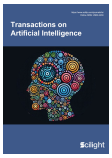
Scilight

Transactions on Artificial Intelligence
https://www.sciltp.com/journals/tai



Article

Check for updates

From Fully-Automatic to Efficient Interaction: A Lightweight Prompting Framework for Jaw Cyst Segmentation

Chao Li [1], Zhe Rao [2,3], Tinglong Tang [2,3], Zhenyang Wang [4], Yalin Wang [5] and Xiaotian Pan [6,*]

[1] School of Information Engineering, Zhijiang College of Zhejiang University of Technology, Shaoxing 312030, China
[2] Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang 443002, China
[3] College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China
[4] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China
[5] School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China
[6] Department of Digital Media Technology, Hangzhou Dianzi University, Hangzhou 310018, China
* Correspondence: xiaotianpan@hdu.edu.cn

How To Cite: Li, C.; Rao, Z.; Tang, T.; et al. From Fully-Automatic to Efficient Interaction: A Lightweight Prompting Framework for Jaw Cyst Segmentation. *Transactions on Artificial Intelligence* **2026**, *2*(1), 231–244. https://doi.org/10.53941/tai.2026.100015

Received: 21 September 2026
Revised: 28 October 2026
Accepted: 2 November 2026
Published: 17 November 2026

Abstract: Faint, interrupted boundaries make jaw cysts difficult to delineate in medical images. Clinicians can help identify the lesion, yet a fully automatic segmentation model has no direct means of using their assessment to guide an individual mask. We developed PromptFormer to accept bounding box prompts through a compact medical segmentation backbone. Its Plug-in Prompt Process Module (PPPM) couples that backbone to the prompt encoder and mask decoder of the Segment Anything Model (SAM); SAM's image encoder is omitted. The backbone is supplemented by a Frequency-domain Global Perception Module (FGPM) for global structure and a Multi-scale Detail Refinement Module (MDRM) for spatial detail. Tests used a clinically collected jaw cyst dataset, BUSI, and ISIC 2017. Across the three datasets, the highest Intersection over Union (IoU) among the compared models was obtained by PromptFormer. Its jaw cyst Dice score was 93.79%. Relative to the SAM-based comparators in the efficiency test, fewer than one seventh as many floating-point operations were needed, and inference was approximately 2.7 to 8.6 times as fast. These computational savings, together with the prompt interface, justify further evaluation on clinical devices with limited resources.

Keywords: jaw cyst; interactive image segmentation; SAM; Fourier transform; multi-scale features

1. Introduction

Visual inspection and palpation form part of the clinical assessment of a jaw cyst. The size, shape, and position of the lesion relative to adjacent anatomy are determined from CT scans or X-rays together with those findings. Jaw cysts are common maxillofacial lesions. Their potential to cause substantial facial deformity makes delineation relevant to diagnosis and treatment planning.

Automatic lesion masks are produced by UNet [1], H2Former [2], and TransUnet [3] for the imaging part of this assessment. Their predictions have no direct input for findings from the clinical examination. Irregular lesion shapes and discontinuous boundaries remain difficult to delineate as well. Providing an interactive input would let the clinician specify the region of interest and supply spatial information for segmentation.

In early interactive networks, the image and prompt entered the encoder together, with the prompt represented as extra channels [4,5]. Learning their correspondence to the anatomical target was left to the encoder. Such fusion may be inadequate with sparse input, allowing surrounding jaw structures to enter the predicted lesion mask where the cyst margin is blurred.

Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

In SAM [6], user inputs are first encoded into prompt embeddings. Cross-attention in the mask decoder relates these embeddings to the image features. The association between a prompt and its corresponding region is learned from a large, varied segmentation dataset. Our interactive component uses this pre-trained prompt encoder and mask decoder to retain those learned associations.

Medical adaptations of SAM often fine-tune or distill a model that retains its image encoder, prompt encoder, and mask decoder. The image encoder accounts for much of the deployment and inference cost. Updating it with a small medical dataset may also alter useful pre-trained representations or lead to overfitting; parameter-efficient adaptation offers alternatives to updating all parameters [7,8]. Retaining the full SAM architecture also leaves limited scope to reuse established medical segmentation backbones such as UNet [1] and H2Former [2].

PromptFormer connects SAM's prompt-processing components to H2Former [2]. The Plug-in Prompt Process Module (PPPM) combines features from this medical segmentation backbone with user prompts through SAM's prompt encoder and mask decoder, replacing SAM's image encoder. We also introduce a Frequency-domain Global Perception Module (FGPM) and a Multi-scale Detail Refinement Module (MDRM) to address low contrast and faint boundaries. FGPM modifies global frequency information, whereas MDRM aggregates local spatial features at different scales. The resulting features support interactive mask prediction within the H2Former architecture.

The contributions of this study are as follows:

- We introduce PPPM, a trainable interface that reuses SAM's prompt encoder and mask decoder in an automatic segmentation backbone. Convolving prompt-derived features with backbone features directs the model toward the region specified by the user.
- FGPM and MDRM address the weak edges found in low-to-medium contrast jaw cyst images. Frequency-domain processing supplies global structural information; aggregation across encoder levels and spatial scales supplies the local detail needed to refine a boundary.
- We compare PromptFormer with 17 other methods across a clinical jaw cyst dataset, and two public benchmarks (BUSI, and ISIC). PromptFormer achieves a jaw cyst IoU that is 1.79 percentage points higher than SAMUS, the strongest prompted comparator. We also report the memory use, computational cost, and inference speed of the models.

2. Related Work

2.1. Medical Image Segmentation

Neural segmentation methods learn image features to localize anatomical structures and pathological regions for computer-aided diagnosis. U-Net [1] retains high-resolution spatial information alongside deeper representations, with later methods extending the use of multiple scales [9–11]. Active contours, thresholding, and clustering are conventional alternatives. Medical image analysis has also been studied under other learning frameworks [12].

Long-range relationships are useful, but a small lesion or a faint boundary still requires detailed representation at several scales. Standard Vision Transformers model distant image regions while remaining computationally expensive and limited in fine-scale representation [13].

Two dominant paradigms have arisen to address the above challenges.

SwinUnet [14] and the CNN-Transformer hybrid H2Former [2] are dedicated segmentation architectures. Frequency processing and detailed feature extraction have also been examined in SF-UNet [15], IRPruneDeXt [16], and TBUnet [17].

Parameter-Efficient Fine-Tuning (PEFT) adapts a pre-trained network with a small set of task-specific parameters while preserving its representations. In the Dual-Kernel Adapter [18], large- and small-kernel depthwise convolutions operate in parallel to widen the receptive field without losing local detail when data are limited. Complex contours remain difficult for automatic methods, which generally provide no direct input through which a clinician's assessment can guide an individual prediction.

