [[CrossRef](#)]
30. Gao, Y.; Xia, W.; Hu, D.; Zhou, S.K. Segment anything in medical images with adapter. *Sci. Rep.* **2024**, *14*, 12461. [[CrossRef](#)]
31. Wu, J.; Fu, R.; Fang, H.; Liu, Y.; Wang, Z.; Xu, Y.; Jin, Y.; Arbel, T. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. *arXiv* **2023**, arXiv:2304.12620.
32. Fazekas, B.; Morano, J.; Lachinov, D.; Aresta, G.; Bogunovic, H. SAMedOCT: Adapting Segment Anything Model (SAM) for Retinal OCT. *arXiv* **2023**, arXiv:2308.09331.
33. Gu, H.; Dong, H.; Yang, J.; Mazurowski, M.A. How to Build the Best Medical Image Segmentation Algorithm Using Foundation Models: A Comprehensive Empirical Study with Segment Anything Model. *Mach. Learn. Biomed. Imaging* **2025**, *3*, 88–120. [[CrossRef](#)]
34. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021*.
35. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022*; pp. 15979–15988.
36. Bao, H.; Dong, L.; Piao, S.; Wei, F. BEiT: BERT Pre-Training of Image Transformers. In *Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022*.
37. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018*; pp. 7132–7141.
38. Mishra, P.; Sarawadekar, K. Polynomial learning rate policy with warm restart for deep neural network. In *Proceedings of the TENCN 2019—2019 IEEE Region 10 Conference (TENCN), Kochi, India, 17–20 October 2019*; pp. 2087–2092.
39. Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; et al. Mixed Precision Training. In *Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018*.
40. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the 18th International Conference 2015, Munich, Germany, 5–9 October 2015; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 9351*, pp. 234–241.
41. Feng, S.; Zhao, H.; Shi, F.; Cheng, X.; Wang, M.; Ma, Y.; Xiang, D.; Zhu, W.; Chen, X. CPFNet: Context pyramid fusion network for medical image segmentation. *IEEE Trans. Med Imaging* **2020**, *39*, 3008–3018. [[CrossRef](#)] [[PubMed](#)]
42. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. UNETR: Transformers for 3D Medical Image Segmentation. In *Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2022*; pp. 1748–1758.
43. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; pp. 770–778.
44. Müller, D.; Soto-Rey, I.; Kramer, F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res. Notes* **2022**, *15*, 210. [[CrossRef](#)] [[PubMed](#)]

45. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007.
46. Abraham, N.; Khan, N.M. A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 8–11 April 2019; pp. 683–687.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.