Domesticating SAM for Breast Ultrasound Image Segmentation via Spatial-frequency Fusion and Uncertainty Correction

Wanting Zhang[1], Huisi Wu[1(✉)], and Jing Qin[2]

[1] College of Computer Science and Software Engineering, Shenzhen University
[2] Centre for Smart Health, The Hong Kong Polytechnic University
hswu@szu.edu.cn

Abstract. Breast ultrasound image segmentation is a challenging task due to the low contrast and blurred boundary between the breast mass and the background. Our goal is to utilize the powerful feature extraction capability of segment anything model (SAM) and make out-of-domain tuning to help SAM distinguish breast masses from background. To this end, we propose a novel model called $SFRecSAM$, which inherits the model architecture of SAM but makes improvements to adapt to breast ultrasound image segmentation. First, we propose a spatial-frequency feature fusion module, which utilizes the fused spatial-frequency features to obtain a more comprehensive feature representation. This fusion feature is used to make up for the shortcomings of SAM's ViT image encoder in extracting low-level feature of masses. It complements the texture details and boundary structure information of masses to better segment targets in low contrast ultrasound images. Second, we propose a dual false corrector, which identifies and corrects false positive and false negative regions using uncertainty estimation, to further improve the segmentation accuracy. Extensive experiments demonstrate that the proposed method significantly outperforms state-of-the-art methods on two representative public breast ultrasound datasets: BUSI and UDIAT. Codes is available at `https://github.com/dodooo1/SFRecSAM`.

Keywords: Breast ultrasound image segmentation · Segment anything model · High-frequency information

1 Introduction

Breast cancer is one of the most common causes of death among women around the world [30]. Ultrasound imaging is extensively utilized in clinical for identifying and diagnosing breast cancer. Interpretation of breast ultrasound images requires extensive domain knowledge associated with benign or malignant breast masses, and hence only experienced radiologists can analyze the images and produce the reports. Manual examination of breast ultrasound images is, however, labor-intensive, time-consuming, and error-prone. To the end, computer-aided diagnosis systems equipped with AI models have been developed to assist radiologists in interpreting breast ultrasound images [36]. One of the key components

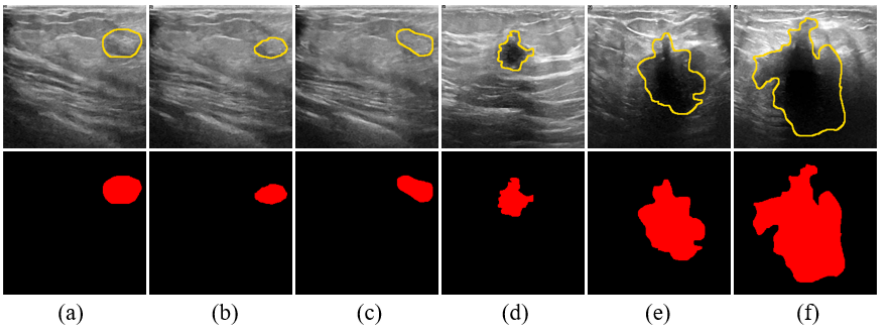

Fig. 1: Challenges of segmenting lesion areas from breast ultrasound image: (a)-(c) low contrast between breast masses and background and (d)-(f) the variations of morphology and size in breast masses. Yellow contours represent the edges of masses, and red regions represent ground truth.

in such a system is a segmentation model, which is able to automatically segment lesions from ultrasound images for further analysis. However, it is very challenging to develop such a segmentation model due to speckle noise and low contrast of ultrasound images (Fig. 1(a)-(c)), which makes it extremely difficult to distinguish lesion boundaries [16, 28]. In addition, the large variation of morphology and size of breast masses (Fig. 1(d)-(f)) makes this task even harder.

Recent years, many deep learning models have been developed to meet these challenges. Most early investigations utilized the local modeling capability of convolutional neural networks (CNNs) to detect, segment, and classify breast masses [2, 35]. Later, transformer-based approaches have been proposed to overcome the limitation of CNNs in capturing global features by harnessing self-attention mechanisms, achieving better results particularly in cases with ambiguous boundaries and irregular shapes [33, 45]. In order to take complementary advantages of CNNs and transformers, some studies combined these two architectures, using CNNs to extract local features while leveraging transformers to capture more remote dependencies [11, 18, 21]. While these models have more powerful representation capability, these approaches still cannot adequately address above-mentioned challenges due to the inadequate network capacity and limited high-quality training data. The recently proposed segment anything model (SAM) [20] has powerful feature extraction capabilities owing to its large network scale and excellent generalization performance because of a large amount of training data. In addition, it is capable of effectively segmenting regions of interest based on prompts. Unfortunately, original SAM is trained only on natural images. As there are significant differences in distribution characteristics between natural and medical images, SAM, in principle, cannot be generalized to medical image segmentation tasks.

Visual out-of-domain optimization is a crucial method for adapting SAM to medical image segmentation. Currently, most research focuses on retraining

SAM's parameters using medical dataset, or adding new learnable adapters to fine-tune some parameters. For example, MedSAM [24] trains SAM on a large medical image dataset and fine-tune SAM's mask decoder, while SAMed [41] applies a low-rank-based (LoRA) strategy to fine-tune SAM with lower computational cost. SAMUS [22] further introduces multiple adapters to generalize the trained SAM image encoder to medical images. In addition, it introduces parallel CNN branches to extract spatial features, injecting local information of images into the image encoder through cross-branch attention. However, most of these methods only consider enhancing or supplementing spatial features of SAM while neglecting frequency features, which are usually of significance importance in medical image analysis due to special imaging mechanisms in medical domain.

