IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 February 2024, accepted 18 March 2024, date of publication 2 April 2024, date of current version 12 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3384383

RESEARCH ARTICLE

SAM-Att: A Prompt-Free SAM-Related Model With an Attention Module for Automatic Segmentation of the Left Ventricle in Echocardiography

YAQI ZHU[1], CHANGCHUN XIONG[1], HENG ZHAO[1], AND YUDONG YAO[2], (Fellow, IEEE)
[1]Research Institute for Medical and Biological Engineering, Ningbo University, Ningbo 315211, China
[2]Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA

Corresponding author: Yudong Yao (yyao@ieee.org)

ABSTRACT Studying the structure and function of the heart through the left ventricle is one of the most common methods for diagnosing heart diseases. The automatic segmentation of the left ventricle can be achieved through deep learning techniques, and researchers have conducted a series of explorations in this field. Recently, the segment anything model (SAM) has achieved significant success in the field of natural images, sparking considerable interest among researchers. This has led them to investigate whether SAM can also be successfully applied in the medical imaging domain. The SAM model's interactive interface enables zero-shot and few-shot learning in the natural image domain, achieving accurate segmentation tasks. However, there are certain limitations in the automatic segmentation of medical images, specifically in the context of natural image cues such as points, boxes, and text prompts. To address this issue, this paper explores the performance of a prompt-free SAM-related model with an attention module for automatic segmentation of the left ventricle in echocardiography, named as SAM-Att. The model employs a low-rank fine-tuning strategy in the upstream, introduces an attention mechanism in the downstream, and successfully accomplishes the automatic segmentation task of the left ventricle with the support of weight files pretrained on the SAM large model. The SAM-Att model achieves dice similarity coefficient (DSC) of 94.49% and hausdorff distance (HD) of 3.505 mm on the test set. The accuracy reaches 98.83%, with precision of 93.65% and recall of 94.77%. A performance comparison of SAM-Att with other SAM-related models (SAM-b, MSA, Sam-CNN, AutoSAM, SAMed) is conducted on the same echocardiography dataset. The results indicate that the left ventricle automatic segmentation achieved the best performance when using SAM-Att.

INDEX TERMS SAM, prompt-free, echocardiography, left ventricle, automatic segmentation.

I. INTRODUCTION

The heart, as a crucial organ within the human body, plays a pivotal role in human exploration, creation, and perception of the world [1]. However, the presence of various heart diseases poses a serious threat to the lives of many people [2], [3]. In order to effectively prevent and treat these diseases, accurate computation, modeling, and analysis of the entire cardiac structure are crucial in the research and application within the medical field [4]. The prerequisite for conducting these studies is to utilize cardiac imaging to segment specific regions of the heart.

In the context of heart segmentation, the diverse morphologies and functions of each region of the heart result in distinct segmentation methods and challenges for each area. Currently, research on the heart generally focuses on the left ventricle [5]. Because the signal intensity within the right ventricle is similar to that of the myocardium, it exhibits a

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

complex crescent shape with variable changes from the base to the apex [6]. Additionally, some thinner ventricular walls may blend with surrounding tissues, increasing the difficulty of segmentation [7], [8]. In contrast, the left ventricle is characterized by a cylindrical region, with a thicker myocardial wall. The left ventricle is crucial for supplying blood to the entire body. Therefore, the segmentation of the left ventricle is typically more common in the study of heart diseases [9].

By segmenting the left ventricle, it is possible to measure the volume, wall thickness, diameter, and shape of the left ventricle, thereby assessing whether the structure of the heart is normal [10], [11]. This is crucial for the detection of structural heart diseases such as myocardial hypertrophy and dilated cardiomyopathy. The segmentation of the left ventricle enables physicians to accurately measure the left ventricle ejection fraction (EF), which represents the percentage of blood pumped out of the left ventricle with each heartbeat [12], [13]. EF is a critical parameter for assessing cardiac contractile function, playing a crucial role in the diagnosis and treatment decisions related to heart diseases [14]. The segmentation of the left ventricle can not only calculate the ejection fraction but also has other clinical significance. The segmentation of the left ventricle aids in assessing the functionality of heart valves [15], such as the mitral valve and aortic valve. It can help detect valve diseases, such as valve stenosis or regurgitation, and determine their impact on the left ventricle. The segmentation of the left ventricle can also be utilized to assess the blood supply of the coronary arteries [16]. By measuring the wall motion and thickness of the left ventricle, it is possible to detect myocardial ischemia caused by coronary artery disease. The segmentation of the left ventricle can also be used to track the therapeutic effects of heart disease treatment [10]. By continuously monitoring the structure and function of the left ventricle, physicians can assess the impact of drug therapy, interventional procedures, or other treatment measures on the heart. In summary, the segmentation of the left ventricle is a crucial step in cardiac imaging, providing healthcare professionals with valuable information to accurately assess the structure and function of the heart for diagnostic and treatment decisions. The most common method currently used for left ventricle segmentation is employing deep learning techniques for automatic segmentation.

In recent years, there has been a series of studies on the automatic segmentation of the left ventricle in echocardiography using deep learning. Liu et al. [17] proposed the deep pyramid local attention neural network (PLANet), which enhances features by capturing supportive information in adjacent contexts. It employs a label consistency learning mechanism to improve the accuracy of pixel prediction. This model is capable of automatically segmenting cardiac structures in 2D echocardiography. Ali et al. [18] proposed a model named ResU, which combines the advantages of ResNet and U-Net. This model exhibits significant advantages in denoising data and can automatically segment the left ventricle. Amer et al. [19] proposed a novel deep learning segmentation method based on U-Net, named ResDUnet. This approach replaces the basic U-Net blocks with residual blocks incorporating squeeze-and-excitation units featuring adaptive channel-wise features. The model achieves automatic segmentation of the left ventricle. Most recently, the segment anything model (SAM) [20] has achieved impressive results in natural image segmentation tasks. Soon after, the research on large-scale models like SAM has reached a new pinnacle in the field of medical image [21]. Many scholars have made improvements to the original SAM, striving to better apply SAM philosophy in medical image segmentation. However, due to the complex morphology, detailed anatomical structures, and uncertain object boundaries in medical images, especially with the limitation of the need of prompts in the SAM encoder, the operator expertise has significant impact in the medical image segmentation [22]. This makes SAM more challenging for medical image segmentation.