2.2. Interactive Segmentation

Early interactive methods used low-level color and texture cues [19]; learned deep features were introduced in later work [4]. RITM [4] and SimpleClick [20] introduce user input at the image encoder, so the same encoder bears much of the work of both visual representation and prompt interpretation. Its computational cost also affects response time. Separate processing of prompts, followed by interaction with image features, is an alternative used in SAM and its medical adaptations [6,21].

The division of work in SAM is useful for our purpose. Prompt embeddings are produced separately from image encoding, then combined with the visual representation by a lightweight mask decoder. Reusing those

components does not require retaining the original image backbone. This leaves room to choose a backbone suited to the memory and response-time limits of a clinical device.

Medical adaptations include global fine-tuning with MedSAM [21], PEFT methods such as MSA and SAMed [22–26], and architecture changes such as SAMUS [27]. These methods retain SAM's large image encoder. PEFT reduces the number of parameters updated during training, but does not necessarily reduce inference cost, and additional adapter branches can increase latency. We instead combine SAM's prompt encoder and mask decoder with a conventional medical segmentation backbone to predict masks from bounding box prompts.

3. Methods

Figure 1 shows the PromptFormer architecture. It retains the hierarchical U-shaped structure of H2Former and adds PPPM, which uses SAM's prompt encoder and mask decoder to process user inputs.

Two further components operate on the backbone features. FGPM introduces frequency-domain processing into the spatial encoder to model global structure alongside local image content. MDRM draws on several encoder levels and spatial scales to refine weak edge responses. Both feed the decoder that predicts the mask, with the aim of resolving lesion margins that remain difficult to distinguish in the image.

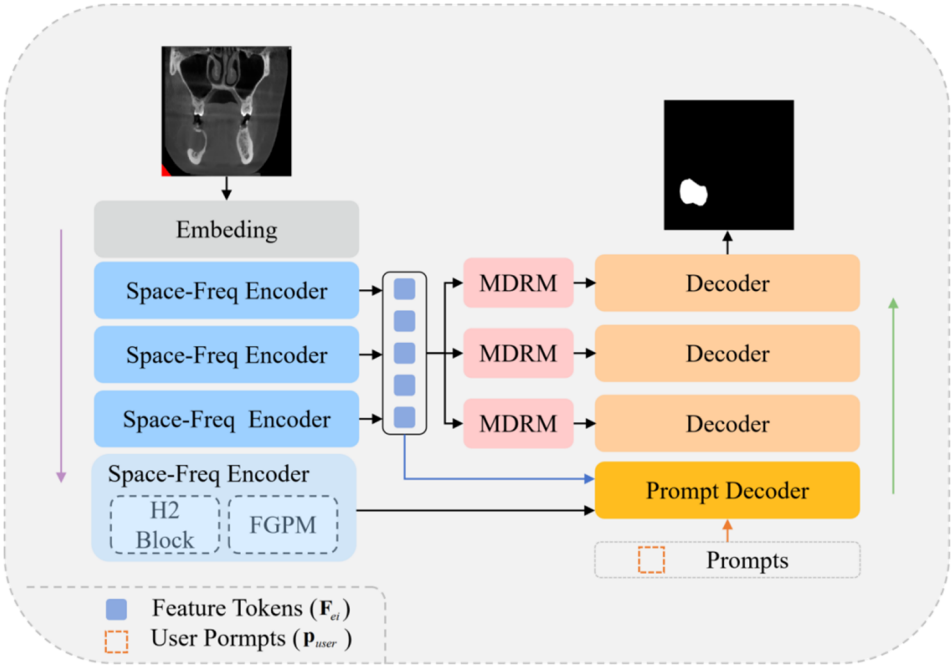


Figure 1. PromptFormer architecture. The Space-Freq Encoder is the Spatial-Frequency Encoder described in Section 3.2; the Prompt Decoder is PPPM (Section 3.1); MDRM is described in Section 3.3.

3.1. Prompt Decoder

PPPM converts sparse prompts into dense features associated with the target region. Figure 2 outlines its operation: extract a prompt descriptor, generate dynamic convolution kernels, and use those kernels to modulate the image features.

We use SAM's pre-trained components to relate the prompt to the target. The external prompts $\boldsymbol{p}_{user}$ and hierarchical image features aggregated by a Feature Pyramid Network (FPN) are supplied to SAM's prompt encoder and mask decoder.

From the decoder output, we select the embedding of the first mask token, which is normally used to generate the primary mask. This embedding combines information about the prompt and target region. We use it as the global prompt descriptor $\boldsymbol{p} \in \mathbb{R}^n$:

$$\boldsymbol{p} = SAM_{block}(FPN(\boldsymbol{F}_e, \boldsymbol{F}_{e2}, \boldsymbol{F}_{e3}), \boldsymbol{p}_{user})[0] \tag{1}$$

where $SAM_{block}(\bullet)$ denotes the SAM components and [0] selects the first mask token. The final encoder-layer output is denoted by $\boldsymbol{F}_e$, and $\boldsymbol{F}_{ei}$ is the output of encoder level i, counted from top to bottom.

A linear projection converts the descriptor $\boldsymbol{p}$ into dynamic convolution kernels $\boldsymbol{K}$:

$$\boldsymbol{K} = \text{Reshape}(\boldsymbol{W}\boldsymbol{p} + \boldsymbol{b}) \tag{2}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are learnable parameters. The projected output is reshaped to $\boldsymbol{K} \in \mathbb{R}^{1\times1\times n\times m}$.

We convolve the kernels $\boldsymbol{K}$ with the encoder features $\boldsymbol{F}_e$ to obtain the target-activated feature map $\boldsymbol{F}_p$. A residual connection and final linear projection then produce the prompt-enhanced features $\boldsymbol{F}_d$:

$$\boldsymbol{F}_p = \boldsymbol{K} * \mathrm{PConv}(\boldsymbol{F}_e) \tag{3}$$

where $\mathrm{PConv}(\bullet)$ denotes convolution with a kernel size of 1.

$$\boldsymbol{F}_d = (\boldsymbol{W}'\boldsymbol{F}_p + \boldsymbol{b}') + \boldsymbol{F}_e \tag{4}$$

Here, $\mathbf{W}' \in \mathbb{R}^{m\times e}$ and $\mathbf{b}' \in \mathbb{R}^{e}$ are the weight and bias parameters, respectively.