In this paper, we propose a novel model adapted from SAM for lesion segmentation in breast ultrasound images. First, we propose a novel spatial-frequency feature fusion (SFF) module, which aims to utilize the fused spatial-frequency features to enhance local information of the segmented targets and hence further eliminate background interference. The SFF module is capable of addressing the challenge of low contrast between lesions and background in breast ultrasound images. Second, as SAM is sensitive to the quality of prompts, we propose a dual false corrector (DFC) to stabilize the segmentation performance of SAM in medical images when the prompts are not good enough to yield satisfactory results. The proposed DFC uses point prompt augmentation to estimate the uncertainty map of generated multiple predictions in order to automatically identify and rectify false positive (FP) and false negative (FN) regions in segmentation results. We call the proposed model $SFRecSAM$. We conduct extensive experiments on two benchmarking datasets of breast ultrasound images: BUSI [1] and UDIAT [39], and comprehensively compare our model with task-specific state-of-the-art (SOTA) methods and recently proposed SAM-based methods. Experimental results show that our $SFRecSAM$ significantly outperforms existing models on these two datasets.

Our main contributions are summarized as follows.

- We propose a novel model for breast ultrasound image segmentation based on SAM, called $SFRecSAM$, which harnesses the powerful feature extraction capability of SAM while equipping SAM with new modules to make it adapt to ultrasound images.
- We propose an innovative spatial-frequency feature fusion (SFF) module and a new dual false corrector (DFC) module and integrate them into SAM; while the SFF generates comprehensive spatial-frequency fused features to supplement the model with detailed local information, the DFC module further employs uncertainty maps of prediction results to identify and correct FP and FN regions in the initial prediction mask.
- Our model significantly outperforms state-of-the-art task-specific methods and recently proposed SAM-based methods on two famous breast ultrasound datasets (BUSI and UDIAT), demonstrating the effectiveness of the proposed model.

2 Related Work

2.1 Breast Ultrasound Image Segmentation

Breast ultrasound (BUS) image segmentation presents a significant challenge, owing to the low contrast and speckle noise inherent in ultrasound images, as well as the significant variations in the size and shape of breast masses [16, 28]. Initially, CNN-based method segments breast masses through excellent local information perception [2, 35], but these methods were constrained by their deficiency in global awareness. Therefore, some researchers combine transformer with CNN to process global and local information simultaneously, thereby enhancing the feature extraction capabilities of the model [11, 18, 21]. However, these task-specific models, affected by the limitation of network architecture and the scarcity of high-quality annotated medical datasets, still cannot achieve fine segmentation of lesion areas.

Different from task-specific models, SAM [20] is a foundation model for computer vision tasks that has been pre-trained on a large set of natural images. SAM decreases the dependence on specialized knowledge for image segmentation tasks and lowers the need for large amounts of annotated data. However, due to differences in characteristics such as pixel intensity, color, and texture between natural images and medical images, recent studies have revealed that SAM shows performance degradation in medical image segmentation [4, 13, 17, 25]. Some researchers begin to explore how to improve the performance of SAM on medical images segmentation tasks [6, 15, 22, 24, 31, 37, 41].

2.2 Spatial-frequency Domain Feature Fusion

The existing network predominantly focuses on spatial features while ignoring the features of the frequency domain. The high-frequency components of the image usually correspond to the drastic intensity changes on the edges of the segmentation targets, which can be used to effectively identify object boundaries. Moreover, high-frequency information reveals image texture and helps distinguish regions of similar brightness but different textures, which is crucial for detecting tumors in low-contrast ultrasound images [32,44]. These characteristics enable segmentation models that exploit high-frequency information can accurately delineate object edges and effectively separate objects from background.

The most commonly used methods to extract high-frequency information from images are Fourier transform and wavelet transform [3]. Wavelet transform, a multi-scale analysis method, can decompose an image into wavelet coefficients used to represent the characteristics of signals or images at different frequencies and locations [40]. Compared with Fourier transform, wavelet transform is more effective in processing multi-scale features of images, and also has better spatial locality and stronger anti-noise ability. Previously, researchers have explored some semantic segmentation methods by combining wavelet transform with deep neural networks [27, 43]. Common strategies include using wavelet transform as pre- or post-processing, and replacing certain layers of the CNNs

with wavelet transform [8,32,43]. But few methods consider both spatial features and frequency features, both of which are useful for segmentation. The convolutional layer in CNN helps reduce noise when extracting spatial features, but it makes the image features too smooth, resulting in the loss of some detailed information. High-frequency image features can retain the edge and texture information of the image, but will introduce strong high-frequency noise. The fused spatial-frequency features creates a more comprehensive image representation, providing a more discriminative and informative feature space for segmentation algorithms. This fusion takes advantage of the strengths of the two domains and compensates for each other's shortcomings to obtain more accurate results.

2.3 Uncertainty Estimation

Uncertainty estimation is the process of evaluating the confidence of network predictions, which is crucial for generating reliable predictions and improving model performance [19]. If the segmentation algorithm can assign a high degree of uncertainty to incorrect predictions, it will enable the model to make better decisions based on these valuable informations. Uncertainty in models can be categorized into two types: aleatoric uncertainty, caused by inherent noise in data, and epistemic uncertainty, resulting from limited knowledge or information about the model or data distribution [7, 46].