This paper investigates several models related to SAM. The segment anything model [20], proposed by the Mata AI Lab team, possesses zero-shot transferability and employs the standard vision transformer (ViT) as the image encoder. There are three variants of ViT utilized in SAM, namely ViT-b, ViT-h, and ViT-l. The model composed of these three types of image encoders is referred to in this paper as SAM-b, SAM-h, and SAM-l. The three models are pretrained on the SA-1B dataset, and corresponding weight files are obtained for each model [20]. Wu et al. [23] proposed the medical SAM adapter (MSA), which integrates domain-specific knowledge segmentation models through a simple yet effective adaptation technique. The MSA demonstrates excellent performance in medical image segmentation tasks across 19 different image modalities. Hu et al. [24] froze the SAM image encoder and explored three types of prompt-free prediction heads based on this, including AutoSAM (removing prompts from the masked decoder of SAM), convolutional neural networks (Sam-CNN), and linear layers. The results demonstrate that when there is a shortage of labels in the dataset, AutoSAM and the Sam-CNN prediction head outperform training from scratch and self-supervised learning methods in terms of segmentation accuracy. Zhang and Liu [25] applied a fine-tuning strategy based on low-rank attention (LoRA) [26] to the SAM image encoder, named as SAMed. After freezing the image encoder, additional trainable LoRA layers were inserted into SAM for feature extraction in medical images. Then, fine-tune the prompt encoder with default embeddings and a masking decoder to achieve precise semantic segmentation of medical images. In this paper, the emphasis is placed on the performance of the model for image segmentation under prompt-free conditions. Applying the above model to the echocardiography dataset under the same conditions, and then making improvements to enhance the model's

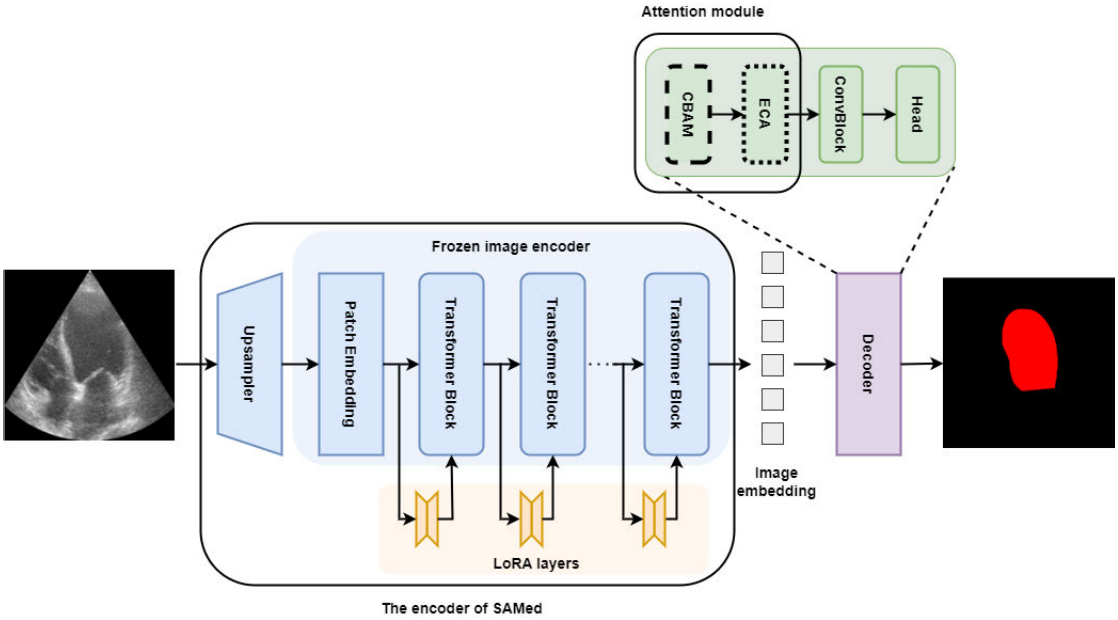


FIGURE 1. Overview of the proposed model (SAM-Att).

performance through adding an attention module in the SAM (SAM-Att). The main tasks and contributions of this paper are as follows.

1) Introducing a SAM-Att model. Building upon the existing SAMed model, the upstream encoder continues to utilize a low-rank fine-tuning strategy, updating the parameters of the Transformer modules. The downstream section removes the prompting encoder and modifies the decoder.
2) Introducing the convolutional block attention module (CBAM) and efficient channel attention (ECA) network in the decoder of the SAM-Att model. Adjusting the weights through learning to better capture the inter-relationships between different channels or among channels, enabling a more effective exploration of correlations between features and focusing on relevant features.
3) Comparing the segmentation performance of SAM-Att model with SAM-b, MSA, Sam-CNN, AutoSAM, and SAMed models on echocardiography.

II. METHOD

For a given echocardiography $x \in R^{512\times512\times3}$, with a spatial resolution of 512 × 512, and three channels. After passing through the model, the output is a segmented image $\hat{S}$ with the corresponding dimensions. The definition of the category corresponding to each pixel is given by $Y = \{y_0, y_1\}$, where y_0 represents the background category, and y_1 represents the left ventricle endocardium category. The overall architecture of the model inherits from SAMed, as illustrated in Figure 1.

The encoder part in the diagram is frozen, and LoRA is employed to update the parameters of the transformer module. Firstly, compress the transformer features into a low-rank space, then reproject the compressed features, aligning them with the output feature channels of the frozen transformer block. In order to better study the automatic segmentation effects of SAM-related algorithms in echocardiography, we maintain consistency throughout the entire experimental process. The present study does not require a prompt encoder, which will facilitate the automation of medical diagnosis. The decoder section differs from SAMed in this study. The decoder will consist of four components: convolutional block attention module, efficient channel attention network, upsampling convolution module, and segmentation head.