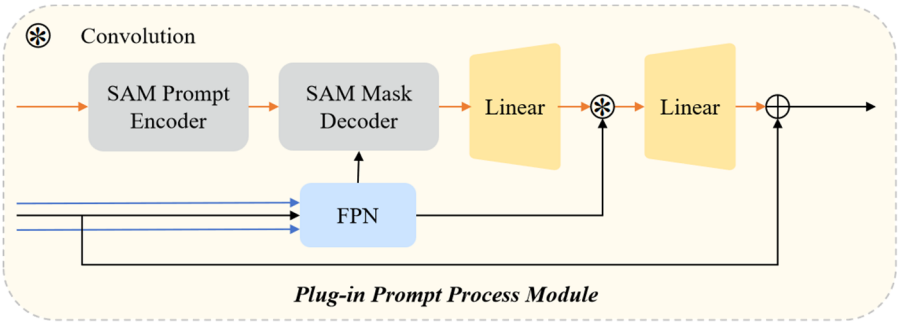


Figure 2. PPPM architecture, with multi-level encoder features and user prompts as inputs.

3.2. Spatial-Frequency Encoder

The Spatial-Frequency Encoder in Figure 3 consists of the H2Former encoder with FGPM added. Local convolutional features are supplemented with global information by transforming the feature map into the frequency domain. Amplitude specifies the strength of its frequency components, and phase their spatial relationships. FGPM implicitly modifies both to refine the structural representation of faint boundaries.

We apply the Fast Fourier Transform (FFT) to the spatial feature $\boldsymbol{F}_e(\boldsymbol{X}) \in \mathbb{R}^{m\times n}$:

$$\boldsymbol{F}_e(\omega) = \mathcal{F}(\boldsymbol{F}_e(\boldsymbol{X})) = \mathrm{Re}(\omega) + i \cdot \mathrm{Im}(\omega) \tag{5}$$

The spectrum can be described either by real and imaginary components or by amplitude $|\boldsymbol{F}_e(\omega)|$ and phase $\angle\boldsymbol{F}_e(\omega)$. FGPM uses the former, filtering each component with its own depthwise convolution (DW). Explicit polar-coordinate operations such as $\arctan$ are thus avoided during filtering, which can ease gradient-based optimization. The learnable filter takes the form:

$$\boldsymbol{F}'_e(\omega) = \mathrm{DW}(\mathrm{Re}(\omega)) + i \cdot \mathrm{DW}(\mathrm{Im}(\omega)) \tag{6}$$

The filtered spectrum is returned to the spatial domain by inverse FFT, giving $\boldsymbol{F}'_e(\boldsymbol{X}) = \mathcal{F}^{-1}(\boldsymbol{F}'_e(\omega))$.

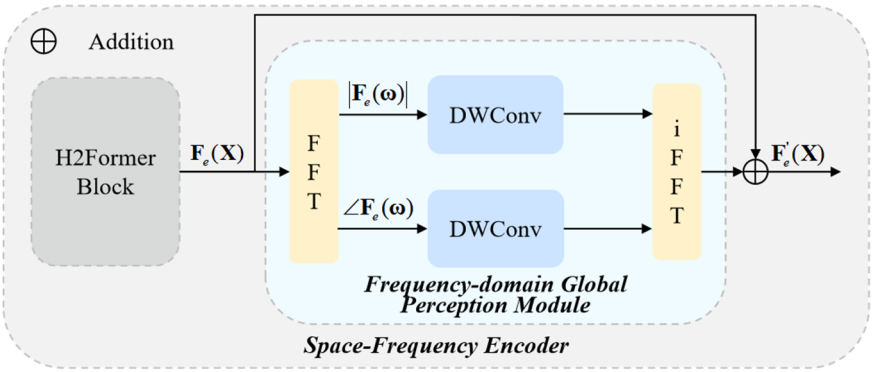


Figure 3. Spatial-Frequency Encoder. FGPM modifies the real and imaginary frequency components, thereby modulating amplitude and phase implicitly.

3.3. Multi-Scale Detail Refinement Module

High-resolution skip features carry spatial detail but relatively little semantic context. For a cyst with an ambiguous margin or variable size, concatenating these features may not identify the detail that belongs to the target. MDRM therefore includes deeper encoder representations to supply semantic context during refinement.

MDRM aggregates in two directions (Figure 4). Across encoder levels, it combines $\boldsymbol{F}_{e1}$, $\boldsymbol{F}_{e2}$, and $\boldsymbol{F}_{e3}$, retaining shallow detail and drawing on deeper semantic information for localization. Within the spatial representation, parallel convolutions gather context from neighborhoods of different sizes. Both cross-level and spatial-scale information thus enter the local refinement.

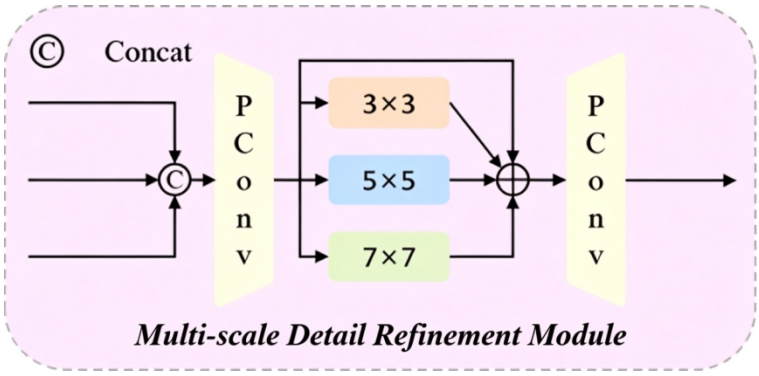


Figure 4. MDRM aggregates features across encoder levels and processes them with parallel kernels at multiple spatial scales.

Concatenating the encoder outputs gives the representation $\boldsymbol{F}_c$:

$$\boldsymbol{F}_c = \text{Concat}(\boldsymbol{F}_{e1}, \boldsymbol{F}_{e2}, \boldsymbol{F}_{e3}) \tag{7}$$

A point-wise convolution $\boldsymbol{F}_c$ reduces the channels of the concatenated representation PConv, yielding $\boldsymbol{F}_{cd}$. Parallel convolutions with 3 × 3, 5 × 5, and 7 × 7 kernels extract context at three spatial scales. Their outputs are added to the base feature $\boldsymbol{F}_{cd}$ through a residual connection, then a final point-wise convolution yields $\boldsymbol{F}'_{cd}$:

$$\boldsymbol{F}_{cd} = \text{PConv}(\boldsymbol{F}_c) \tag{8}$$

$$\boldsymbol{F}'_{cd} = \text{PConv}\left(\boldsymbol{F}_{cd} + \sum_{k\in\{3,5,7\}} C_{k\times k}(\boldsymbol{F}_{cd})\right) \tag{9}$$

where $C_{k\times k}(\bullet)$ denotes convolution and k specifies its kernel size.

4. Experiment

4.1. Datasets

The clinical images were obtained at Quzhou People's Hospital in a retrospective study approved by its Ethics Committee. There were 648 patients with pathologically confirmed jaw cysts and 2592 cone-beam computed tomography (CBCT) images. Cases with severe motion artifacts, extreme blurring, or incomplete lesion boundaries were excluded. We standardized the slices to 256 × 256 pixels and re-annotated them according to [28]. Two senior stomatologists, each with more than five years of clinical experience, drew the contours independently. A chief physician resolved disagreements and finalized the masks. The training, validation, and test sets contained 1558, 517, and 517 images, respectively.

