2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

SAMCA: Segment Anything Model with Double Click Training and Shared Weight Adapter for Medcial Ultrasound Image Segmentation

YiRu Huo[1,2†], YiChen Shi[3,4†], Jun Feng[1,2*], Liu Yang[1,2] and Na Liu[1,2]

[1] School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang, China
[2] Key Laboratory of Electromagnetic Environmental Effects and Information Processing, Shijiazhuang, China
[3] Shanghai Jiao Tong University, Shanghai, China
[4] Eastern Institute of Advanced Study, Ningbo, China

Abstract. Segmentation of medical ultrasound images is crucial for clinical diagnosis. However, challenges such as low contrast and blurred boundaries make obtaining large-scale labeled data for model training difficult. The Segment Anything Model (SAM), excelling at prompt-based segmentation in natural images, shows promise for ultrasound applications. In light of this, we propose SAMCA, a promptable medical ultrasound image segmentation model. SAMCA incorporates a shared weight adapter designed to efficiently transfer information between layers, allowing SAM to adapt to the complexities of medical ultrasound imaging. Additionally, we introduce a double click training strategy, where the first set of click prompts is used to provide guidance information for the initial target area, and the second set focuses on correcting local errors in the segmentation error-prone areas. A dynamic fusion mechanism ensures that the second set leverages the global context of the first set during refinement. Experimental comparisons with classic and recent segmentation networks demonstrate that SAMCA achieves state-of-the-art (SOTA) performance on the challenging TN3K and BUSI datasets, with DSC scores of 86.36% and 89.55%, respectively. Moreover, SAMCA is significantly more lightweight, requiring only 3% of parameter updates compared to SAM-Med2d. Our code will be publicly available at here.

Keywords: Medical ultrasound image segmentation, Segment anything model, Shared weight adapter, Double click training.

1 Introduction

Medical ultrasound image segmentation has become a cornerstone of medical image analysis and diagnosis due to its advantages, including real-time imaging, radiation-free, and low cost [2]. However, accurate segmentation typically requires a large amount of expert-annotated data, which must be carefully labeled by trained medical professionals. Compared to other medical imaging technologies, such as CT and MRI

[16], ultrasound images present unique challenges due to their inherent characteristics, including low contrast, blurred boundaries, and tissue abnormalities, which complicate the segmentation process [17] Therefore, improving segmentation accuracy, overcoming challenges related to image quality, and reducing the workload of medical annotators are crucial objectives.

In recent years, interactive image segmentation algorithms have made significant strides by integrating user prior knowledge and iteratively refining segmentation masks through direct interactions [21]. Deep learning-based interactive segmentation algorithms offer a substantial improvement over traditional manual pixel-level annotation techniques, allowing users to obtain detailed pixel-level annotations with simple and intuitive interactions. This approach holds great promise for reducing annotation costs and improving applications in medical image analysis [14]. Building on this progress, the emergence of foundation models has revolutionized the development of intelligent models. Adapting pre-trained, large-scale models for various downstream tasks is becoming increasingly popular due to their superior generalization capabilities and efficient training on smaller datasets [27]. A prime example of this trend is the Segment Anything Model (SAM) [12], a state-of-the-art visual foundation model designed for promptable image segmentation. Trained on a large-scale natural image dataset, SAM has demonstrated impressive zero-shot performance on various tasks in natural image contexts, offering promising avenues for accelerating medical data annotation [20].

However, recent evaluations [11] show that direct application of SAM to medical images, whether prompted or not, often leads to suboptimal results. This is largely due to the domain gap between natural and medical images. Ultrasound images differ significantly in texture, contrast, and noise patterns. Unlike natural images, which typically exhibit sharp edges, rich textures, and consistent lighting, ultrasound images are characterized by speckle noise, low contrast, blurred anatomical structures, and susceptibility to artifacts—features stemming from their acoustic imaging mechanism. These factors, along with the high structural variability and lack of spatial regularity in ultrasound data, hinder the direct transferability of models like SAM trained on natural images. Therefore, structural modifications and tailored training strategies are essential for adapting SAM to the unique challenges of ultrasound imaging.

A straightforward approach to bridge the domain gap between natural and medical images is to fine-tune SAM using medical images [26]. Several studies [25] have used parameter-efficient fine-tuning (PEFT) [31], demonstrating promising performance in medical imaging tasks. However, despite the advantages of parametric efficiency, training on specific datasets often leads to limited generalization of feature representations [35]. In addition, regions such as lesions or heterogeneous tissues in medical images often exhibit variable shapes and low contrast, making segmentation of these areas prone to errors and instability [28]. Existing methods [24] rely on random clicks for training, which may prevent the model from fully capturing and optimizing the unique features of these areas. As a result, the inherent capabilities of SAM may be compromised, leading to a decrease in segmentation performance and requiring more manual correction to achieve accurate results.

These problems are particularly prominent in ultrasound image segmentation tasks. The inherent noise interference and low contrast characteristics of ultrasound images

make existing methods face greater challenges in data annotation efficiency and computing resource consumption. To address this challenge, the present study adapts SAM for ultrasound image segmentation by using PEFT techniques and integrating adapter structure to fine-tune SAM. Additionally, a double click training strategy is introduced to direct the model's attention to areas prone to errors, thereby enhancing both local and global segmentation accuracy. Our specific contributions include:

1. We introduce SAMCA, a promptable segmentation method of medical ultrasound images that significantly enhances the capability of SAM for medical applications. In this work, we propose a shared weight adapter that facilitates efficient cross-domain knowledge transfer and captures fine-grained features, all at a low training cost.
2. We propose a double click training strategy to enable the model to learn the areas prone to segmentation errors in a targeted manner. This strategy optimizes the segmentation results through the synergy of two sets of click prompts.
3. We introduce a dynamic fusion mechanism to facilitate the collaboration between the prompt information, ensuring that global information is effectively utilized during segmentation refinement.
4. We conducted comprehensive quantitative and qualitative experiments on four different ultrasound datasets to evaluate the effectiveness of the proposed method.