A. ENCODER

The encoder part of the model is inspired by the research conducted by Zhang et al. [25] Initially, the image encoder is frozen, and subsequently, a fine-tuning strategy based on low-rank updates is introduced, allowing parameter updates for the transformer modules. Due to the SAM being trained on large-scale datasets, it possesses excellent feature extraction capabilities. Therefore, it is worth exploring the use of the SAM to guide downstream tasks. The introduction of LoRA enables the SAM to utilize newly acquired knowledge during the training process of echocardiography, allowing for the updating of a small portion of parameters [27]. This not only saves computational costs but also reduces the deployment and storage challenges during the model fine-tuning process, while ensuring segmentation performance. The working principle of LoRA is illustrated in Figure 2. For a pre-trained matrix $W_0 \in R^{d\times k}$, matrix parameter updates can be performed through low-rank decomposition,

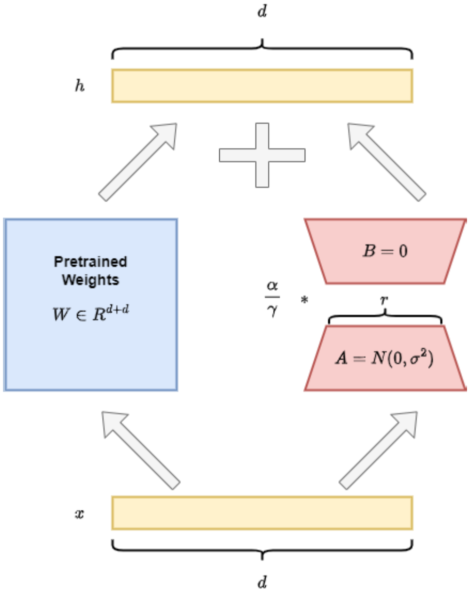


FIGURE 2. LoRA module.

the formula is as follows,

$$W_0 + \Delta = W_0 + BA \tag{1}$$

where $B \in R^{d\times k}$, $A \in R^{r\times k}$, and $r \ll \min(d, k)$, with the dimension of r significantly smaller than that of d and k. When conducting fine-tuning training, the parameters of W_0 remain unchanged, and only the parameters for learning A and B are updated. Initialize matrix A with a Gaussian distribution, and initialize matrix B with zeros. During the training process, constrain the update ΔW using $\frac{\alpha}{\gamma}$, where α is adjusted during training similar to learning rate.

B. DECODER

The SAM model's decoder consists of lightweight transformer layers and segmentation heads. SAMed transforms SAM's ambiguous segmentation heads into deterministic output heads and employs LoRA for fine-tuning. This study will utilize convolutional modules for upsampling, eliminating the need for prompts to make the operation more convenient. To enhance the model's focus on key feature information, an attention mechanism is introduced to learn the importance weights on specific positions and channels, thereby improving the quality of image reconstruction. The initial portion is also adjusted appropriately. Through three layers of pixel-wise convolution operations, the mapping from feature maps to segmentation masks is achieved, enhancing the accuracy of segmentation.

1) CONVOLUTIONAL BLOCK ATTENTION MODULE

Convolutional block attention module consists of two submodules, namely the channel attention module and the spatial attention module, which respectively perform attention operations on the channel and spatial dimensions [28]. As shown in Figure 3, the input feature $F \in R^{256\times32\times32}$ enters the module. Firstly, it undergoes a one-dimensional convolution in the channel attention module, resulting in $M_c \in R^{256\times1\times1}$. The output of this operation is multiplied element-wise with the input feature F, yielding the output F' of the channel attention module. Subsequently, F' undergoes a two-dimensional convolution in the spatial attention module, resulting in $M_s \in R^{2\times7\times7}$. The output of this operation is then multiplied element-wise with F', resulting in the final output F'' of the spatial attention module. This process is illustrated by the following formula,

$$F' = M_c(F) \otimes F \tag{2}$$

$$F'' = M_s(F') \otimes F' \tag{3}$$

Channel attention modules compress spatial dimensions only without altering channel dimensions. The green box in Figure 4 illustrates the workflow of this module. Firstly, the input feature map undergoes two parallel operations: MaxPool and AvgPool layers. This transforms the feature map's dimensions from 256 × 32 × 32 to 256 × 1 × 1. Then, after passing through the Shared MLP (multi-layer perceptron) module, the channel count is first compressed to one-fourth of its original value to reduce computational and storage complexity. Subsequently, the compressed features are processed using the rectified linear unit (ReLU) to learn the importance of inter-channel relationships. Finally, the channel count is expanded back to its original value. Adding the two output results element-wise, then applying a sigmoid activation function to obtain the channel attention's output. Finally, multiplying this output with the original image to restore it to a size of 256 × 32 × 32. The formula for channel attention is as follows,

$$M_c(F) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F)) \tag{4}$$

The spatial attention module compresses only the channel dimension without altering the spatial dimension. The purple box in Figure 4 illustrates the workflow of this module. Firstly, take the output of channel attention and obtain two 1 × 32 × 32 feature maps through max pooling and average pooling. Then, concatenate the two feature maps through the concat operation, and pass the concatenated feature map through a 7 × 7 convolutional layer to obtain a feature map with a channel number of 1. The process involves passing through a sigmoid function to obtain the feature map of spatial attention. Finally, this map is multiplied with the input results to restore it to a size of 256 × 32 × 32. The spatial attention formula is as follows,

$$M_s(F) = \sigma(f^{7\times7}([\mathrm{AvgPool}(F); \mathrm{MaxPool}(F)])) \tag{5}$$

2) EFFICIENT CHANNEL ATTENTION NETWORK

The efficient channel attention module is a non-reductive local cross-channel interaction strategy [29]. Appropriate cross-channel interaction can significantly reduce the complexity of the model while maintaining performance. The working principle of the ECA is illustrated in Figure 5. First,

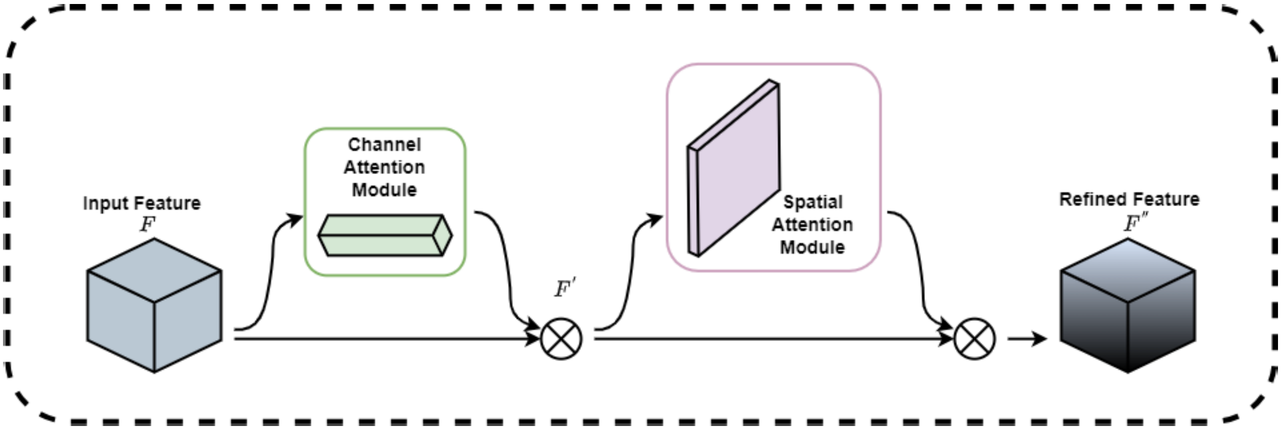


FIGURE 3. Convolutional block attention module.