For broader evaluation, we conducted domain-specific experiments independently on two public datasets:

- For the breast ultrasound experiments, BUSI [29] supplied 780 images of benign tumors, malignant tumors, and normal tissue. The 7:1:2 division produced 546 training images, 78 validation images, and 156 test images.
- For dermoscopic skin lesion segmentation, we used ISIC 2017 [30] (https://challenge.isic-archive.com/data/#2017, accessed on 15 May 2026). Training used 2000 images; validation and testing used 150 and 750 images, respectively.

The training and evaluation procedure was repeated independently on each dataset. These experiments compare the architecture across lesion types and imaging modalities; they do not test cross-dataset transfer.

4.2. Experimental Setup

We trained with batches of 8 images on an NVIDIA GeForce RTX 3090 GPU with 24 GB of VRAM. All inputs were 256 × 256 pixels, and box prompts were generated according to the MedSAM protocol [21]. The objective combined Dice loss and binary cross-entropy (BCE), with a weight of 0.5 for each term. H2Former and

the added modules were initialized without pre-training. Only the prompt-processing components in PPPM used pre-trained SAM weights [6]. With Adam, the base learning rate was 0.0001; the pre-trained components used 0.00001, following [31], to limit changes to their learned representations (see Appendix A).

4.3. Evaluation Metrics

Segmentation was evaluated by IoU, Dice, and Recall. The notation in the following equations counts pixels: TP denotes target pixels correctly included in the mask, FP denotes background pixels included by mistake, and FN denotes target pixels left out.

IoU is the intersection of the predicted and reference masks divided by their union. False-positive and false-negative pixels both reduce this value:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{10}$$

Dice normalizes twice the intersection by the total number of foreground pixels in the two masks:

$$\mathrm{Dice} = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{11}$$

Recall gives the proportion of reference target pixels recovered in the prediction. It does not penalize FP, so IoU and Dice are needed alongside it to account for erroneous foreground pixels:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{12}$$

5. Results

5.1. Comparative Experiments

5.1.1. Quantitative Comparison

We compared automatic and prompted segmentation on the jaw cyst images (Table 1). The Transformer, CNN-Transformer, and CNN-KAN automatic methods comprise SwinUnet [14], DS-TransUnet [32], TransUnet [3], H2Former [2], Trans-SAM [33], SAM2-UNet [34], and UKAN [9]. For prompted segmentation, we used SAM [6], MedSAM [21], SAMUS [27], and the non-SAM model SimpleClick [20].

Table 1 compares complete models and does not isolate the causes of their performance differences. SimpleClick and some SAM-based variants fall below several automatic methods. Input-level fusion may leave the association between a prompt and complex anatomy inadequately represented in SimpleClick. For SAM-based variants, natural-image features may remain poorly matched to the medical task when little adaptation data are available.

The best Dice, IoU, and Recall in Table 1 are 93.79%, 88.67%, and 93.89%, respectively, all obtained by PromptFormer. This configuration combines prompt-based localization with an image backbone trained from scratch on the medical images.

Table 1. Comparison on cyst dataset.

Method		Dice	IoU	Recall
Automatic	UKAN [9]	90.11	84.10	91.21
	TransUnet [3]	90.63	84.48	89.91
	SwinUnet [14]	89.68	82.68	90.53
	DS-TransUnet [32]	92.83	87.15	93.42
	H2Former [2]	92.39	86.44	93.69
	Trans-SAM [33]	77.59	69.16	77.67
	SAM2-UNet [34]	87.37	81.54	86.80
Prompted	SimpleClick [20]	90.32	83.75	88.10
	SAM [6]	90.54	83.22	91.85
	MedSAM [21]	91.87	85.80	91.45
	SAMUS [27]	92.83	86.88	93.09
	Ours	93.79	88.67	93.89

5.1.2. Qualitative Comparison

Figure 5 compares the contours at blurred cyst margins. Some methods produce fairly regular outlines in these examples, whereas PromptFormer follows more of the irregular lesion structure. This gives a view of contour shape in addition to the aggregate overlap scores.

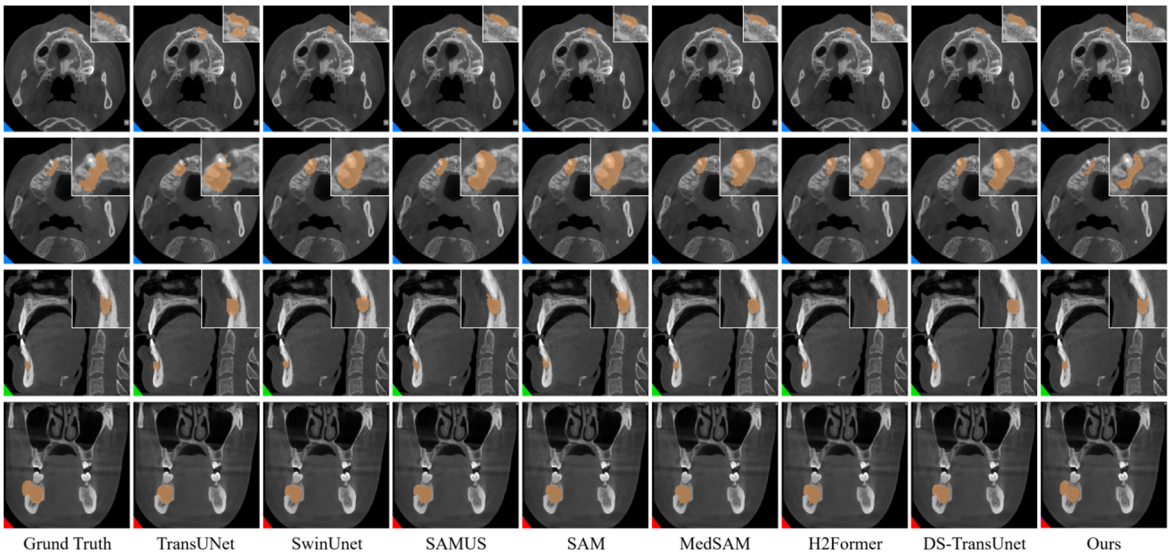


Figure 5. Visualization of comparative experiments on jaw cyst.

We tested the modules separately to distinguish prompt-based localization from the contribution of spatial-frequency and multi-scale features. The latter provide the information about lesion shape and weak boundaries used alongside the prompted location.

5.1.3. Ablation Study

Table 2 reports the jaw cyst configurations. H2Former is the baseline, to which PPPM, FGPM, and MDRM are added separately and together.