2 Related Work

2.1 Medical Ultrasound Image Segmentation

Medical ultrasound image segmentation is crucial for identifying structures such as lesions and organs. Early methods, including thresholding [18] and clustering [22], often lacked robustness and accuracy. With the rise of deep learning, CNN and Transformer based models have become dominant. CNN-based models like U-Net [23] laid the foundation for modern segmentation networks. Subsequent variants such as CE-Net [9] and CA-Net [6] introduced dilated convolutions and attention mechanisms to enhance feature representation and segmentation precision. Transformer-based models have shown strong performance. TransUNet [4] combines CNNs and Transformers in a hybrid encoder, while SwinUnet [3] leverages hierarchical Swin Transformer blocks. Other methods, including TransFuse [36] and H2Former [10], use parallel CNN-Transformer branches to balance local detail and global context modeling. These approaches typically follow a U-shaped architecture with skip connections, improving segmentation accuracy across various scenarios.

2.2 SAM for Medical

The Segment Anything Model (SAM) has shown remarkable zero-shot segmentation capabilities with diverse input prompts [33]. However, medical images—especially ultrasound images—differ significantly from natural images in terms of texture and noise characteristics, necessitating specially designed adaptations are needed to accommodate SAM [15].

MedSAM [13] fine-tuned SAM on a self-collected multi modal dataset for general medical image segmentation. To improve adaptability, recent studies have explored parameter-efficient fine-tuning (PEFT) techniques. SAMed [34] employed a low-rank adaptation (LoRA) strategy, while MSA [26] introduced lightweight adapters to enhance performance with reduced training cost. Building on MSA, SAMIHS [25] optimized adapter design for downstream tasks with fewer parameters. SAM-Med2D [5] further leveraged SAM for medical image analysis and fine-tuned it on the Sa-med2d-20m dataset[30] using an adapter adjustment technique.

Inspired by these advances, SAMCA is developed with a novel adapter architecture that integrates and extends prior designs, aiming to address the unique challenges of ultrasound image segmentation. In addition, SAMCA adopts a targeted training strategy tailored to the characteristics of ultrasound data. These design choices are intended to improve adaptability to domain-specific tasks and further explore the potential of SAM in medical imaging.

3 Methodology

3.1 Macro architecture of SAMCA

As shown in Fig.1, the overall structure of SAMCA is inherited from SAM. To enhance the adaptability of the model to medical ultrasound images, two adapter modules are introduced between each transformer module. By sharing a part of the adapter parameters at the same position in different Transformer layers in the image encoder, the model can learn a more general feature representation of ultrasound images while reducing the total number of parameters.

For prompt input, considering the fuzzy boundaries and irregular shapes commonly found in medical ultrasound images, we use point prompts in this paper and employ the double click training strategy. During training, the model receives two sets of click prompts in each iteration. The first set of prompts, called the initial click, is primarily used to provide guidance information for the initial target area. By randomly selecting a foreground area from the ground truth (GT) as the click position, this prompt helps guide the model to perform global segmentation in each iteration. The second set of prompts, called the guided click, focuses on refining the model's performance in areas prone to errors. This set of prompts is dynamically adjusted based on the segmentation error from the first set, effectively correcting the model's performance in local, complex regions.

In order to ensure that the guided click prompt can use the global information provided by the initial click prompt when refining the segmentation and avoid the spread of local errors, we fusion the embedding of the initial click prompt into the embedding of the guided click prompt (as indicated by the blue arrows in Fig.1). Through this interactive training and optimization process, the double click training strategy significantly improves the model performance in medical ultrasound image segmentation tasks. The effectiveness of the model is evaluated by comparing the GT with the guided mask and the initial mask, using two sets of losses, where the guided mask is considered the model's output.

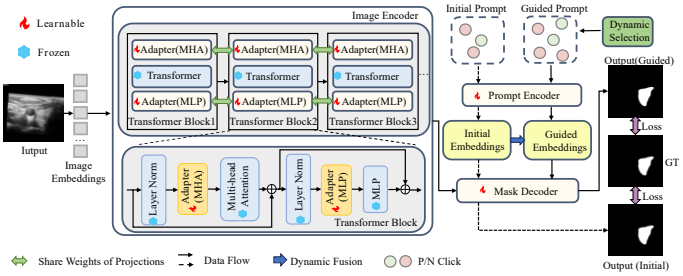

Fig. 1. The structure of SAMCA. During training, the model uses two prompt sets-initial clicks and guided clicks, shown as solid and dashed lines. Two losses are computed: one from the initial output and Ground Truth, and another from the guidance output, which serves as the final output.

3.2 Comparison and Design of the Shared Weights Adapter

We compared four different adapter structures, as shown in Fig. 2. The MSA adapter [26] (Fig. 2 (a)) adopts a simple design with two projection layers and an activation function, enabling efficient feature transformation and core feature focus, particularly suitable for small datasets and ultrasound segmentation.

In contrast, the SAM-Med2D adapter [5] (Fig. 2 (b)) introduces channel and spatial attention to enhance feature expression but increases computational complexity. Its reliance on global average pooling and convolution leads to detail loss, especially near lesion boundaries, potentially degrading segmentation accuracy.

The SAMIHS adapter [25] (Fig. 2 (c)) improves parameter efficiency via projection sharing and feature adaptation within transformer layers. However, its lack of local detail extraction limits performance on low-contrast or structurally complex ultrasound images.