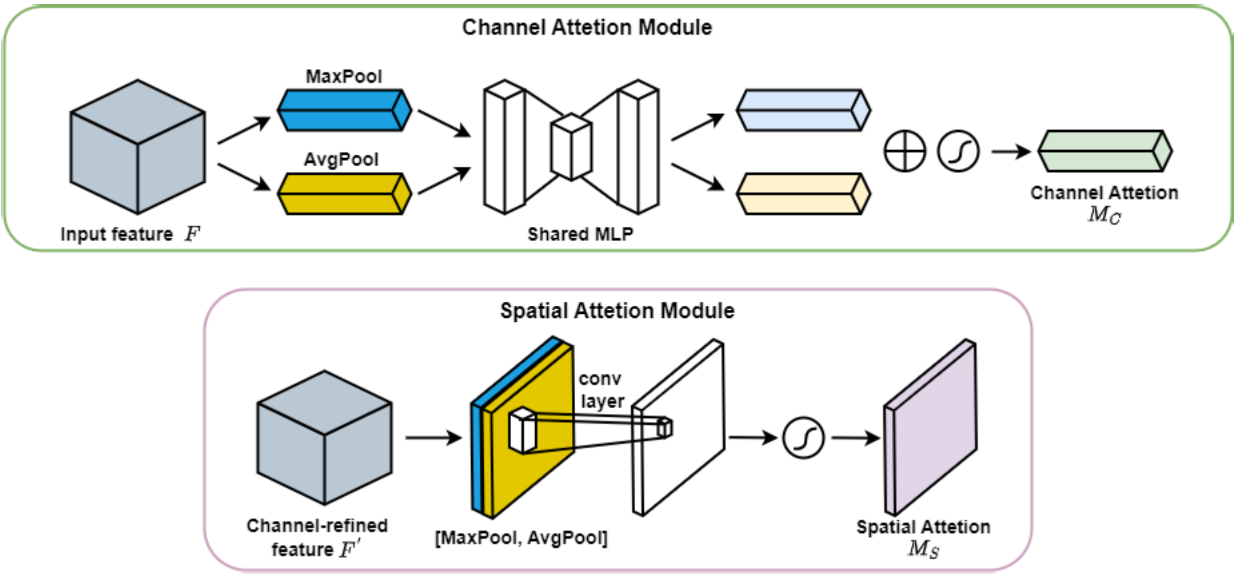


FIGURE 4. Channel attention module and spartial attention module.

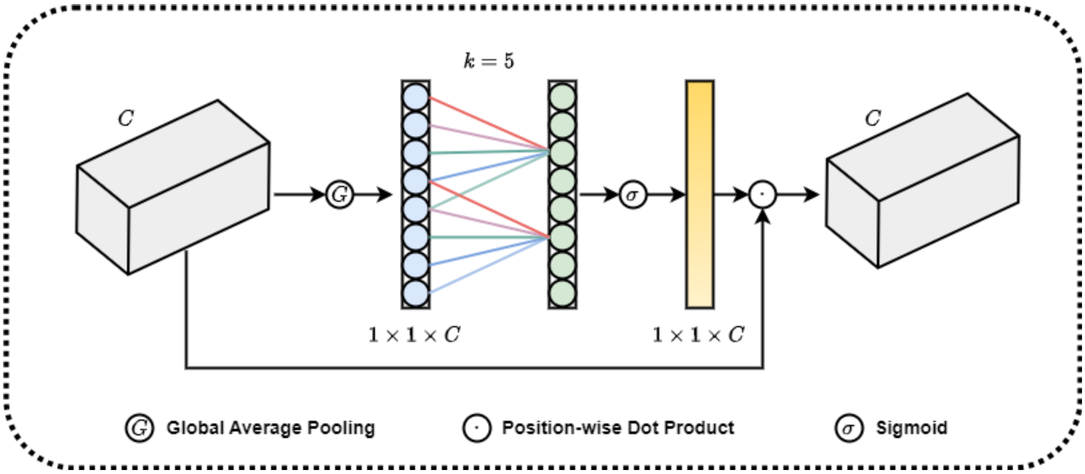


FIGURE 5. Efficient channel attention network.

compress spatial information of the input feature map with dimensions 32 × 32 × 256 through global average pooling, resulting in an 1 × 1 × 256 feature map. Then, use an 1 × 1 convolution to learn the importance between different channels, outputting an 1 × 1x256 feature map. The final step involves element-wise multiplication with the input feature map of size 32 × 32 × 256. This results in a feature map with channel-wise attention.

In the process of channel feature learning, the ECA employs dynamic convolutional kernels. The size of its kernel undergoes adaptive changes through a function, with larger convolutional kernels employed in layers with a higher number of channels. This facilitates greater cross-channel interaction through the use of 1 × 1 convolutions. Using smaller convolutional kernels in layers with fewer channels, performing 1 × 1 convolutions, minimizes inter-channel interactions to a lesser extent. The size of the convolutional kernel, denoted as k, is defined by the following formula,

$$k = \psi(C) = |\frac{\log_2 C}{\gamma} + \frac{b}{\gamma}|_{\text{odd}} \tag{6}$$

In the context, C represents the number of channels, $||_{\text{odd}}$ indicates that k can only take odd values, and γ and b are set to 2 and 1, respectively, in the current research process, to adjust the ratio between the number of channels C and the convolutional kernel size k.