PPPM gives the largest gain when modules are tested individually. Baseline Dice and IoU are 92.39% and 86.44%; with PPPM, they reach 93.07% and 87.57%. This configuration introduces the prompt's spatial prior through PPPM.

FGPM alone raises Dice from 92.39% to 92.94% and IoU from 86.44% to 87.49%. With PPPM present, adding FGPM changes Dice from 93.07% to 93.58% and IoU from 87.57% to 88.41%. Its frequency information improves both overlap measures in each setting, consistent with the intended representation of faint and interrupted edges.

Without PPPM, MDRM gives 91.47% Dice in experiment 4, below the 92.39% baseline. Adding FGPM recovers part of the overlap loss: experiment 6 reaches 92.78% Dice, although FGPM alone in experiment 3 gives 92.94%. The effect changes with prompt guidance. MDRM then raises Dice from 93.07% to 93.33% in experiment 7 and from 93.58% to 93.79% in experiment 8. Uncertain localization may allow redundant background features into multi-level aggregation, while the prompt helps associate details with the cyst. Figure 6 shows corresponding masks.

Table 2. Ablation experiments on Cyst.

Method			Cyst		
PPPM	**FGPM**	**MDRM**	**Dice↑**	**IoU↑**	**Recall↑**
×	×	×	92.39	86.44	93.69
√	×	×	93.07	87.57	92.90
×	√	×	92.94	87.49	92.63
×	×	√	91.47	85.86	92.01
√	√	×	93.58	88.41	93.56
×	√	√	92.78	87.24	88.53
√	×	√	93.33	87.86	93.26
√	√	√	93.79	88.67	93.89

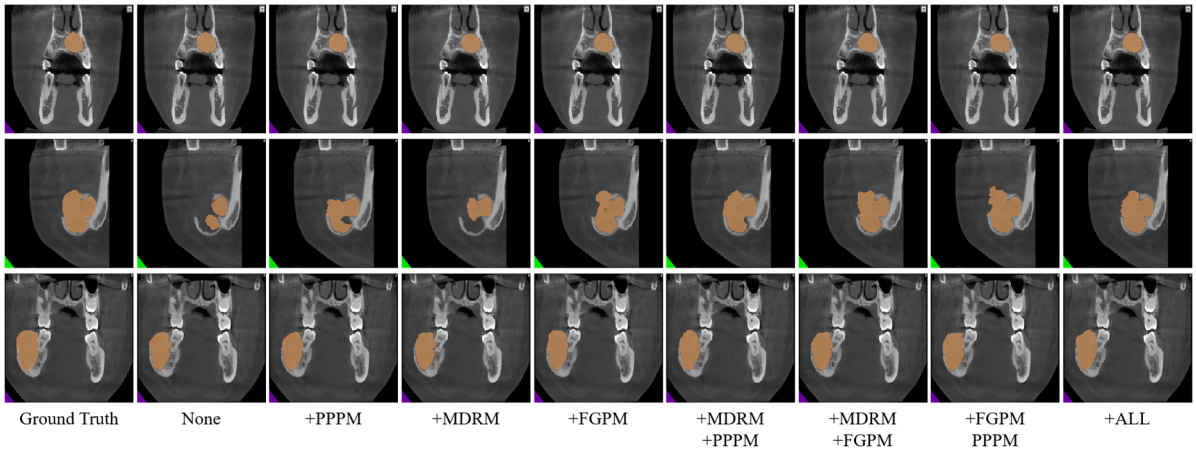


Figure 6. Visualization of Ablation experiments on jaw cyst.

The masks in Figure 6 place the PPPM response in the indicated region. More faint or broken contour segments are followed when FGPM is present. With PPPM, MDRM adds local detail and leaves fewer erroneous regions than the less effective configurations shown.

5.2. Additional Experiments

5.2.1. Comparison and Ablation on BUSI

Breast ultrasound images in BUSI provide a test with artifacts, noise, and low contrast that can obscure lesion margins. In Table 3, we compare the architecture with CNN, Transformer, and hybrid automatic models and with prompted SAM-based methods.

Both SAMUS and PromptFormer score above the automatic models on BUSI. The Dice, IoU, and Recall values for PromptFormer are 90.93%, 83.94%, and 95.42%; those for SAMUS are 90.54%, 83.13%, and 91.78%. The usefulness of a localization prompt in noisy ultrasound is a possible reason for this ordering. Its contribution, together with the effects of FGPM and MDRM, is examined in Table 4.

Table 3. Quantitative comparison on the BUSI dataset.

Method		Dice↑	IoU↑	Recall↑
Automatic	Unet [1]	78.11	69.60	83.16
	CPFNET [35]	80.56	72.28	79.74
	CA-NET [36]	81.88	73.49	84.33
	CE-NET [37]	81.60	73.62	81.72
	AAU-Net [38]	80.81	72.69	82.77
	SwinUnet [14]	67.23	55.58	78.94
	SETR [39]	68.22	57.68	77.41
	MISSFormer [40]	78.43	69.73	83.24
	TransUnet [3]	82.22	74.77	83.19
	TransFuse [41]	73.52	64.28	78.61
	FAT-Net [42]	82.16	74.26	82.13
	H2Former [2]	81.48	73.34	85.16
	Trans-SAM [33]	79.42	70.93	80.42
	SAM2-UNet [34]	82.70	74.86	84.56
Prompted	MedSAM [21]	77.75	66.57	86.59
	MSA [22]	81.66	72.42	89.21
	SAMUS [27]	90.54	83.13	91.78
	Ours	90.93	83.94	95.42

On BUSI, the three-module combinations were evaluated again using the ablation design in Table 4.

PPPM alone raises BUSI Dice from 81.48% to 86.82%, a 5.34-percentage-point gain. Both the jaw cyst and breast ultrasound experiments benefit from prompt guidance, with the larger improvement occurring on BUSI.

FGPM changes Dice only from 81.48% to 81.69% when added without PPPM (Table 4). In the prompted configuration, the change is from 86.82% to 90.87%, or 4.05 percentage points. Speckle may interfere with the

weak lesion contour in an unguided frequency representation. A localization prompt could help focus the refinement on that contour, accounting for the difference between these configurations.

The 82.61% Dice obtained with MDRM alone is above the BUSI baseline. Yet adding it to PPPM without FGPM lowers Dice from 86.82% to 85.44%. The full three-module combination has the highest Dice, IoU, and Recall. Larger breast lesions may supply more spatial context for multi-scale aggregation than smaller jaw cysts; this hypothesis would have to be checked by analyzing performance against lesion size.

Table 4. Ablation study on the BUSI dataset.