Due to the variability of disease pathology and the poor imaging quality of medical images, uncertainty estimation is crucial in the field of medical image analysis, such as lesion detection [26, 29], lung node segmentation [14] and brain tumor segmentation [34]. Uncertainty in medical image segmentation can be divided into two categories: (1) unclear regions or boundaries surrounding tissues, and (2) semantic ambiguity of the regions or boundaries [46]. Methods for estimating uncertainty can be divided into several types based on the number and nature of the deep neural networks [9]. Single deterministic method, which provides uncertainty for each pixel in a medical image with only a single forward pass, have received much attention. It can be roughly divided into two implementations, one is to model and train a single network to quantify uncertainty; the other one is to use additional modules to estimate uncertainty of prediction [9]. Ultimately, the reliability and accuracy of model predictions can be enhanced by processing those high-uncertainty pixels based on the uncertainty associated with each pixel's classification.

3 Method

3.1 Overview

The original SAM's framework consists of an image encoder, a prompt encoder, and a mask decoder that merge image and prompt embeddings to generate segmentation masks. Starting from two key problems, we adapt SAM to make it suitable for breast ultrasound image segmentation. First, ultrasound images are

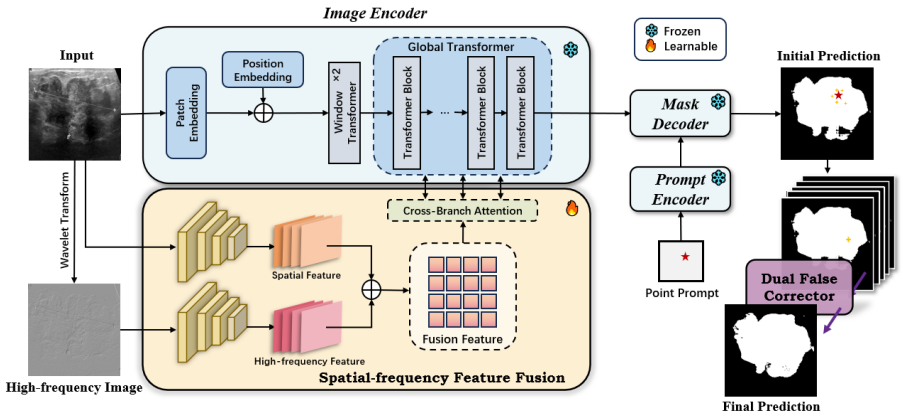

Fig. 2: Overview of our proposed method, which mainly introduces a spatial-frequency feature fusion (SFF) module and a dual false corrector (DFC). SFF fuses spatial features and high-frequency features to generate comprehensive fused features, supplementing more discriminative information of the segmented target for SAM's ViT image encoder. DFC uses the high uncertainty map to find and correct the FP and FN regions in the initial prediction.

characterized by strong noise and blurred boundaries, and it is crucial to obtain fine-grained features of segmented targets to supplement the local information lacking in SAM's ViT image encoder. Second, since SAM is strongly affected by the quality of prompts, a single prediction will inevitably produce FP and FN regions. Based on these problems, we introduce a spatial-frequency feature fusion (SFF) module and a dual false corrector (DFC) into the SAM's framework.

The SFF module integrates the fusion features obtained by merging high-frequency image features and spatial image features into the model. It aims to combine the advantages of the spatial features and the frequency features to supplement the model with comprehensive information. The DFC aims to further correct FP and FN regions in the initial prediction by estimating uncertainty, which improves the accuracy and stability of predictions. As shown in Fig. 2, our model's framework inherits the original architecture of SAM and freezes all its original parameters. We introduce an additional branch to extract high-frequency components of ultrasound images, learn task-specific knowledge from the high-frequency components and fuse them with valuable spatial features extracted by multi-scale convolution. Then the features extracted by ViT image encoder are used as query to establish global dependencies with the fused features, so that fine-grained features and global information complement each other. Finally, the outputs of the two branches are merged as the final image embedding and input into the mask decoder, which is combined with prompt embedding to obtain the prediction. We further utilize point prompt augmentation to generate multiple predictions, find and correct the FN and FP regions of the initial prediction by estimating aleatoric uncertainty to obtain the final prediction.

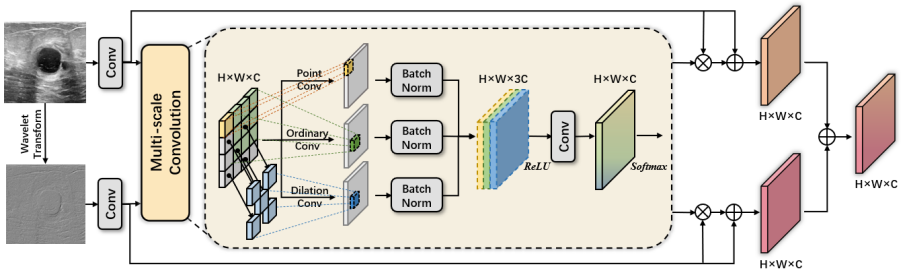

Fig. 3: Illustration of the spatial-frequency feature fusion (SFF) module. The SFF extracts and fuses spatial features and high-frequency features through multi-scale convolution to obtain powerful feature representations.

3.2 Spatial-frequency Feature Fusion

Transformer has strong global perception capability, but only global information is not enough when facing inputs with rich spatial structure such as images. It also needs to be able to capture local features, which often carry important visual information, such as edges and textures of objects. Notably, the high-frequency information inherent in the image captures sudden changes in signal intensity at the boundaries of the segmentation targets, corresponding to fine details and sharp edges in ultrasound images, providing useful information for breast masses segmentation. SAMUS [22] introduces CNN branches to extract spatial features to make up for the lack of local features of SAM's ViT image encoder, but the spatial features may be too smooth to lose some detailed informations. Inspired by this, we fuse the spatial image features with high-frequency image features to obtain the more comprehensive and robust features. This provides the model with more effective discriminative information and improves performance.