3) UPSAMPLING CONVOLUTION MODULE

In this module, the first step involves a transposed convolutional layer, which enlarges the dimensions of the input feature map to twice its original size and reduces the channel count from 256 to 64. Then, apply layer normalization to normalize the output, stabilize the training process, and enhance the model's generalization ability. The process continues by applying the gausssian error linear unit (GELU) activation function for non-linear transformation. Subsequently, the output undergoes another transposed convolutional layer, akin to the first transposed convolutional layer. However, in this layer, the number of channels in the input features is reduced from 64 to 32, while simultaneously doubling its spatial dimensions. Finally, implement the upsampling operation.

4) SEGMENTATION HEAD

Before outputting the segmentation categories, add two layers of convolutional modules. The first convolutional module reduces the number of channels from 32 to 8, with a convolutional kernel size of 3. The second convolutional module maintains the same number of channels, while the third convolutional module transforms the number of channels from 8 to 2, effectively segregating the pixel categories.

C. LOSS FUNCTION

SAM-Att fine-tunes using the cross-entropy loss function during the training process [30]. The principle formula for the loss function is as follows,

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)] \tag{7}$$

In the above, $y_i \in \{0, 1\}$ represents the true label of the i-th sample, and $\hat{y}_i \in \{0, 1\}$ represents the predicted value of the neural network model for the i-th sample. The smaller the value of the loss function, the smaller the difference between the true probability distribution and the predicted probability distribution.

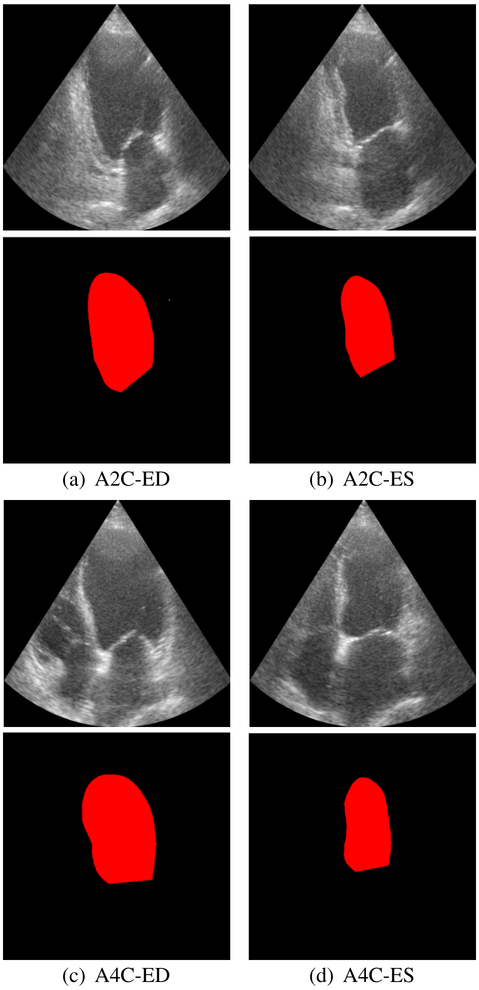

(a) A2C-ED
(b) A2C-ES
(c) A4C-ED
(d) A4C-ES

FIGURE 6. Echocardiography and labels in various views and states.

III. EXPERIMENT

A. DATASET

This study utilized the cardiac acquisition for multi-structure ultrasound segmentation (CAMUS) dataset, consisting of end-systolic and end-diastolic frames from both apical two-chamber and apical four-chamber views [31]. A2C-ED represents the end-diastolic state of the left ventricle in the apical two-chamber view, A2C-ES represents the end-systolic state of the left ventricle in the apical two-chamber view, A4C-ED represents the end-diastolic state

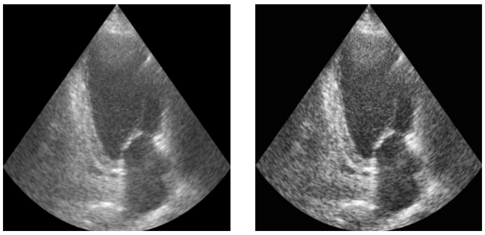


FIGURE 7. Comparison of image enhancement before and after using CLAHE function.

of the left ventricle in the apical four-chamber view, A4C-ES represents the end-systolic state of the left ventricle in the apical four-chamber view. As shown in Figure 6, the 1800 ultrasound images for this experiment will be randomly divided into training, validation, and test sets in a ratio of 6:2:2, resulting in 1080 training images and 360 images each for both the validation and test sets. Before training the input model, preprocess the images by first computing the histogram of each image. Perform histogram equalization on local regions of the image based on the histogram distribution, thereby enhancing local details. The comparison between the image before and after enhancement is shown in Figure 7. Due to variations in the original sizes of collected echocardiography, the images are first resized to a uniform 550 × 600 dimensions using interpolation. Subsequently, random cropping is applied to achieve a final size of 512 × 512. Due to the limitation on the number of images, it is decided to augment data diversity by applying horizontal and vertical flips to the images with a certain probability. The above operations provide the model with more variations and perspectives, thereby enhancing the model's generalization ability. The validation and test sets did not undergo random cropping and flipping operations. The image size is kept consistent with the input size of the model during training, directly adjusted to 512 × 512.

B. TRAINING CONFIGURATION

The model training workstation is equipped with an Intel Core i9-10900X CPU and a NVIDIA RTX 3090 GPU with 24GB of VRAM. During the training process, the initial learning rate is set to 0.001, the batch size is 8, and the number of training epochs is set to 250. The optimizer employs the adaptive moment estimation (Adam) algorithm, which utilizes an adaptive learning rate mechanism capable of automatically adjusting the learning rates of parameters. This feature facilitates faster convergence to the global minimum during the training process. In this experiment, the learning rate will decrease by 50% every 30 training epochs. After the training is complete, the weights that exhibit the best performance will be saved to a file for subsequent testing.

C. EVALUATION METRICS

Dice similarity coefficient (DSC): This is a measure for quantifying the similarity between two sets, with values ranging from 0 to 1 [32]. A higher value indicates a greater similarity between the two sets. The formula is as follows,

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{8}$$

Hausdorff distance (HD): It is used to measure the maximum distance between two segmentation results, it reflects the inconsistency of the boundaries in the segmentation results [33]. A smaller value of HD indicates a closer segmentation result. The formula is as follows,

$$\mathrm{HD} = \max(h(A, B), h(B, A)) \tag{9}$$

In this context, $h(A, B)$ represents the distance from each point in set A to the nearest point in set B, $h(B, A)$ represents the distance from each point in set B to the nearest point in set A, and max denotes the maximum distance between the two point sets.