Method			BUSI		
PPPM	**FGPM**	**MDRM**	**Dice↑**	**IoU↑**	**Recall↑**
×	×	×	81.48	73.34	85.16
√	×	×	86.82	79.28	88.08
×	√	×	81.69	73.57	84.32
×	×	√	82.61	74.39	84.19
√	√	×	90.87	83.93	90.80
×	√	√	82.64	74.60	85.72
√	×	√	85.44	77.62	87.64
√	√	√	90.93	83.94	95.42

5.2.2. Comparison on ISIC

We tested both automatic and prompted models on the ISIC 2017 dermoscopy dataset; their results appear in Table 5.

Table 5. Quantitative comparison on the ISIC2017 dataset.

ISIC		Dice↑	IoU↑	Recall↑
Automatic	TransUnet [3]	84.32	75.56	83.20
	SwinUnet [14]	85.12	76.73	84.54
	DS-TransUnet [32]	85.13	77.32	84.74
	H2Former [2]	84.83	76.36	80.56
Prompted	SAM [6]	83.85	74.87	82.47
	MedSAM [21]	84.15	76.12	83.76
	SAMUS [27]	90.57	83.43	91.01
	Ours	90.49	83.57	92.25

For ISIC, SAMUS records 90.57% Dice and PromptFormer 90.49%, a difference of 0.08 percentage points. The ranking changes for IoU: PromptFormer is highest at 83.57%. It also has the highest Recall, 92.25% (Table 5). Thus, its advantage on this dataset covers IoU and Recall, while SAMUS has the better Dice score.

5.2.3. Comparison of Schemes for the Frequency-domain Global Perception Module

The frequency variants in Table 6 compare filtering in real-imaginary and amplitude-phase coordinates. Real+Img gives the highest Dice, IoU, and Recall, ahead of amplitude alone, phase alone, and their joint polar representation.

For amplitude alone, Dice falls from the baseline 93.33% to 93.14%. Increasing spectral strength may amplify noise along with the weak structural response of a cyst boundary. Phase alone instead reaches 93.36%, only slightly above the baseline. Joint amplitude-phase filtering gives 93.22%, so using both polar components does not recover the baseline performance.

Table 6. FGPM processing schemes. Amp denotes amplitude, Pha phase, Real the real component, and Img the imaginary component.

Method	Dice↑	IoU↑	Recall↑
Baseline	93.33	87.86	93.26
Amplitude	93.14	87.84	92.98
Phase	93.36	88.04	93.70
Amp+Pha	93.22	87.82	93.17
Real+Img	93.79	88.67	93.89

The real-imaginary representation retains the spectral information without explicitly transforming to amplitude and phase during filtering. Gradient propagation may be more difficult through nonlinear polar-coordinate operations, offering a possible explanation for the lower polar-variant scores. Final segmentation scores cannot establish that optimization mechanism by themselves. They do, however, favor Cartesian processing for the present implementation.

5.2.4. Comparison Across Different Segmentation Backbones Integrated with PPPM

To examine the effect of different feature representations, we integrated PPPM with five backbones (Table 7). They comprise U-Net, a CNN; SwinUnet, a Transformer; the CNN-Transformer hybrids TransUnet and H2Former; and UKAN, a CNN-KAN hybrid.

Table 7. Impact of integrating the PPPM module into various segmentation backbones.

Backbone	Backbone Only			Integrated PPPM		
	Dice↑	IoU↑	Recall↑	Dice↑	IoU↑	Recall↑
U-Net [1]	89.67	83.50	91.50	91.66	85.65	90.83
SwinUnet [14]	89.68	82.68	90.53	92.19	85.87	93.00
UKAN [9]	90.11	84.10	91.21	90.50	84.60	90.35
TransUnet [3]	91.89	84.99	92.15	93.55	88.12	92.77
H2Former [2]	92.39	86.44	93.69	93.07	87.57	92.90

In Table 7, every backbone gains in Dice and IoU with PPPM. SwinUnet has the largest IoU gain, 3.19 percentage points, followed by TransUnet at 3.13 points. TransUnet reaches the highest integrated Dice, 93.55%. For UKAN, Dice changes from 90.11% to 90.50% and IoU from 84.10% to 84.60%, gains of 0.39 and 0.50 points. A mismatch between KAN representations and prompt features may explain this smaller response, although it was not isolated experimentally. U-Net, UKAN, and H2Former have lower Recall after integration despite their higher Dice and IoU.

5.2.5. Computational Resource Comparison with Prompting Methods

The resource comparison in Table 8 includes SimpleClick and the SAM-based MSA, SAMUS, and MedSAM alongside PromptFormer. The measured quantities are GPU memory during inference (MEM, GB), computational cost (Comp, GFLOPs), inference speed (Speed, FPS), and trainable parameters (Params, millions).

PromptFormer uses 29.84 GFLOPs per prediction and achieves 36.29 FPS, with 1.72 GB of inference memory and 48.62 million trainable parameters. Compared with every method in Table 8, it has lower memory use, computation, and parameter count, as well as higher inference speed. Relative to the SAM-based methods, the computational ratio is approximately 1/27.0 to 1/9.8 and the inference-speed ratio is 2.7 to 8.6. Under these test conditions, replacing the original SAM image encoder reduces the resources needed for interactive prediction.

Table 8. Comparison of computational efficiency and model complexity.

Method	MEM↓	Params↓	Comp↓	Speed↑
SimpleClick [20]	2.03	322.18	532.87	10.52
MSA [22]	13.20	97.56	805.80	4.23
SAMUS [27]	4.30	130.10	291.76	13.36
MedSAM [21]	4.51	90.49	743.98	4.54
Ours	1.72	48.62	29.84	36.29

6. Discussion

On jaw cyst images, the combination of a medical backbone, box guidance, and feature refinement gives 93.79% Dice. Box guidance provides a direct input for a clinician's spatial assessment. The pattern across the other modalities depends on the metric: PromptFormer has the best reported BUSI scores and the best ISIC 2017 IoU and Recall, but its ISIC Dice is 0.08 percentage points below SAMUS. These are separate training-and-evaluation experiments within each dataset, rather than tests of cross-dataset transfer. Replacing a natural-image encoder may reduce a mismatch in pre-trained features; the present comparisons do not isolate domain shift or show that it has been eliminated.

The detail modules need to be considered with their localization input. The complete configuration is best on Jaw Cyst and BUSI, but some unprompted configurations lose accuracy when MDRM is added. Module

interaction is also apparent on BUSI (Table 4): the PPPM-only Dice of 86.82% falls to 85.44% with MDRM but without FGPM, whereas the full combination gives 90.93%. Prompt guidance may help select lesion features from the information being aggregated, although the other modules also affect the outcome. Inference requires 29.84 GFLOPs and reaches 36.29 FPS in the reported tests. These measurements support evaluation on less powerful devices, with deployment testing and clinical validation still needed to determine practical usefulness.

UKAN benefits little from PPPM compared with the other backbones. Compatibility between its image representations and the prompt features may limit the gain. The same interface should therefore not be assumed equally useful across architectures. Very small or early-stage lesions may also supply too little spatial detail for effective multi-scale refinement; this remains a limitation requiring targeted evaluation.