The shared weights adapter (Fig. 2 (d)) combines the advantages of previous structures. It retains the MSA adapter's core feature focus, while integrating the SAMIHS adapter's parameter reconstruction and shared projection mechanism for better feature perception. This cross-layer mechanism similarity to densely connected neural architectures [32], which have been shown to facilitate feature reuse and improve information flow. By sharing weights across layers, the adapter can capture both low-level and high-level abstractions without increasing model complexity. Inspired by SAM-Med2D's attention module, we added a convolution layer after the dimensionality reduction layer to improve feature utilization. This adapter shows higher stability and efficiency in smaller datasets and medical ultrasound segmentation. Given an input feature $m \in R^{h \times w \times c}$, it sequentially passes through the symmetric down projection W_{down}, the activation function, the convolution layer, and the up projection W_{up}. The scaling factor R_{l} and the offset factor B_{l} are used to ensure personalized adjustment of the feature distribution of each layer. The forward process in the adapter can be expressed as:

$$\text{Adapter}(m) = m + \sigma\left(Conv\left(mW_{\text{down}}B_{\text{l}}\right)\right)W_{\text{up}} + R_{\text{l}} \tag{1}$$

where σ denotes the GeLU activation function, and *Conv* denotes 2D convolution.

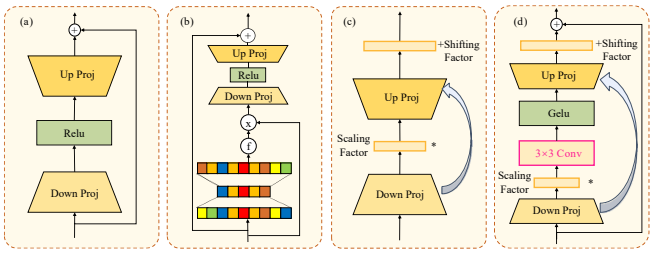

Fig. 2. Comparison between the shared weights adapter and other adapter structures is presented. (a) illustrates the MSA adapter structure; (b) illustrates the SAM-Med2d adapter structure; (c) illustrates the SAMIHS adapter structure; (d) illustrates our shared weights adapter structure.

3.3 Double Click Training

Traditional interactive image segmentation methods [21] typically involve iterations during the training process, where clicks are placed directly in false positive and false negative regions. Although this approach is straightforward, the random selection of click locations in error regions may lead to inadequate learning in some error-prone areas. To address this limitation, we propose a double click training (DCT) strategy that does not restrict click generation to predefined error regions. Instead, it dynamically adjusts click placement, achieving more flexible and comprehensive coverage of error-prone areas.

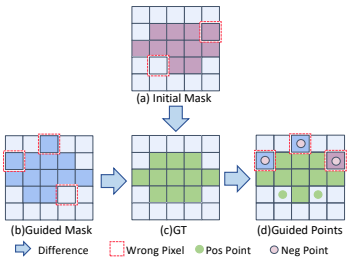

Fig. 3. Guided points dynamic selection. (a)shows the generated mask for the initial click in a particular iteration; (b)shows the generated mask for the guided click in a particular iteration; (c)represents the GT; (d)shows the selection of guided points.

The double click training process for generating guided clicks is illustrated in Fig. 3, where (a), (b), (c) and (d) correspond to the initial mask, the guided mask, the GT, and the click points produced by DCT, respectively. During the iterative sampling process, DCT utilizes two distinct sampling strategies: one that produces positive and negative clicks based on the initial mask, and the other one is informed by the guided mask.
As depicted in Fig. 3(d), guided click generation is performed by comprehensively analyzing the initial mask, the guided mask, and the GT. Specifically, during the training iteration phase, the model first locates the common area where the initial mask and the GT label are not aligned (that is, (a) to (c)), and preferentially applies positive clicks to correct the error. Next, the model locates the common area where the guided mask and

the GT label are not aligned (that is, (b) to (c)), and on this basis applies both positive clicks and negative clicks to optimize the areas prone to segmentation error. Finally, by integrating the above processes, an optimized click position is generated, so that the model can focus on local error correction in error-prone areas and consider global segmentation optimization during the training iteration process.

To ensure that the guided click prompt fully utilizes the global information from the initial click prompt, we propose a dynamic fusion mechanism that combines the embedding information from the initial click prompt with that of the guided click prompt, allowing for local adjustment while maintaining global consistency. The dynamic weight adjustment mechanism is shown in Fig. 4, where E_1 and E_2 represent the embeddings of the initial and guided clicks, respectively, and E'_2 represents the final embedding of the guided click prompt.

Compared to simply using E_2 and E_1+E_2, the dynamic weights adjust the influence of the first and second sets of prompt embeddings based on the loss values returned from different training rounds. Specifically, as segmentation accuracy improves, the dynamic fusion mechanism gradually reduces the weight of the first set of prompt embedding and increases the weight of the second set, thereby correcting detail errors at a deeper level. The sensitivity of the weight adjustment is controlled by a smoothing scaling factor to ensure stable weight changes during each iteration. Eventually, as the model converges, the weights of both sets of prompt's approach 1. The weight adjustment formula is as follows:

$$\lambda = \lambda_{\min} + \left(\lambda_{\max} - \lambda_{\min}\right) \cdot \sigma\left(-\zeta \cdot \Delta L\right) \tag{2}$$

$$\mu = \mu_{\min} + \left(\mu_{\max} - \mu_{\min}\right) \cdot \sigma\left(\zeta \cdot \Delta L\right) \tag{3}$$

where $\sigma(x)$ is the Sigmoid function; ζ is a scaling factor that controls the sensitivity to the change in loss ΔL. λ_{min} and μ_{min} are set to 0.75; λ_{max} and μ_{max} are set to 1.25.

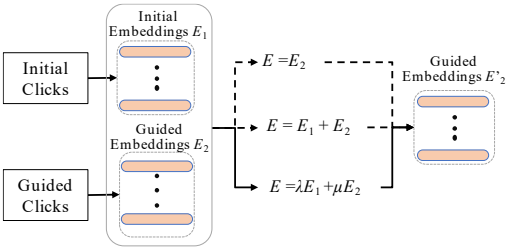

Fig. 4. Comparison of the dynamic fusion mechanism with the E_2 and E_1+E_2 methods.