High-frequency Image Extraction. 2D images, which are essentially discrete non-stationary signals, contain rich frequency range and spatial location information. The Fourier transform, although able to provide global frequency information, is unable to effectively capture local spatio-temporal features when dealing with such signals. The wavelet transform, as a multi-scale analysis method, can effectively preserve these local features while decomposing the images. We apply wavelet transform to decompose the original image into four components: low-frequency (LL), horizontal high-frequency (HL), vertical high-frequency (LH), and diagonal high-frequency (HH). The high-frequency image is represented by combining the high-frequency components from different directions.

Fusion of High-frequency Features and Spatial Features. To capture different levels of features in ultrasound images, we use a multi-scale convolution module to extract valuable features and suppress unimportant features, as shown in Fig. 3. It consists of three types of convolutions with different receptive fields:

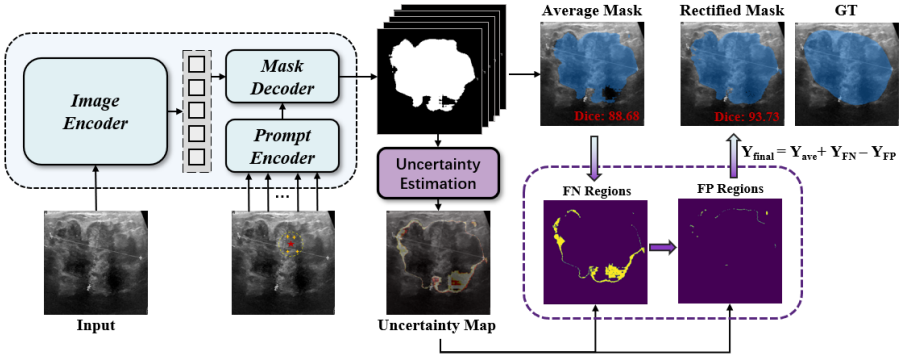

Fig. 4: Illustration of the dual false corrector (DFC). The DFC finds and corrects false positive and false negative regions in the initial prediction by uncertainty estimation.

point convolution, ordinary convolution (kernel size is 3×3, stride is 1, padding is 1), and dilated convolution (kernel size is 3×3, stride is 1, padding is 2, and dilation rate is 2). The final feature embedding is obtained by concatenating the outputs of the three convolutions. We extract features in the original image and high-frequency image through multi-scale convolution, and merge the feature embeddings of the two branches as the final image embedding.

3.3 Dual False Correction

Due to the sensitivity of SAM predictions to prompt quality and the low contrast and obvious shadows of ultrasound images, a single prediction has poor stability and may include a wide range of FP and FN regions [42]. FNPC [38] estimates uncertainty based on bounding box prompt to correct predictions, but this prompt is more difficult to deploy in clinical applications than simple point prompt. Inspired by this, we use a dual false corrector (DFC) to rectify the initial prediction of the model. We use a random single-point prompt as the initial prompt and adopt point prompt enhancement strategy to obtain multiple predictions which are used for uncertainty estimation. We apply a predefined threshold to identify high uncertainty mask, and the intersection of this mask with the segmentation mask and the background mask is considered as potential FP and FN regions. The final prediction is obtained by adding potential FN regions and removing potential FP regions from the initial prediction.

Point Prompt Enhancement Strategy. Since SAM's predictions are highly correlated with prompt quality, point prompts with changed positions will lead to differences in segmentation masks. As shown in Fig. 4, our sampling method is to perturb the initial prompt by random shifting, generating N point prompts $p = \{p^1, p^2, \cdots, p^N\}$ within a radius $R = \frac{1}{M} \min(H, W)$ around the center of the initial point prompt, where M is the radius ratio. We set M and N to 5 and

10 respectively. H and W denote height and width of the initial segmentation mask. This perturbation strategy enables the model to perform interactive segmentation with each point prompt at various locations, and generates different segmentation masks $Y = \{Y^1, Y^2, \cdots, Y^N\}$ for the subsequent correction.

Uncertainty Estimation. We estimate uncertainty in the network's predictions using an external uncertainty quantification method, which is separate from the underlying prediction task. By integrating multiple predictions generated by extra point prompts, we can obtain the combined segmentation result $\hat{Y} = \frac{1}{N+1}\sum_{i=0}^{N} Y^i$, with Y^0 denoting the initial prediction. The pixel value at each position (j, k) can be expressed as $\hat{Y}_{j,k} = \frac{1}{N+1}\sum_{i=0}^{N} Y^i_{j,k}$. Entropy can describe the degree of uncertainty in the state of a system, so we approximate the uncertainty of each pixel using the uncertainty form of entropy, with $\epsilon = 10^{-7}$:

$$U_{j,k} = -0.5 \cdot [\hat{Y}_{j,k} \cdot \log(\hat{Y}_{j,k} + \epsilon) + (1 - \hat{Y}_{j,k}) \cdot \log(1 - \hat{Y}_{j,k} + \epsilon)], \tag{1}$$

where U is the aleatoric uncertainty mask. We set a custom threshold for selecting high uncertainty areas as follows:

$$T_u = \min(U) + \gamma \cdot [\max(U) - \min(U)], \tag{2}$$

where γ is the threshold ratio, which we set to 0.5. We extract the high uncertainty mask through threshold T_u: $U_h = U > T_u$, with U_h being the final uncertainty mask used in the correction process.