The present results apply to 2D images with bounding boxes assumed to be reasonably accurate. The model has no direct representation of context across a 3D volume, and its interface does not support point clicks or diagnostic descriptions. Other input conditions have not been evaluated here.

7. Conclusions

PromptFormer provides bounding box guidance to a medical segmentation backbone through PPPM, using SAM's prompt encoder and mask decoder without its original image encoder. This retains an interactive input while avoiding the computation associated with that encoder. FGPM processes global frequency information, and MDRM combines encoder levels and spatial scales to refine weak margins and local detail under prompt guidance. On Jaw Cyst, the framework obtains 93.79% Dice and 88.67% IoU, with competitive segmentation results also obtained on BUSI and ISIC 2017. Computational demand is lower than that of the prompted comparators in the resource test.

Backbone choice affects the gain from PPPM, and the benefit of refinement depends on the accompanying modules. Further medical imaging evaluations need to consider these dependencies. We plan to test additional prompt types, including textual descriptions, for interaction with PPPM in clinical workflows.

Author Contributions

C.L.: conceptualization, methodology, software, writing—original draft; Z.R.: investigation, validation, visualization, writing—original draft; T.T.: data curation, formal analysis; Z.W.: software, validation; Y.W.: investigation, formal analysis, funding acquisition; X.P.: supervision, writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62503209, in part by the Gansu Province Science and Technology Program under Grants 26YFFA024 and 25JRRA721, and in part by the Open Research Fund of Anhui Province Key Laboratory of Machine Vision Detection and Perception (No. KLMVI-2025-HIT-08).

Institutional Review Board Statement

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Medical Ethics Committee of Quzhou People's Hospital (protocol code 2026-093). Approved on June 10, 2026.

Informed Consent Statement

Patient consent was waived because this was a retrospective study and all patient data were fully anonymized prior to analysis.

Data Availability Statement

Dataset 1 (jaw cyst) is proprietary and cannot be publicly shared owing to privacy and ethical constraints. Dataset 2 (BUSI, https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset) and Dataset 3 (ISIC 2017, https://challenge.isic-archive.com/landing/2017/) are publicly available from their official repositories.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used ChatGPT for language polishing and grammatical refinement. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Appendix A: Model Card

Table A1. Model card for PromptFormer.

Section	Details
Model Details	
Model Name	PromptFormer
Architecture	Plug-in Prompt Process Module (PPPM) + H2Former backbone + FGPM + MDRM
Intended Use	
Primary Use	Interactive medical image segmentation, especially for jaw cyst, breast ultrasound, and skin lesion images
Intended Users	Clinicians, radiologists, medical AI researchers
Out-of-scope	Non-medical natural images; fully automatic segmentation without user prompts
Training Data	
Dataset 1	Jaw cyst (proprietary, 2592 images, collected from Quzhou City People's Hospital)
Dataset 2	BUSI (public, 780 breast ultrasound images)
Dataset 3	ISIC 2017 (public, 2900 dermoscopy images)
Data Preprocessing	Resized to 256×256; bounding box prompts as per MedSAM protocol
Training Procedure	
Optimizer	Adam, learning rate = 1×10^{-4}
Loss	0.5 Dice + 0.5 Binary Cross-Entropy
Batch size	8
Hardware	NVIDIA GeForce RTX 3090 (24 GB)
Training duration	~20 epochs (early stopping)
Evaluation Metrics	
Primary	Dice Similarity Coefficient, Intersection over Union (IoU), Recall
Evaluation Results (Test Sets)	
Jaw Cyst	Dice 93.79%, IoU 88.67%, Recall 93.89%
BUSI	Dice 90.93%, IoU 83.94%, Recall 95.42%
ISIC 2017	Dice 90.49%, IoU 83.57%, Recall 92.25%
Limitations and Biases	
Known Limitations	Performance may degrade on extremely small lesions; requires accurate bounding box prompts; trained on single-center jaw cyst data, may not generalize to all populations
Potential Biases	Data may reflect demographic and imaging protocol biases from the source hospitals; public datasets are curated and may differ from real-world clinical distributions

References

1. 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, Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
2. He, A.; Wang, K.; Li, T.; et al. H2Former: An Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2023**, *42*, 2763–2775. https://doi.org/10.1109/tmi.2023.3264513.
3. Chen, J.; Lu, Y.; Yu, Q.; et al. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
4. Sofiiuk, K.; Petrov, I.A.; Konushin, A. Reviving Iterative Training with Mask Guidance for Interactive Segmentation. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 3141–3145. https://doi.org/10.1109/icip46576.2022.9897365.
5. Marinov, Z.; Jäger, P.F.; Egger, J.; et al. Deep Interactive Segmentation of Medical Images: A Systematic Review and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10998–11018. https://doi.org/10.1109/tpami.2024.3452629.
6. Kirillov, A.; Mintun, E.; Ravi, N.; 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. https://doi.org/10.1109/iccv51070.2023.00371.
7. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; et al. Parameter-Efficient Transfer Learning for NLP. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 2790–2799.
8. Hu, E.J.; Shen, Y.; Wallis, P.; et al. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
9. Li, C.; Liu, X.; Li, W.; et al. U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 4652–4660.