To address the challenges of ultrasound image segmentation—such as low contrast, blurred boundaries, and complex, small lesions—a hybrid loss function is employed, combining Dice Loss, Binary Cross-Entropy (BCE) Loss, and Intersection over Union (IoU) Loss. Each component is designed to target a specific limitation of the task: Dice Loss effectively mitigates class imbalance by focusing on the overlap between predicted and ground truth regions; BCE Loss strengthens pixel-wise classification by penalizing incorrect predictions at the individual pixel level; and IoU Loss emphasizes

global shape alignment and boundary consistency. The hybrid loss function is expressed as follows, the weights α, β, and γ are set to 6:1:3.

$$Loss = \alpha L_{Dice} + \beta L_{BCE} + \gamma L_{IoU} \tag{4}$$

3.4 The Agorithm of SAMCA

We illustrate the method in Algorithm 1. The input includes an ultrasound image $I \in R^{H\times W\times 3}$, along with the number of initial clicks N_{initial} and guided clicks N_{Guided}. The encoder ϕ_{enc} first extracts the image feature map F_{img} from I.

Before iteration, the initial and guided click prompts, P_{initial} and P_{guided}, are generated based on GT. These point-based prompts are encoded into feature tensors T_{initial} and T_{guided}. These are decoded by ϕ_{decoder} is then used to generate two segmentation masks: the initial mask $F_{\text{i-mask}}$ and guided mask $F_{\text{g-mask}}$.

Table 1. Comparison of ten segmentation methods on the TN3K Dataset. Best results in each category are highlighted in **bold**, second-best are underlined.

Methods	Dsc (%) ↑	mIoU (%) ↑	HD (mm) ↓	Acc (%) ↑
U-Net [23]	79.01±21.87	69.40±23.09	34.12±23.77	96.44±4.17
CE-Net [9]	80.37±19.74	70.66±21.51	32.79±24.28	96.41±4.56
SwinUnet [3]	70.08±23.29	58.19±24.11	44.13±25.61	94.94±4.35
CA-Net [6]	80.52±19.35	70.78±21.28	33.65±24.93	96.41±4.21
TransFuse [35]	78.50±21.60	68.60±22.86	32.44±23.17	96.44±3.96
H2Former [10]	**82.48±18.30**	**73.31±20.41**	**30.58±22.10**	**96.95±3.59**
TransUNet [4]	81.44±19.31	72.31±21.08	30.98±21.68	96.94±3.26
SAM(1p) [12]	26.69±21.74	17.47±16.70	150.95±114.86	87.74±11.58
Med2d(1p) [5]	68.50±30.33	55.33±28.93	74.17±90.21	94.76±12.01
SAMCA(1p)	78.79±19.44	68.51±22.01	57.65±71.13	96.18±5.17
SAM(3p) [12]	28.64±20.30	18.45±15.07	165.11±108.28	87.39±11.85
Med2d(3p) [5]	73.21±23.43	60.70±25.04	56.82±75.71	95.69±10.39
SAMCA(3p)	83.82±13.91	74.03±17.09	41.39±48.37	97.18±3.25
SAM(5p) [12]	45.95±20.21	32.09±17.39	141.10±92.69	89.28±10.19
Med2d(5p) [5]	76.24±16.58	64.22±18.77	44.98±56.08	96.30±3.49
SAMCA(5p)	**86.36±12.10**	**77.71±15.02**	**32.31±37.59**	**97.79±2.40**

To guide the iterative refinement, error regions between the predicted masks and the ground truth are analyzed. Based on these discrepancies, two types of click regions are identified: Based on discrepancies between the masks and GT, two regions P_{pri} and P_{sec} are defined to update P_{guided}. In each iteration, a dynamic fusion process combines T_{initial} and T_{guided} into T_{fused}, enhancing segmentation accuracy. This iterative process continues, with click prompts being updated at each step based on feedback from the

current prediction, gradually improving alignment with the ground truth. Once convergence is achieved or the maximum number of iterations is reached, the final guided mask $F_{\mathrm{g-mask}}$is returned as the output segmentation mask M.

Algorithm 1 SAMCA

Input: Image $I \in R^{H\times W\times 3}$, Initial Clicks Number N_{initial}, Guided Clicks Number N_{Guided}
Output: Segmentation Mask M

$F_{\mathrm{img}} \leftarrow \phi_{enc}(I)$
while not converged **do**
  P_{initial}, P_{guided} ← Generate_Clicks (N_{initial}, N_{Guided}, GT)
  T_{initial}, $T_{\mathrm{guided}} \leftarrow \phi_{\mathrm{enc}}(P_{\mathrm{initial}}, P_{\mathrm{guided}})$
  $F_{\mathrm{i-mask}}$, $F_{\mathrm{g-mask}} \leftarrow \phi_{\mathrm{decoder}}(F_{\mathrm{img}}, P_{\mathrm{initial}}, P_{\mathrm{guided}})$
  Adjust based on initial and guided mask:
  P_{pri} ← Identify_Difference_Regions ($F_{\mathrm{i-mask}}$, GT)
  P_{sec} ← Identify_Difference_Regions ($F_{\mathrm{g-mask}}$, GT)
  P_{guided} ← Apply_Guided_Clicks (P_{pri}, P_{sec})
  Dynamic Fusion for Iterative Optimization:
  for $i = 1$ **to** N_{iter} **do**
    $T_{\mathrm{initial}} \leftarrow \phi_{\mathrm{enc}}(P_{\mathrm{initial}})$
    $T_{\mathrm{guided}} \leftarrow \phi_{\mathrm{enc}}(P_{\mathrm{guided}})$
    T_{fused} ← Dynamic_Fusion (T_{initial}, T_{guided})
    $F_{\mathrm{i-mask}} \leftarrow \phi_{\mathrm{decoder}}(F_{\mathrm{img}}, T_{\mathrm{initial}})$
    $F_{\mathrm{g-mask}} \leftarrow \phi_{\mathrm{decoder}}(F_{\mathrm{img}}, T_{\mathrm{fused}})$
    Update clicks based on segmentation feedback:
    P_{initial} ← Update_Clicks ($F_{\mathrm{i-mask}}$, GT)
    P_{guided} ← Update_Clicks ($F_{\mathrm{i-mask}}$, $F_{\mathrm{g-mask}}$, GT)
  end for
end while
$M \leftarrow F_{\mathrm{g-mask}}$
return M