False Negative and False Positive Correction. The correction is based on the assumption that pixels with similar intensity are more likely to belong to the same class. We estimate the high uncertainty mask U_h of the input image I, which highlights potential FN and FP regions. We utilize the initial prediction to estimate the average intensity of the images in the target and background regions as I_t and I_b. If the intensity of a pixel in U_h is similar to the average intensity I_t of the pixel points contained in the recognized target region, then we consider that pixel point should be included in the final segmentation mask. For false negative correction, $(1 - \hat{Y}) \cdot U_h$ is utilized to find possible FN regions. The range of similar intensity (I_{tl}, I_{th}) belonging to the segmentation target is determined based on I_t, and only pixel points with intensity within this range are retained in the final FN mask Y_{FN}. For false positive correction, $\hat{Y} \cdot U_h$ is utilized to find possible FP regions, and the final FP mask Y_{FP} is also determined within the range of background intensity (I_{bl}, I_{bh}) decided by I_b. The correction process is shown in Fig. 4, and the final prediction mask can be calculated as: $Y_{final} = \hat{Y} + Y_{FN} - Y_{FP}$.

4 Experiment

4.1 Experimental Setup

Datasets. We evaluate the proposed model using two publicly available datasets: BUSI [1] and UDIAT [39]. BUSI consists of 780 images and corresponding anno-

Table 1: Statistical comparison with different state-of-the-art methods on two breast ultrasound image datasets: BUSI and UDIAT.

Method	Year	Backbone	BUSI			UDIAT		
			mDice	mIoU	mHD	mDice	mIoU	mHD
TransUNet [5]	2021	ResNet-50	82.22	74.53	4.51	84.42	75.00	4.14
HiFormer [12]	2023	ResNet-50	82.72	74.38	4.92	86.57	78.52	4.10
H2Former [10]	2023	ResNet-34	81.48	72.67	7.76	89.95	82.15	5.78
MNFE-Net [23]	2023	ResNet-34	85.49	77.72	4.68	91.39	84.39	4.02
SAM [20]	2023	ViT-B	55.48	45.50	10.99	80.40	70.67	6.16
MSA [37]	2023	ViT-B	81.95	73.13	5.50	88.09	79.16	4.43
SAM-Med2D [6]	2023	ViT-B	80.12	69.28	8.50	84.76	75.28	5.82
SAMed [41]	2023	ViT-B	73.02	64.16	8.86	80.28	70.12	6.82
SAMUS [22]	2023	ViT-B	85.62	76.82	4.80	90.46	82.94	3.98
BUSSAM [31]	2024	ViT-B	86.68	78.23	4.72	90.93	83.01	3.95
Ours	-	ViT-B	**87.14**	**78.58**	**4.66**	**91.80**	**85.14**	**3.92**

tations covering 600 female patients aged 25 to 75, with the average size of each image being 500 × 500 pixels. We only use samples of benign and malignant lesions (437 benign and 210 malignant) in the experiment. UDIAT consists of 163 images corresponding to 110 benign and 53 malignant breast masses. We divide datasets into train set, test set and validation set according to 8:1:1.

Evaluation Metrics. We evaluate model performance using mean dice (mDice), mean interaction over union (mIoU), and mean hausdorff distance (mHD).

Implementation Details. The model is trained on a single RTX 3090 GPU. We use the original parameters of the SAM for initialization and randomly initialize the remaining parameters. The parameters of the prompt encoder and mask decoder are frozen during training, while the parameters of the image encoder can be learned. We adjust the ultrasound images to a resolution of 256×256 when training model. We select the AdamW optimizer for training, and set the initial learning rate to 0.0005, batch size to 4, and epoch to 400. The prompt mode is set to a random single-point prompt.

4.2 Comparison with State-of-the-art Methods

We compare with four task-specific methods and five SAM-based methods to verify the effectiveness of our model. Task-specific methods include TransUNet [5], HiFormer [12], H2Former [10] and MNFE-Net [23]. SAM-based methods include SAM-Med2D [6], SAMed [41], MSA [37], SAMUS [22] and BUSSAM [31].

Comparison with SAM-based Methods. As shown in Table 1, the SAM without fine-tuning shows predictable performance degradation on both datasets

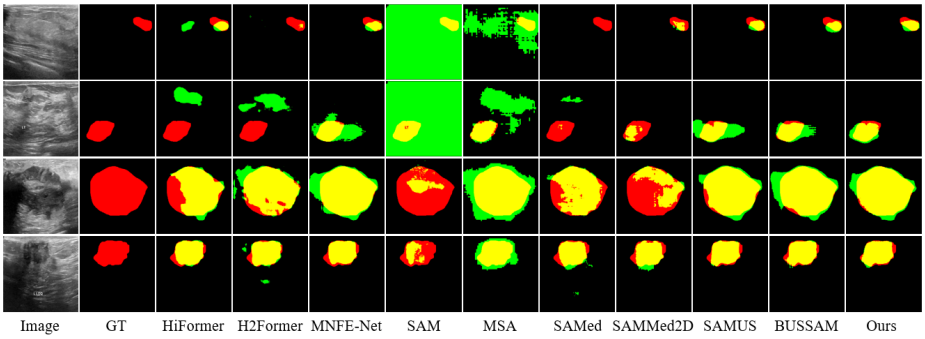

Fig. 5: Visual comparison with different methods on the BUSI test set. Red, green and yellow represent ground truth, prediction and their overlapping regions, respectively.

Table 2: Performance comparison. Efficiency on BUSI based on a single RTX 3090 GPU and batch size equal to 1.

Method	Resolution	BUSI		
		mDice	Params(M)	GFLOPs
SAM [20]	1024 × 1024	55.48	90.49	**371.99**
SAMed [41]	512 × 512	73.02	**90.36**	103.41
SAM-Med2D [6]	256 × 256	80.12	270.99	65.37
MSA [37]	256 × 256	81.95	100.92	35.36
SAMUS [22]	256 × 256	85.62	130.14	145.88
Ours	256 × 256	**87.14**	130.23	167.07

Table 3: Ablation studies on SFF and DFC on two breast ultrasound image datasets. SFF: spatial-frequency feature fusion. DFC: dual false corrector.