Accuracy: The ratio of correctly classified pixels to the total number of pixels in the segmentation results. The formula is as follows,

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

In the context provided, TP (true positive) represents the number of pixels correctly classified as the target, TN (true negative) represents the number of pixels correctly classified as the background, FP (false positive) represents the number of pixels incorrectly classified as the target, and FN (false negative) represents the number of pixels incorrectly classified as the background. The accuracy value ranges from 0 to 1, with a higher value indicating higher classification accuracy.

Precision: The proportion of pixels predicted by the model to be the target that actually belong to the target. The formula is as follows,

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{11}$$

Precision measures the accuracy of a model in predicting the target. A higher value indicates that the model is less likely to incorrectly classify background pixels as the target.

Recall: The model successfully predicts the proportion of pixels that are true positives for the target within the actual target pixels. The formula is as follows,

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

Recall measures the model's ability to recognize the target. A higher value indicates that the model is less likely to incorrectly classify target pixels as background.

IV. RESULTS

In this study, the performance comparison of six models (SAM-b [20], MSA [23], Sam-CNN [24], AutoSAM [24], SAMed [25], SAM-Att), as well as the optimal mechanism of incorporating attention modules, is conducted by keeping the same variables and utilizing the aforementioned 20% test set.

TABLE 1. The comparison of segmentation performance among different types of image encoders in the SAM model.

Model	DSC	HD (mm)	Accuracy	Precision	Recall
SAM-b	68.54	43.003	93.54	64.76	79.69
SAM-l	69.60	41.242	94.07	69.36	76.10
SAM-h	69.97	43.007	93.79	66.83	77.86

TABLE 2. Compare the ablative experimental results of SAMed-noP baseline model with the addition of an attention module.

Model	DSC	HD (mm)	Accuracy	Precision	Recall
SAMed-noP	92.29	4.943	98.61	91.19	93.15
SAMed-noP + CBAM	92.36	4.579	98.62	92.04	93.33
SAMed-noP + ECA	92.40	4.715	98.63	91.37	94.15
SAMed-noP + ECA +CBAM	93.32	3.838	98.81	92.77	94.37
SAMed-noP + CBAM +ECA (SAM-Att, proposed)	**93.49**	**3.505**	**98.83**	**93.65**	**94.77**

TABLE 3. Compare the performance of different SAM-related models.

Model	DSC	HD (mm)	Accuracy	Precision	Recall
SAM-b	68.54	43.003	93.54	64.76	79.69
MSA	68.83	40.764	93.98	69.43	74.32
Sam-CNN	87.70	18.531	97.85	89.80	87.18
AutoSAM	89.12	9.428	98.09	90.60	88.45
SAMed	92.47	4.502	98.65	93.20	92.36
SAM-Att (proposed)	**93.49**	**3.505**	**98.83**	**93.65**	**94.77**

A. ABLATION EXPERIMENT

Before conducting ablation experiments, the segmentation performance of three models (SAM-b, SAM-l, and SAM-h) composed of three types of image encoders (ViT-b, ViT-l, and ViT-h) was compared in echocardiography datasets. The specific results are shown in Table 1.

From Table 1, it can be observed that the SAM-h model has the highest DSC value among the three models. The SAM-l model has the lowest HD value among the three models, while the HD values of SAM-b and SAM-h models are relatively close. The SAM-l model has the highest accuracy and precision among the three models. The SAM-b model exhibits the highest recall value among the three models. The overall performance of the three models is quite similar. But Hu et al. conducted a comparison of the performance of the SAM model using three image encoders, ViT-b, ViT-l, and ViT-h, in the process of skin cancer segmentation. They found that fine-tuning the model with ViT-b as the encoder resulted in better performance [34]. Therefore, all the improvement experiments on the SAM model in this study are based on the ViT-b image encoder.

In this research, the aim is to explore the impact of attention modules on the performance of SAM. This paper conducts ablation experiments to analyze their effects. This study primarily focuses on two different attention mechanisms: CBAM and ECA. The SAMed is modified by removing the prompt encoder and replacing the decoder with transposed convolution. This modified model is named SAMed-noP and serves as the baseline model. Then, introduce CBAM and ECA separately, named SAMed-noP + CBAM and SAMed-noP + ECA, respectively. Afterward, introducing both modules simultaneously, named SAMed-noP + ECA + CBAM based on the order of introduction, and SAMed-noP + CBAM + ECA. The ablation experiment results of introducing CBAM and ECA networks into the SAM-Att model decoder are shown in Table 2.

The first row of Table 2 represents the test results in the baseline model, indicating the performance without any attention modules added. The second-line model incorporates CBAM on top of the baseline model. It can be observed that the DSC value for the model with only CBAM is 92.36, the HD value is 4.579, the accuracy is 98.62, the precision is 92.04, and the recall is 93.33. The performance shows a slight improvement compared to the baseline model. The third-line model incorporates an ECA on top of the baseline model. This model exhibits greater improvements in DSC, accuracy, and recall values compared to a model with only CBAM added. However, the improvements in the other two metrics are not as pronounced as those observed in the model with only CBAM. The fourth-line model undergoes ECA first and then CBAM. It exhibits higher HD value, accuracy, precision, and recall compared to models with individual introductions of these techniques. The fifth-layer model first undergoes CBAM and then ECA. All four evaluation metrics are higher than those of the previous models.

B. THE COMPARATIVE RESULTS WITH OTHER MODELS

To enhance the persuasiveness of the experimental results in this study, the performance of other SAM-related models will be compared within the same dataset. Table 3 shows the segmentation accuracy and precision of six models: SAM-b, MSA, Sam-CNN, AutoSAM, SAMed, and SAM-Att (proposed).

From Table 3, it can be observed that the segmentation results of the proposed original SAM model and the model with added adapters are not satisfactory. The DSC values are only in the sixties, while the HD values exceed forty. The accuracy values are around ninety-three percent, and precision values are in the sixties, with recall values in the seventies. The performance of the Sam-CNN model, integrated into the CNN network, exhibits a significant improvement compared to the previous two models. The DSC has increased from just over 60% to over 80%, and the HD has decreased from over 40 millimeters to just over 10 millimeters. The AutoSAM has shown significant improvement in HD values for the deletion mask decoder prompt, while the improvement in other evaluation metrics is relatively small. The SAMed strategy, based on low-rank fine-tuning, has also made contributions in terms of performance improvement. The SAM-Att proposed in this study exhibits outstanding performance, with DSC of 93.49%, HD of only 3.505, accuracy of 98.83%, precision of 93.65%, and recall of 94.77%.