4 Experiments

4.1 Dataset and Setup

To evaluate the proposed model, we conduct experiments on four public ultrasound datasets: TN3K [7], BUSI [1], DDTI [19], and UDIAT [29]. Dataset partitioning and preprocessing follow the protocol in [8]. Since DDTI and UDIAT share segmentation targets with TN3K and BUSI, they are treated as unseen datasets to assess generalization.

Experiments are conducted using an NVIDIA RTX 3090 GPU with Python 3.9 and PyTorch. Training is performed with a batch size of 4, a learning rate of 0.0001, and 20 epochs using the Adam optimizer. All images are resized to 256×256. Evaluation metrics include Dice Similarity Coefficient (DSC), mean Intersection over Union (mIoU),

Hausdorff Distance (HD), Accuracy (Acc), Sensitivity (Sen), and Specificity (Spe). The model results reported as the mean ± standard deviation.

4.2 Comparison with the state-of-the-art methods

Experiments on TN3K and BUSI dataset We extensively compare our method with various SOTA CNN and Transformer-based medical image segmentation methods, including U-Net [23], CE-Net [9], SwinUnet [3], CA-Net [6], TransFuse [35], H2Former [10], and TransUNet [4] and two SAM-based methods (SAM [12] and SAM-Med2d [12]). Non-SAM methods use existing public performance data as a reference [8]. Tables 1 present the experimental results in theTN3K datasets, showing that the proposed method achieves significant performance in various evaluation metrics.

Table 2. Comparison of ten segmentation methods on the TN3K Dataset. Best results in each category are highlighted in **bold**, second-best are underlined.

Methods	Dsc (%) ↑	mIoU (%) ↑	HD (mm) ↓	Acc (%) ↑
U-Net [23]	78.11±25.45	69.60±26.81	33.60±32.78	96.69±4.71
CE-Net [9]	81.68±23.53	73.62±24.03	29.19±31.03	96.86±5.47
SwinUnet [3]	67.23±25.79	55.58±25.88	47.02±34.18	95.28±5.23
CA-Net [6]	81.68±21.55	73.49±23.19	28.67±28.25	96.85±5.31
TransFuse [35]	73.52±28.16	64.28±28.05	34.95±37.18	96.21±5.85
H2Former [10]	81.48±22.91	73.34±24.13	27.84±27.02	96.85±5.50
TransUNet [4]	**82.22±24.08**	**74.77±24.57**	**27.54±28.25**	**97.26±4.84**
SAM(1p) [12]	39.54±24.78	27.83±20.85	131.01±113.18	91.45±9.92
Med2d(1p) [5]	82.39±14.18	73.24±16.93	44.18±58.78	97.01±4.06
SAMCA(1p)	84.28±15.71	75.27±18.36	41.54±52.20	97.42±3.92
SAM(3p) [12]	40.77±25.22	29.51±21.69	154.24±126.18	90.60±10.30
Med2d(3p) [5]	87.11±11.92	78.95±14.32	32.20±43.92	97.66±2.91
SAMCA(3p)	87.61±13.76	79.58±16.53	33.38±49.92	98.02±3.77
SAM(5p) [12]	59.38±25.59	46.83±25.58	132.18±114.11	92.57±8.74
Med2d(5p) [5]	89.49±9.78	81.89±12.57	**26.70±40.34**	98.04±3.27
SAMCA(5p)	**89.55±11.14**	**82.28±14.13**	29.09±41.19	**98.26±2.39**

On the TN3K dataset, SAMCA method shows competitive performance compared to other SAM-based methods in key metrics, On the TN3K dataset, SAMCA outperforms other SAM-based methods across all key metrics. With increasing prompts (1p/3p/5p), its DSC improved from 78.79% to 86.36%, surpassing SAM and SAM-Med2d. SAMCA(5p) achieved a lower HD (32.31 mm) and higher mIoU (77.71%) and Acc (97.79%) than SAM-Med2d(5p). Compared to SOTA CNN and Transformer based method, SAMCA(3p) achieves optimal performance. It delivers competitive results in Dsc, HD, mIoU, and Acc These results underscore the advantages of SAMCA in both segmentation accuracy and boundary precision.

The effectiveness of SAMCA in ultrasound image segmentation is further validated on the BUSI dataset, as shown in Table 2, achieving a DSC of 84.2% with a single click and improving to 89.55% with five clicks. Compared to SAM-Med2d, SAMCA consistently achieves higher DSC, mIoU, and Accuracy, along with lower boundary error, indicating improved segmentation accuracy and boundary precision. When compared to CNN and Transformer based methods, SAMCA(5p) significantly outperforms U-Net, CE-Net, SwinUnet, and TransFuse in both DSC and HD, demonstrating its effectiveness in segmentation quality and classification accuracy.

High DSC scores on multiple ultrasound datasets indicate accurate structural localization and boundary delineation. In clinical practice, this can help doctors more efficiently identify and measure target regions, improving diagnostic efficiency and reducing errors from manual annotation. Moreover, the segmentation results can assist in semi-automated labeling, easing annotation workload and supporting large-scale dataset construction.