SFF	DFC	BUSI			UDIAT		
		mDice	mIoU	mHD	mDice	mIoU	mHD
		85.62	76.82	4.80	90.46	82.94	3.98
✓		86.27	77.54	4.70	91.68	84.93	3.95
	✓	86.30	77.61	5.07	90.74	83.38	3.94
✓	✓	**87.14**	**78.58**	**4.66**	**91.80**	**85.14**	**3.92**

due to its lack of specific knowledge related to medical images. After fine-tuning, all SAM-based methods achieve varying degrees of performance improvement. Among all tuning methods, our method achieves better performance with dice scores of 87.14% and 91.80% on the two datasets respectively, an improvement of 31.66% and 11.40% compared to SAM. Compared with SAMUS, which is specially tuned for ultrasound image segmentation, the dice scores increased by 1.52% and 1.34% respectively. The kernel density estimation in Fig. 6 shows that our method has a higher probability density in the dice interval, indicating the robustness of our model. This shows that the powerful fused spatial-frequency features supplement the detailed information of masses for SAM, helping the network learn more discriminative and informative features. In addition, based on more accurate prediction results obtained by utilizing high-frequency information, our method further applies the DFC to find and effectively correct the FP and FN regions in the initial prediction, resulting in better performance.

Comparison with Task-specific Methods. As shown in Table 1, our method exhibits performance improvements across all metrics for both datasets, surpassing four task-specific methods. This is not only due to SAM's powerful feature extraction capability and inherent model framework advantages, but also to the

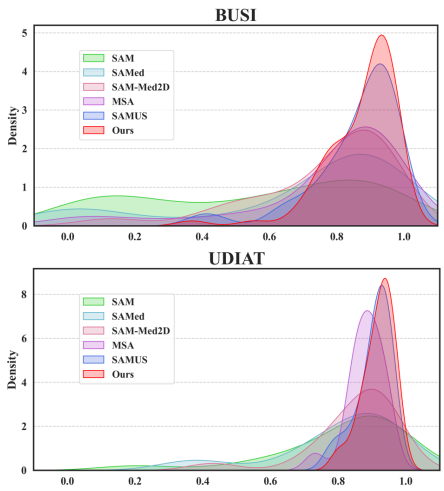

Fig. 6: Comparison of kernel density estimation with SAM-based methods on BUSI and UDIAT.

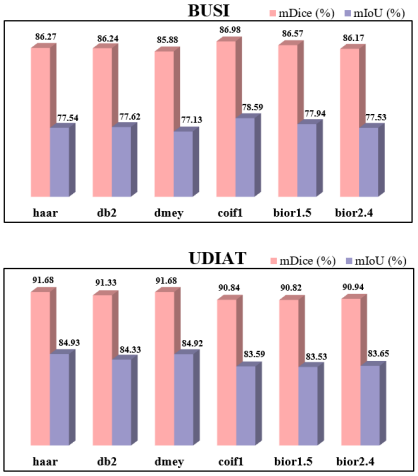

Fig. 7: Comparison of segmentation performance using different wavelet bases to extract high-frequency image.

specific designs we introduced to address the challenges of ultrasound image segmentation tasks. SAM can achieve segmentation performance equivalent to or even better than that of task-specific models with a small amount of parameter adjustment. It effectively avoids the performance limitations of task-specific models due to inadequate network capacity or complex training calculations.

Visual Comparison with SOTA. Fig. 5 is the visual segmentation results of breast masses under various methods. Our proposed model achieves fine segmentation of breast masses by introducing the SFF module and the DFC into SAM's framework. Especially in low-contrast situations, the performance improvement is significant. The reason is that our method can extract more comprehensive features in breast ultrasound images and perform fine rectification. The SFF module enables model to have strong perception of the texture details and boundary structure of segmented targets, while the DFC can correct results through point prompt enhancement and uncertainty estimation based on more accurate prediction, further improving the reliability of predictions.

Performance-efficiency Comparison with SOTA. As shown in Table 2, we conducted comparisons of performance efficiency. Due to the introduction of SFF and DFC, the efficiency is reduced compared to SAM, but our model achieves a significant performance improvement with small fine-tuning overhead. The main overhead comes from extracting the high-frequency features of images and processing the texture details and edge structure information contained therein.

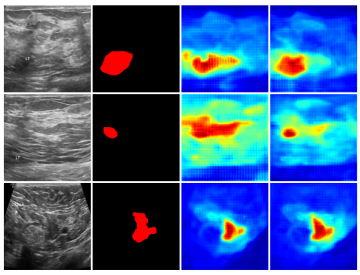

Fig. 8: Visual feature map comparison of adding SFF. (a) Input image. (b) GT. (c) w/o SFF (d) w/ SFF.

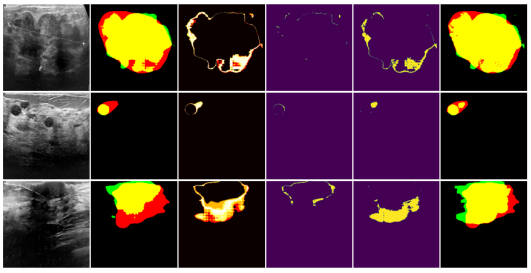

Fig. 9: Visual the correction process and prediciton comparison of adding DFC. (a) Input image. (b) w/o DFC. (c) Uncertainty map. (d) FN condition mask. (e) FP condition mask. (f) w/ DFC.