10. Li, P.; Qiu, Z.; Zhan, Y.; et al. Multi-Scale Bottleneck Residual Network for Retinal Vessel Segmentation. *J. Med. Syst.* **2023**, *47*, 102. https://doi.org/10.1007/s10916-023-01992-7.
11. Hu, J.; Yang, Z.; Wang, P.; et al. Federated Learning for Medical Image Analysis: Privacy-Preserving Paradigms and Clinical Challenges. *Trans. Artif. Intell.* **2025**, *1*, 153–169. https://doi.org/10.53941/tai.2025.100010.
12. Piao, H.; Wu, D.O. Federated Continual Learning for Privacy-Preserving, Reliable and Interpretable Multi-Center Corneal Diseases Diagnosis. *Trans. Artif. Intell.* **2026**, *2*, 119–130. https://doi.org/10.53941/tai.2026.100008.
13. Takahashi, S.; Sakaguchi, Y.; Kouno, N.; et al. Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. *J. Med. Syst.* **2024**, *48*, 84. https://doi.org/10.1007/s10916-024-02105-8.
14. Cao, H.; Wang, Y.; Chen, J.; et al. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. In *Lecture Notes in Computer Science, Proceedings of the Computer Vision–ECCV 2022 Workshops, Tel Aviv, Israel, 23–27 October 2022*; Springer: Cham, Switzerland, 2023; pp. 205–218. https://doi.org/10.1007/978-3-031-25066-8_9.
15. Zhou, Z.; He, A.; Wu, Y.; et al. Spatial-Frequency Dual Domain Attention Network for Medical Image Segmentation. In Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Lisbon, Portugal, 3–6 December 2024; pp. 4076–4081. https://doi.org/10.1109/bibm62325.2024.10822613.
16. Zhang, M.; Feng, J.; Yang, H.; et al. IRPruneDeXt: Efficient Infrared Small Target Detection via Musical Wavelet-Regularized Channel Pruning. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 20229–20242. https://doi.org/10.1109/tnnls.2025.3594958.
17. Chen, L.; Li, J.; Ge, H. TBUnet: A Pure Convolutional U-Net Capable of Multifaceted Feature Extraction for Medical Image Segmentation. *J. Med. Syst.* **2023**, *47*, 122. https://doi.org/10.1007/s10916-023-02014-2.
18. Zhu, Z.; Zhu, H.; Lu, S.Y.; et al. Dual-Kernel Adapter: Expanding Spatial Horizons for Data-Constrained Medical Image Analysis. In Proceedings of the International Conference on Learning Representations, Rio de Janeiro, Brazil, 23–27 April 2026; Volume 2026, pp. 116057–116088.
19. Rother, C.; Kolmogorov, V.; Blake, A. GrabCut: Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Trans. Graph.* **2004**, *23*, 309–314. https://doi.org/10.1145/1015706.1015720.
20. Liu, Q.; Xu, Z.; Bertasius, G.; et al. SimpleClick: Interactive Image Segmentation with Simple Vision Transformers. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 22233–22243. https://doi.org/10.1109/iccv51070.2023.02037.
21. Ma, J.; He, Y.; Li, F.; et al. Segment Anything in Medical Images. *Nat. Commun.* **2024**, *15*, 654. https://doi.org/10.1038/s41467-024-44824-z.
22. Wu, J.; Wang, Z.; Hong, M.; et al. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. *Med. Image Anal.* **2025**, *102*, 103547. https://doi.org/10.1016/j.media.2025.103547.
23. Zhang, K.; Liu, D. Customized Segment Anything Model for Medical Image Segmentation. *arXiv* **2023**, arXiv:2304.13785.
24. Peng, Z.; Xu, Z.; Zeng, Z.; et al. SAM-PARSER: Fine-Tuning SAM Efficiently by Parameter Space Reconstruction. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 4515–4523.
25. Cai, Y.; Fan, D.J.; Gao, F.; et al. BoxPolypSAM: Leveraging SAM in Box-Supervised Polyp Segmentation. In Proceedings of the 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), Houston, TX, USA, 14–17 April 2025; pp. 1–5. https://doi.org/10.1109/isbi60581.2025.10980797.
26. Mazurowski, M.A.; Dong, H.; Gu, H.; et al. Segment Anything Model for Medical Image Analysis: An Experimental Study. *Med. Image Anal.* **2023**, *89*, 102918. https://doi.org/10.1016/j.media.2023.102918.
27. Lin, X.; Xiang, Y.; Yu, L.; et al. Beyond Adapting SAM: Towards End-to-End Ultrasound Image Segmentation via Auto Prompting. In *Lecture Notes in Computer Science, Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2024, Marrakesh, Morocco, 6–10 October 2024*; Springer: Cham, Switzerland, 2024; pp. 24–34. https://doi.org/10.1007/978-3-031-72111-3_3.
28. Zheng, H.; Jiang, X.; Xu, X.; et al. RBMDC-Net: Effective Jaw Cyst Segmentation Network Using Residual Bottleneck and Multiscale Dilated Convolution. *IEEE Access* **2025**, *13*, 3683–3695. https://doi.org/10.1109/access.2024.3525408.
29. Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; et al. Deep Learning Approaches for Data Augmentation and Classification of Breast Masses Using Ultrasound Images. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 1–11. https://doi.org/10.14569/ijacsa.2019.0100579.
30. Codella, N.C.F.; Gutman, D.; Celebi, M.E.; et al. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 168–172. https://doi.org/10.1109/isbi.2018.8363547.
31. Wang, Y.; Han, W.; Peng, H.; et al. SAM-SS: Straightforward and Efficient Designs Based on Segment Anything Model for Semantic Segmentation. *IEEE Trans. Comput. Soc. Syst.* **2026**, *13*, 5323–5331. https://doi.org/10.1109/tcss.2026.3686911.
32. Lin, A.; Chen, B.; Xu, J.; et al. DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–15. https://doi.org/10.1109/tim.2022.3178991.

33. Wu, Y.; Wang, Z.; Yang, X.; et al. Trans-SAM: Transfer Segment Anything Model to Medical Image Segmentation with Parameter-Efficient Fine-Tuning. *Knowl.-Based Syst.* **2025**, *310*, 112909. https://doi.org/10.1016/j.knosys.2024.112909.
34. Xiong, X.; Wu, Z.; Tan, S.; et al. SAM2-UNet: Segment Anything 2 Makes Strong Encoder for Natural and Medical Image Segmentation. *Vis. Intell.* **2026**, *4*, 2. https://doi.org/10.1007/s44267-025-00106-w.
35. Feng, S.; Zhao, H.; Shi, F.; et al. CPFNet: Context Pyramid Fusion Network for Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 3008–3018. https://doi.org/10.1109/tmi.2020.2983721.
36. Gu, R.; Wang, G.; Song, T.; et al. CA-Net: Comprehensive Attention Convolutional Neural Networks for Explainable Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2021**, *40*, 699–711. https://doi.org/10.1109/tmi.2020.3035253.
37. Gu, Z.; Cheng, J.; Fu, H.; et al. CE-Net: Context Encoder Network for 2D Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 2281–2292. https://doi.org/10.1109/tmi.2019.2903562.
38. Chen, G.; Li, L.; Dai, Y.; et al. AAU-Net: An Adaptive Attention U-Net for Breast Lesions Segmentation in Ultrasound Images. *IEEE Trans. Med. Imaging* **2023**, *42*, 1289–1300. https://doi.org/10.1109/tmi.2022.3226268.
39. Zheng, S.; Lu, J.; Zhao, H.; et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6877–6886. https://doi.org/10.1109/cvpr46437.2021.00681.
40. Huang, X.; Deng, Z.; Li, D.; et al. MISSFormer: An Effective Medical Image Segmentation Transformer. *arXiv* **2021**, arXiv:2109.07162.
41. Zhang, Y.; Liu, H.; Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In *Lecture Notes in Computer Science, Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021, Strasbourg, France, 27 September–1 October 2021*; Springer: Cham, Switzerland, 2021; pp. 14–24. https://doi.org/10.1007/978-3-030-87193-2_2.
42. Wu, H.; Chen, S.; Chen, G.; et al. FAT-Net: Feature Adaptive Transformers for Automated Skin Lesion Segmentation. *Med. Image Anal.* **2022**, *76*, 102327. https://doi.org/10.1016/j.media.2021.102327.