Comparison of sensitivity and specificity of different methods. Fig. 5 shows the experimental results of the SAM-based method in terms of Sen and Spe metrics. The subfigures (a) and (b) display the test results on the TN3K and BUSI datasets, respectively. SAMCA with different numbers of clicks has achieved the best results compared to other SAM methods, which shows that this method is more sensitive to lesions. SAMCA demonstrates excellent performance in both sensitivity and specificity, highlighting its high potential for medical image annotation tasks—particularly in scenarios demanding precise segmentation, such as tumor delineation and lesion area identification.

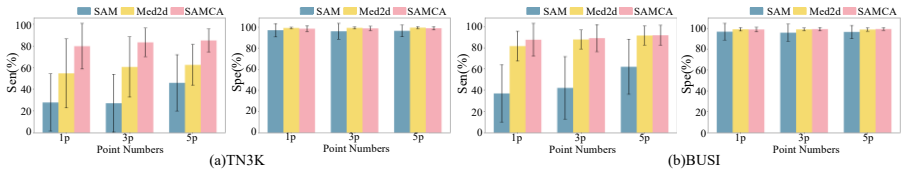

Fig. 5. Comparison of sensitivity and specificity of SAM-based methods.

Generalization Ability. Assessing the generalization capability of task-specific methods is critical, as it helps determine how effectively these models perform on new, unseen datasets. As shown in Fig. 6, we quantitatively evaluate this aspect. When comparing performance on familiar (seen) and unfamiliar (unseen) datasets, SAMCA exhibits the smallest performance degradation across both segmentation tasks, while maintaining high DSC scores. This not only reflects strong generalizability but also suggests potential clinical applicability, as stable and accurate segmentation across diverse data is important for supporting reliable diagnosis and treatment planning.

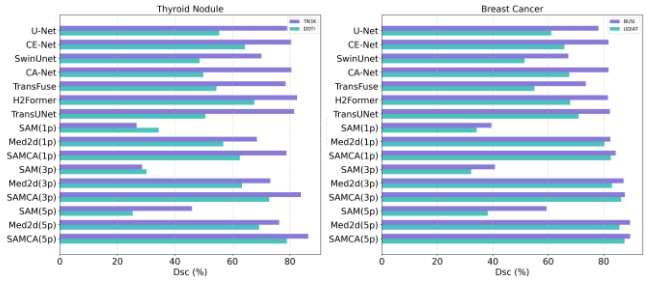

Fig. 6. Comparison of SAMCA with different methods on see datasets (highlighted in purple) and unseen datasets not previously encountered (indicated in blue). Higher blue bars indicate stronger generalization ability.

Visualization and analysis. Fig. 7 provides a visual comparison of six methods, including U-Net, CE-Net, TransUNet, H2Former, SAM-Med2d, and the proposed SAMCA. Rows 1 and 2 display the results of the BUSI dataset, row 3 presents results from the TN3K dataset, and rows 4 and 5 show results from the DDTI and UDIAT datasets, respectively. The proposed SAMCA method performs exceptionally well across all datasets, accurately capturing target areas while maintaining clear boundaries. Compared to other methods, SAMCA effectively avoids boundary inconsistencies and segmentation artifacts. On the BUSI dataset, the SAMCA method accurately segments tumor regions with clear boundaries and no obvious artifacts. On the TN3K dataset, SAMCA also performs excellently in complex images, handling noise and inconsistencies in the data effectively. Moreover, SAMCA demonstrates superior generalization capabilities on the DDTI and UDIAT datasets, handling targets with higher irregularity and complexity, demonstrating strong adaptability.

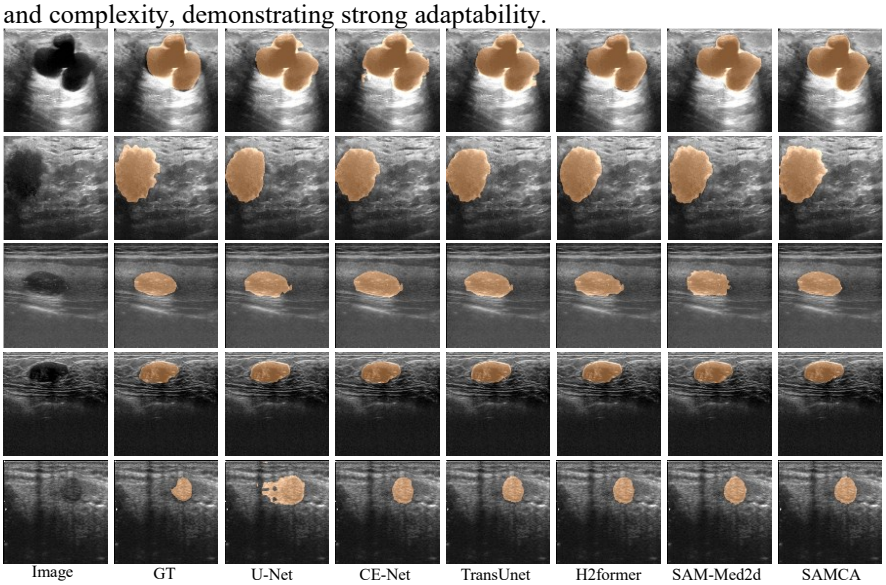

Fig. 7. Visualization of comparative experiments.

4.3 Comparison of Model Training Parameter

Table 3 compares SAMCA with other SAM-based methods on the BUSI dataset in terms of training parameters and DSC performance. While SAM-Med2D delivers competitive results, its high parameter count increases computational cost, limiting its practicality in resource-constrained settings. SAMCA achieves consistently better DSC with different adapters, reaching the best performance while maintaining low training complexity. This parameter sharing strategy enhances regularization and improves generalization on small-scale medical datasets that are often noisy and poorly annotated.

Table 3. Compare the effects of different adapter structures on model performance and training parameters. 'TP': trainable parameters.