To visually assess the segmentation results, an echocardiography is selected for prediction. Figure 8 displays the segmentation results of six different models on this echocardiography. (a) and (b) are the original images of the echocardiography for testing, along with their corresponding labels. (c) to (h) represent the result images obtained from the segmentation of six different models. It can be observed that the results obtained by the SAM-b and MSA for segmentation are the worst, with almost no segmentation achieved. The results segmented by the Sam-CNN, AutoSAM, and SAMed exhibit a general similarity in structure to the labels, but there is room for improvement, particularly in the edge regions. The segmentation results obtained from the SAM-Att model are closest to the ground truth labels, and the edges are also smoother.

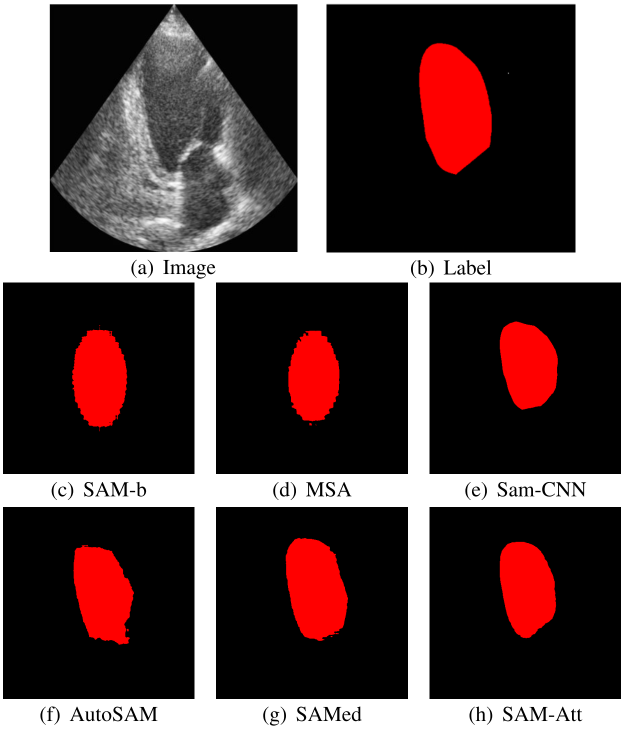

(a) Image
(b) Label
(c) SAM-b
(d) MSA
(e) Sam-CNN
(f) AutoSAM
(g) SAMed
(h) SAM-Att

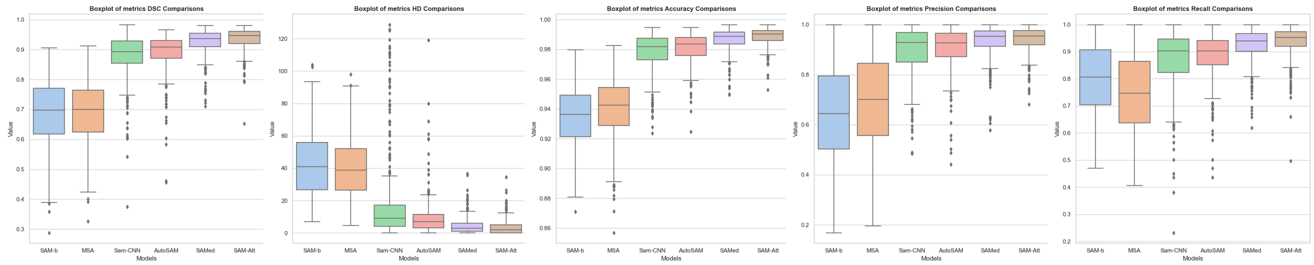
FIGURE 8. Segmentation results of different SAM-related models. (a) and (b) respectively represent the original echocardiography image and its label. (c) to (h) represent segmentation result images of six different SAM-related models for echocardiography.

V. DISCUSSION

Our experiments improve upon SAMed based on a low-rank fine-tuning strategy and optimize the decoder section to automatically output segmentation results without the need for prompts. This enhancement makes the model more widely applicable in the field of medical image segmentation, with a more convenient operation for a broader user base. The encoder part of the model adopts the low-rank fine-tuning strategy from the SAMed model, where the encoder is frozen, and fine-tuning is applied to update the parameters of the transformer modules. This ensures the effectiveness of feature information extraction. The decoder utilizes convolutional modules to achieve the up-sampling of feature maps, incorporating both CBAM and ECA attention mechanisms. This allows the model to learn importance weights for specific locations and channels, enabling the decoder to better focus on crucial features and consequently enhancing the quality of image reconstruction. The final improvement involves refining the segmentation head by employing three layers of convolutional operations to map the feature maps to a segmentation mask of the same size as the input image. Through pixel-wise convolution operations, the network is able to model the boundaries and details of both the background and segmented foreground in the feature space, thereby enhancing the accuracy of segmentation.