4.3 Ablation Studies

Effectiveness of SFF. The SFF module uses a feature extractor to extract high-frequency features from high frequency images, which are further integrated with spatial features to obtain more robust features. As shown in the Table 3, the performance degradation caused by removing this module shows that our SFF is effective for breast masses segmentation. The reason is that incorporating the fused spatial-frequency features into the model can form a more comprehensive feature representation, allowing SAM to explicitly perceive the structural information of the segmented targets, adapting to breast ultrasound image segmentation. Fig. 8 demonstrates the feature map visualization of our ablation study for the SFF. It shows that network pays more attention to the real lesion regions and can accurately identify the segmentation target even with low contrast under the influence of the SFF module.

Effectiveness of DFC. The DFC mainly selects high uncertainty regions in the segmentation mask by setting a threshold, and then identifies and corrects possible FN and FP regions based on the high uncertainty mask. As shown in the Table 3, the correction strategy resulted in improved segmentation performance of the test set. Fig. 9 shows the correction process and the comparison of predictions before and after correction. It can be seen that the DFC effectively corrects the potential FP and FN regions, making the final prediction are closer to the ground truth (GT) than the initial prediction.

Effect of Wavelet Bases. Fig. 7 shows the impact of high-frequency feature extraction on segmentation accuracy using different wavelet bases of Haar, Daubechies 2 (Db 2), Dmey, Coiflets 1 (Coif 1), bior1.5, and bior2.4. Comprehensive comparison shows that the prediction using Haar wavelet has higher and more stable accuracy, so we use Haar as the wavelet base for our experiments.

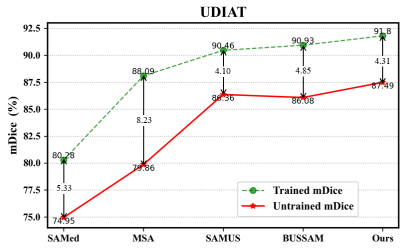

Fig. 10: Generalization ability. Untrained mDice represents mean dice of training model on BUSI and testing it on UDIAT.

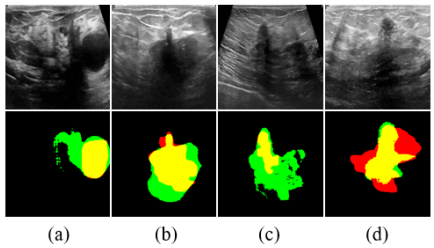

Fig. 11: Failure cases. Red, green and yellow represent ground truth, prediction and their overlapping regions, respectively.

Generalization Ability. We utilize the model trained on BUSI to test on the UDIAT dataset to evaluate the model's generalization ability. The performance comparison of different SAM-based models on the untrained UDIAT dataset is shown in Fig. 10. From the gap between the mDice of trained and untrained data of each model, it can be seen that our *SFRecSAM* not only achieve the best performance after training, but also successfully control the reduction of mDice when applied to unseen datasets. This generalization ability is attributed to the powerful feature extraction capability of SAM and the effectiveness of our method in handling the challenges of low contrast and strong noise.

Discussions and Limitations. Although we only conducted experiments on breast ultrasound images, we believe that our *SFRecSAM* can be used to analyze other datasets with similar challenges or other ultrasound images. However, as shown in the failure cases in Fig. 11, our method has some limitations. When lesions and artifacts are overly similar, SFF struggles to extract clear structural information, and DFC may consider many irrelevant artifact when correcting misclassified pixels based on pixel differences, thus limiting model performance.

5 Conclusion

In this paper, we propose a novel model *SFRecSAM* for masses segmentation in breast ultrasound images, which utilizes a spatial-frequency feature fusion (SFF) module and a dual false corrector (DFC) to make innovative improvements to SAM. The SFF fuses the extracted high-frequency features of the ultrasound image with the spatial features to achieve more powerful feature extraction capabilities, supplementing SAM with the texture details and edge structure information of masses. In addition, the DFC corrects the FP and FN regions in the initial prediction by uncertainty estimation to reduce the sensitivity of SAM to prompt's quality and improve the stability of segmentation results. With the help of the SFF and the DFC, our method improves the performance of SAM on breast ultrasound image segmentation. Extensive experiments performed on BUSI and UDIAT demonstrate the effectiveness of our proposed method.

Acknowledgements

This work was supported in part by grants from the National Natural Science Foundation of China (No. 62273241), the Natural Science Foundation of Guangdong Province, China (No. 2024A1515011946), the Shenzhen Institute of Artificial Intelligence and Robotics for Society, and the Grant under RGC Theme-based Research Scheme (No. T45-401/22-N).

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Alzahrani, Y.A.: Convolutional Neural Networks for Breast Ultrasound Image Segmentation. Ph.D. thesis, University of Windsor (Canada) (2022)
3. Borisagar, K.R., Thanki, R.M., Sedani, B.S., Borisagar, K.R., Thanki, R.M., Sedani, B.S.: Fourier transform, short-time fourier transform, and wavelet transform. Speech enhancement techniques for digital hearing aids pp. 63–74 (2019)
4. Chen, F., Chen, L., Han, H., Zhang, S., Zhang, D., Liao, H.: The ability of segmenting anything model (sam) to segment ultrasound images. BioScience Trends **17**(3), 211–218 (2023)
5. 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)
6. 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)
7. Der Kiureghian, A., Ditlevsen, O.: Aleatory or epistemic? does it matter? Structural safety **31**(2), 105–112 (2009)
8. Gao, F., Wang, X., Gao, Y., Dong, J., Wang, S.: Sea ice change detection in sar images based on convolutional-wavelet neural networks. IEEE Geoscience and Remote Sensing Letters **16**(8), 1240–1244 (2019)
9. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al.: A survey of uncertainty in deep neural networks. Artificial Intelligence Review **56**(Suppl 1), 1513–1589 (2023)
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 **42**(9), 2763–2775 (2023)
11. He, Q., Yang, Q., Xie, M.: Hctnet: A hybrid cnn-transformer network for breast ultrasound image segmentation. Computers in Biology and Medicine **155**, 106629 (2023)
12. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., Cohen-Adad, J., Merhof, D.: Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 6202–6212 (2023)
13. Hu, M., Li, Y., Yang, X.: Breastsam: A study of segment anything model for breast tumor detection in ultrasound images. arXiv preprint arXiv:2305.12447 (2023)
14. Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., Welling, M.: Supervised uncertainty quantification for segmentation with multiple annotations. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 137–145. Springer (2019)