Method	Adapter	Dsc (%) ↑	TP↓
Med2d	-	82.39±14.18	184.56M
SAMCA	+adapter(MSA)[26]	82.95±16.48	7.61M
	+adapter(share)[25]	83.73±15.89	4.07M
	+adapter(Ours)	84.28±15.71	5.15M

Fig. 8 further compares SAMCA with SOTA CNN and Transformer based method. SAMCA outperforms these methods in DSC and mIoU, using significantly fewer parameters, which highlights its computational efficiency. In terms of segmentation accuracy and boundary precision, SAMCA clearly surpasses other models, while methods like SwinUnet and CE-Net perform worse in HD, indicating weaker boundary handling.

SAMCA outperforms SOTA methods in both segmentation accuracy and computational efficiency, demonstrating its wide applicability and potential in interactive segmentation of medical ultrasound images. Its lightweight design significantly reduces the computational resources required for training, making it a promising candidate for deployment in resource-constrained environments. This may help facilitate the practical adoption of intelligent medical imaging tools in real-world clinical settings.

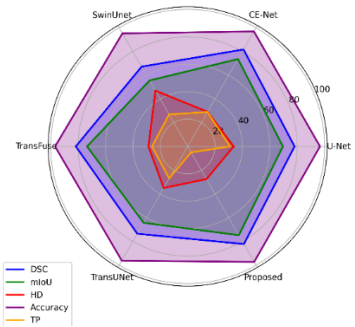

Fig. 8. Comparison of CNN and Transformer-based model performance and training parameters on TN3K. 'TP': trainable parameters.

4.4 Ablation experiments

Effectiveness of each component. To evaluate the contribution of each module in SAMCA, we conducted ablation experiments on the TN3K and BUSI datasets by testing different combinations of Adapter (MHA), Adapter (MLP), and DCT strategy. These components were incrementally integrated into the original SAM architecture for training and evaluation. As shown in Table 4, Models 2 and 3 introduce only a single adapter module, each resulting in notable improvements in model performance. With the addition of DCT in Model 5, performance is further enhanced. Fig. 9 shows the visualization results of the ablation experiment of different modules. These results demonstrate that the integration of adapters and DCT enables SAMCA to achieve superior segmentation performance, confirming its effectiveness in ultrasound image segmentation.

Table 4. Ablation study on different component combinations of SAMCA. 'A1' and 'A2' represent Adapter (MHA) and Adapter (MLP), and 'DCT' Stands for double click training.

Model	A1	A2	DCT	TN3K		BUSI	
				Dsc(%)↑	HD(mm)↓	Dsc(%)↑	HD(mm)↓
Model1	×	×	×	26.69±21.74	150.95±114.86	39.54±24.78	131.01±113.18
Model2	√	×	×	76.26±23.31	73.68±75.54	81.56±19.83	67.69±68.39
Model3	×	√	×	76.83±23.80	63.62±76.97	81.28±19.48	62.58±61.76
Model4	√	√	×	77.36±23.27	62.75±76.15	82.09±18.62	62.40±59.35
Model5	√	√	√	**78.79±19.44**	**57.65±71.13**	**84.28±15.71**	**41.54±52.20**

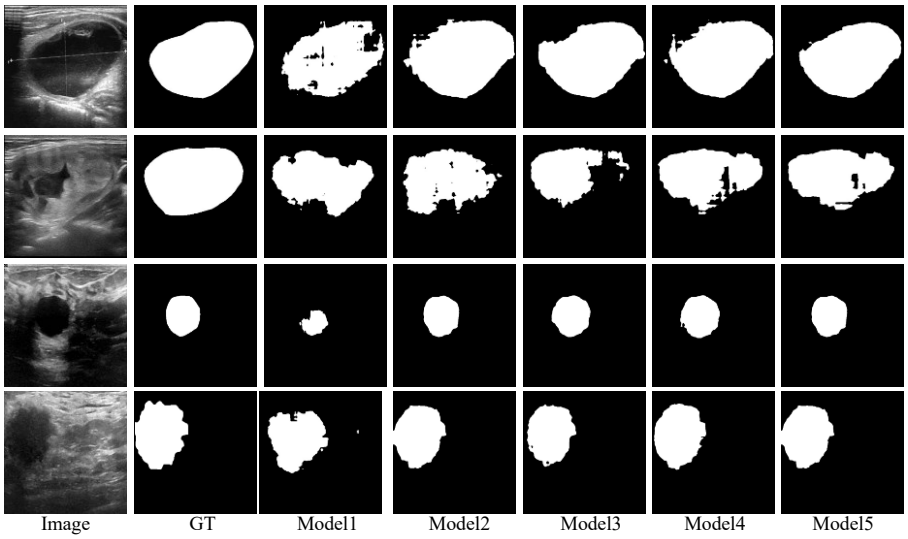

Fig. 9. Visualization of segmentation results for different models in our framework.

Effectiveness of the Dynamic Fusion. To validate the effectiveness of the dynamic fusion mechanism, we evaluated SAMCA on the BUSI dataset using different fusion strategies, as illustrated in Fig. 4, with results summarized in Table 5. Without fusion, using only guided prompt embedding E_2, the model achieved a DSC of 86.22%. Applying simple addition (E_1+E_2) increased the DSC to 88.99%. With the dynamic fusion method ($\lambda E_1 + \mu E_2$), performance improved further, reaching a maximum DSC of 89.55%. These results highlight that dynamic fusion effectively balances the influence of initial and guided prompts, leading to more accurate segmentation.

Table 5. Comparison of Results for Different Fusion Methods

Methods	Dsc(%) ↑	mIoU(%) ↑	HD(mm)↓	Acc(%) ↑
E_2	82.83±17.75	72.97±22.68	62.13±56.21	96.46±4.52
E_1+E_2	83.27±17.56	73.66±19.87	48.36±56.49	96.29±4.58
$\lambda E_1 + \mu E_2$	**84.28±15.71**	**75.27±18.36**	**41.54±52.20**	**97.42±3.92**

5 Conclusion

The Segment Anything Model (SAM), though effective in natural image segmentation, faces challenges in medical imaging due to low contrast and complex anatomical structures. To address these limitations, we propose SAMCA, which integrates SAM with a shared-weight adapter to enhance cross-domain knowledge transfer for medical image segmentation. A double click training strategy is introduced, where the first click identifies the target region and the second refines difficult areas, improving local accuracy. Furthermore, a dynamic fusion mechanism strengthens the interaction between prompts, mitigating local errors and enhancing overall segmentation performance. Experiments on four public ultrasound datasets demonstrate that SAMCA surpasses existing mainstream methods in both segmentation accuracy and computational efficiency, highlighting its potential for interactive ultrasound image segmentation.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief 28, 104863 (2020)
2. Ansari, M.Y., Mangalote, I.A.C., Meher, P.K., Aboumarzouk, O., Al-Ansari, A., Halabi, O., Dakua, S.P.: Advancements in deep learning for b-mode ultrasound segmentation: a comprehensive review. IEEE Transactions on emerging topics in computational intelligence (2024)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swinunet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

5. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
6. Das, D., Nayak, D.R., Pachori, R.B.: Ca-net: A novel cascaded attention-based network for multi-stage glaucoma classification using fundus images. IEEE Transactions on Instrumentation and Measurement (2023)
7. Gong, H., Chen, J., Chen, G., Li, H., Li, G., Chen, F.: Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. Computers in biology and medicine 155, 106389 (2023)
8. Gowda, S.N., Clifton, D.A.: Cc-sam: Sam with cross-feature attention and context for ultrasound image segmentation. In: European Conference on Computer Vision. pp. 108–124. Springer (2024)
9. Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., Liu, J.: Ce-net: Context encoder network for 2d medical image segmentation. IEEE transactions on medical imaging 38(10), 2281–2292 (2019)
10. He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H.: H2former: An efficient hierarchical hybrid transformer for medical image segmentation. IEEE Transactions on Medical Imaging (2023)
11. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? Medical Image Analysis 92, 103061 (2024)
12. 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 IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
13. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications 15(1), 654 (2024)
14. Mahadevan, S., Voigtlaender, P., Leibe, B.: Iteratively trained interactive segmentation. arXiv preprint arXiv:1805.04398 (2018)
15. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Medical Image Analysis 89, 102918 (2023)
16. Mecheter, I., Abbod, M., Amira, A., Zaidi, H.: Deep learning with multiresolution handcrafted features for brain mri segmentation. Artificial intelligence in medicine 131, 102365 (2022)
17. Mishra, D., Chaudhury, S., Sarkar, M., Soin, A.S.: Ultrasound image segmentation: a deeply supervised network with attention to boundaries. IEEE Transactions on Biomedical Engineering 66(6), 1637–1648 (2018)
18. Pare, S., Kumar, A., Singh, G.K., Bajaj, V.: Image segmentation using multilevel thresholding: a research review. Iranian Journal of Science and Technology, Transactions of Electrical Engineering 44(1), 1–29 (2020)
19. Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E.: An open access thyroid ultrasound image database. In: 10th International symposium on medical information processing and analysis. vol. 9287, pp. 188–193. SPIE (2015)
20. Quan, Q., Tang, F., Xu, Z., Zhu, H., Zhou, S.K.: Slide-sam: Medical sam meets sliding window. arXiv preprint arXiv:2311.10121 (2023)
21. Ramadan, H., Lachqar, C., Tairi, H.: A survey of recent interactive image segmentation methods. Computational visual media 6(4), 355–384 (2020)
22. Ramesh, K., Kumar, G.K., Swapna, K., Datta, D., Rajest, S.S.: A review of medical image segmentation algorithms. EAI Endorsed Transactions on Pervasive Health & Technology 7(27) (2021)

23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
24. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d: towards general-purpose segmentation models for volumetric medical images. arXiv preprint arXiv:2310.15161 (2023)
25. Wang, Y., Chen, K., Yuan, W., Tang, Z., Meng, C., Bai, X.: Samihs: adaptation of segment anything model for intracranial hemorrhage segmentation. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
26. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. Medical image analysis p. 103547 (2025)
27. Xin, Y., Luo, S., Zhou, H., Du, J., Liu, X., Fan, Y., Li, Q., Du, Y.: Parameterefficient fine-tuning for pre-trained vision models: A survey. arXiv preprint arXiv:2402.02242 (2024)
28. Xu, M., Ma, Q., Zhang, H., Kong, D., Zeng, T.: Mef-unet: An end-to-end ultrasound image segmentation algorithm based on multi-scale feature extraction and fusion. Computerized Medical Imaging and Graphics 114, 102370 (2024)
29. Yap, M.H., Goyal, M., Osman, F., Martí, R., Denton, E., Juette, A., Zwiggelaar, R.: Breast ultrasound region of interest detection and lesion localisation. Artificial Intelligence in Medicine 107, 101880 (2020)
30. Ye, J., Cheng, J., Chen, J., Deng, Z., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. arXiv preprint arXiv:2311.11969 (2023)
31. Yuan, Y., Zhan, Y., Xiong, Z.: Parameter-efficient transfer learning for remote sensing image-text retrieval. IEEE Transactions on Geoscience and Remote Sensing (2023)
32. Yue G, Han W, Jiang B, et al. Boundary constraint network with cross layer feature integration for polyp segmentation[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(8): 4090-4099.
33. Zhang, C., Puspitasari, F.D., Zheng, S., Li, C., Qiao, Y., Kang, T., Shan, X., Zhang, C., Qin, C., Rameau, F., et al.: A survey on segment anything model (sam): Vision foundation model meets prompt engineering. arXiv preprint arXiv:2306.06211 (2023)
34. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
35. Zhang, Y., Shen, Z., Jiao, R.: Segment anything model for medical image segmentation: Current applications and future directions. Computers in Biology and Medicine p. 108238 (2024)
36. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 14–24. Springer (2021)