15. Hu, X., Xu, X., Shi, Y.: How to efficiently adapt large segmentation model (sam) to medical images. arXiv preprint arXiv:2306.13731 (2023)
16. Huang, Q., Luo, Y., Zhang, Q.: Breast ultrasound image segmentation: a survey. International journal of computer assisted radiology and surgery **12**, 493–507 (2017)
17. 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)
18. Kai, H., Feng, Z.Y., Meng, H., Baoping, F.Y., Han, Y.R.: Ultrasound image segmentation of breast tumors based on swin-transformerv2. In: Proceedings of the 2022 10th International Conference on Information Technology: IoT and Smart City. pp. 106–111 (2022)
19. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
20. 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)
21. Lin, G., Chen, M., Tan, M., Chen, L., Chen, J.: A dual-stage transformer and mlp-based network for breast ultrasound image segmentation. Biocybernetics and Biomedical Engineering **43**(4), 656–671 (2023)
22. Lin, X., Xiang, Y., Zhang, L., Yang, X., Yan, Z., Yu, L.: Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. arXiv preprint arXiv:2309.06824 (2023)
23. Liu, G., Wang, J., Liu, D., Chang, B.: A multiscale nonlocal feature extraction network for breast lesion segmentation in ultrasound images. IEEE Transactions on Instrumentation and Measurement **72**, 1–12 (2023)
24. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
25. 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)
26. Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical image analysis **59**, 101557 (2020)
27. Narváez, P., Gutierrez, S., Percybrooks, W.S.: Automatic segmentation and classification of heart sounds using modified empirical wavelet transform and power features. Applied Sciences **10**(14), 4791 (2020)
28. Noble, J.A., Boukerroui, D.: Ultrasound image segmentation: a survey. IEEE Transactions on medical imaging **25**(8), 987–1010 (2006)
29. Seeböck, P., Orlando, J.I., Schlegl, T., Waldstein, S.M., Bogunović, H., Klimscha, S., Langs, G., Schmidt-Erfurth, U.: Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. IEEE transactions on medical imaging **39**(1), 87–98 (2019)
30. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F.: Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **71**(3), 209–249 (2021)
31. Tu, Z., Gu, L., Wang, X., Jiang, B.: Ultrasound sam adapter: Adapting sam for breast lesion segmentation in ultrasound images. arXiv preprint arXiv:2404.14837 (2024)

32. Upadhyay, K., Agrawal, M., Vashist, P.: Wavelet based fine-to-coarse retinal blood vessel extraction using u-net model. In: 2020 International Conference on Signal Processing and Communications (SPCOM). pp. 1–5. IEEE (2020)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
34. Wang, G., Li, W., Ourselin, S., Vercauteren, T.: Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4. pp. 61–72. Springer (2019)
35. Wang, K., Liang, S., Zhong, S., Feng, Q., Ning, Z., Zhang, Y.: Breast ultrasound image segmentation: a coarse-to-fine fusion convolutional neural network. Medical Physics **48**(8), 4262–4278 (2021)
36. Wu, G.G., Zhou, L.Q., Xu, J.W., Wang, J.Y., Wei, Q., Deng, Y.B., Cui, X.W., Dietrich, C.F.: Artificial intelligence in breast ultrasound. World Journal of Radiology **11**(2), 19 (2019)
37. Wu, J., Ji, W., Liu, Y., Fu, H., Xu, M., Xu, Y., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
38. Yao, X., Liu, H., Hu, D., Lu, D., Lou, A., Li, H., Deng, R., Arenas, G., Oguz, B., Schwartz, N., et al.: False negative/positive control for sam on noisy medical images. arXiv preprint arXiv:2308.10382 (2023)
39. Yap, M.H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A.K., Marti, R.: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics **22**(4), 1218–1226 (2017)
40. Zhang, D.: Wavelet transform. Texts in Computer Science (2021)
41. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
42. Zhang, Y., Hu, S., Jiang, C., Cheng, Y., Qi, Y.: Segment anything model with uncertainty rectification for auto-prompting medical image segmentation. arXiv preprint arXiv:2311.10529 (2023)
43. Zhao, C., Xia, B., Chen, W., Guo, L., Du, J., Wang, T., Lei, B.: Multi-scale wavelet network algorithm for pediatric echocardiographic segmentation via hierarchical feature guided fusion. Applied Soft Computing **107**, 107386 (2021)
44. Zhou, Y., Huang, J., Wang, C., Song, L., Yang, G.: Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21085–21096 (2023)
45. Zhu, X., Hu, H., Wang, H., Yao, J., Li, W., Ou, D., Xu, D.: Region aware transformer for automatic breast ultrasound tumor segmentation. In: International Conference on Medical Imaging with Deep Learning. pp. 1523–1537. PMLR (2022)
46. Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H.: A review of uncertainty estimation and its application in medical imaging. Meta-Radiology p. 100003 (2023)