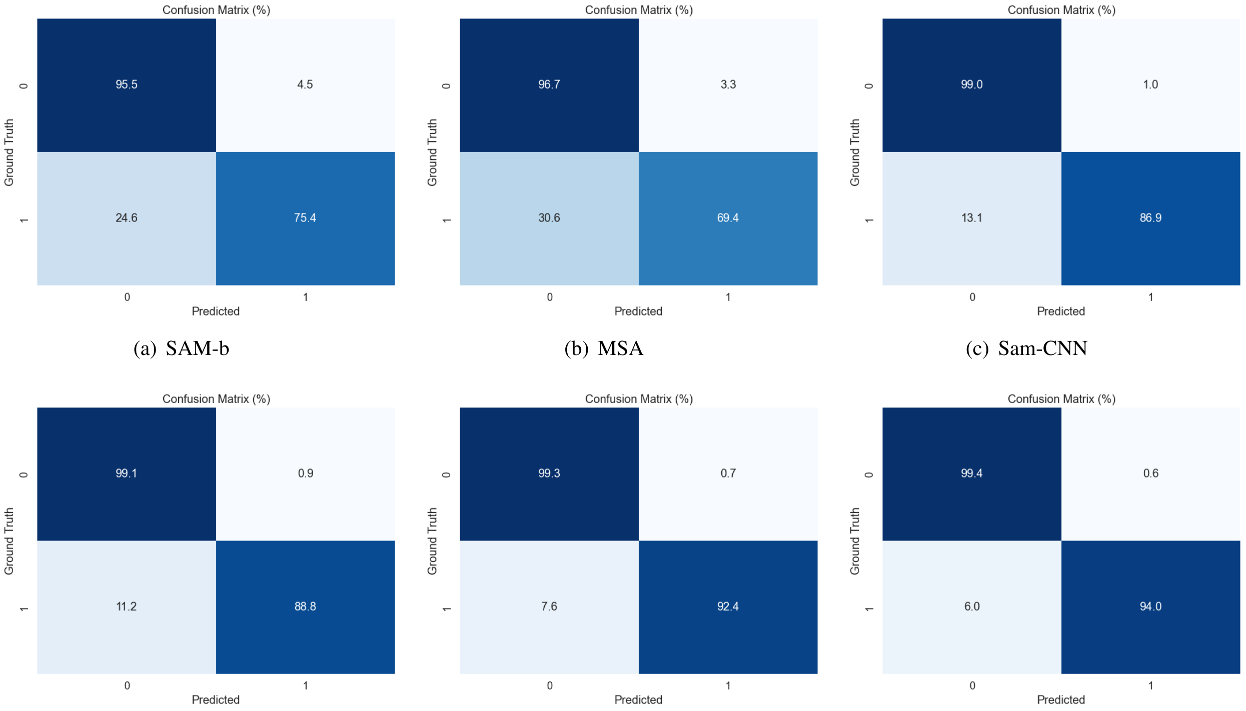
The boxplots in Figure 9 illustrate the variation in various evaluation metrics across different models. The horizontal axis of each boxplot represents the six models for comparison, while the vertical axis represents the numerical values obtained in the test set. Figure (a) displays boxplot of DSC values for different models. It is seen that the median values under SAM-b and MSA models are worse than the other four, and the data exhibit higher variability. In the SAM-Att model, the median is the highest, and the data dispersion is the smallest. Figure (b) displays boxplot of HD values for different models. It can be observed that the median values under the SAM-b and MSA models are higher than the other four, and the data also exhibits greater dispersion. The SAM-Att model has the smallest median and also the smallest data dispersion. The boxplot in Figures (c) to (e) depict the performance of various models in terms of accuracy, precision, and recall values. It is seen that both SAM-b and MSA exhibit worse performance across all three evaluation metrics compared to the other four models. Conversely, the SAM-Att outperforms the other models in terms of accuracy, precision, and recall across the board.

The confusion matrix in Figure 10 illustrates the pixel-wise classification performance of different models on the test set. The horizontal axis of each confusion matrix represents the predicted values, while the vertical axis represents the



(a) DSC (b) HD (c) Accuracy (d) Precision (e) Recall

FIGURE 9. The boxplots of evaluation metrics under different models. (a) to (e) respectively depict comparisons of evaluation metrics DSC, HD, Accuracy, Precision, and Recall.



(a) SAM-b (b) MSA (c) Sam-CNN

(d) AutoSAM (e) SAMed (f) SAM-Att

FIGURE 10. The confusion matrix for different models during execution.

true values. Among them, 0 represents the background, and 1 represents the left ventricular endocardium to be segmented. It can be observed that the SAM-b model has a percentage of 24.6% of pixels where the true value is 1 and the predicted value is 0, a percentage of 4.5% where the true value is 0 and the predicted value is 1, and a percentage of 75.4% where the true value is 1 and the predicted value is also 1. The MSA model has a 30.6% proportion of pixels where the true value is 1 while the predicted value is 0, a 3.3% proportion where the true value is 0 while the predicted value is 1, and a 69.4% proportion where both the true and predicted values are 1. The Sam-CNN model has a pixel misclassification rate of 13.1% for pixels where the ground truth is 1 and the predicted value is 0. For pixels where the ground truth is 0 and the predicted value is 1, the misclassification rate is 1.0%. Additionally, the model achieves an accuracy of 86.9% for pixels where both the ground truth and predicted values are 1. The AutoSAM model has a pixel misclassification rate of 11.2% for pixels where the ground truth is 1 but the predicted value is 0, and a misclassification rate of 0.9% for pixels where the ground truth is 0 but the predicted value is 1. Additionally, the percentage of pixels where both the ground truth and the predicted value are 1 is 88.8%. The SAMed model has a 7.6% proportion of pixels where the true value is 1 but the predicted value is 0, a 0.7% proportion where the true value is 0 but the predicted value is 1, and a 92.4% proportion where both the true and predicted values are 1. The SAM-Att model exhibits excellent performance in pixel-wise classification compared to five other models. The percentage of pixels with true values of 1 and predicted values of 0 is

6.0%, while the percentage of pixels with true values of 0 and predicted values of 1 is 0.6%. Additionally, the percentage of pixels with true values of 1 and predicted values of 1 is 94%.

VI. CONCLUSION

This paper builds upon the existing SAMed model, maintaining the use of a low-rank fine-tuning strategy for the upstream encoder. During this phase, the transformer component undergoes parameter updates to enhance its applicability for ultrasound image segmentation tasks. In the downstream process, remove the prompt encoding section to implement automatic segmentation functionality that is more tailored to the practical situations in medical image segmentation. The introduced CBAM and ECA attention mechanisms in the decoder section further enhance the model's performance. In the CAMUS echocardiography dataset, this study compares the segmentation performance of the proposed SAM-Att model with five other SAM-related models under the prompt-free condition. The results indicate that the automatic segmentation of the left ventricle by SAM-Att is the most effective.

While achieving high segmentation accuracy and precision, there is no doubt that training speed is compromised when using weight files pre-trained on such a large-scale dataset. The models related to SAM still face significant challenges in the field of medical image segmentation, which holds crucial value for researching heart diseases.

REFERENCES

[1] C. Renowden, T. Beer, and L. Mata, "Exploring integrated artscience experiences to foster nature connectedness through head, heart and hand," *People Nature*, vol. 4, no. 2, pp. 519–533, Apr. 2022.

[2] W.-Y. Low, Y.-K. Lee, and A. L. Samy, "Non-communicable diseases in the Asia–Pacific region: Prevalence, risk factors and community-based prevention," *Int. J. Occupational Med. Environ. Health*, vol. 28, no. 1, pp. 20–26, 2015.

[3] G. Sun, X. Li, J. Wei, T. Zhang, B. Li, M. Chen, Y. Wang, K. Chen, and Y. Li, "Pharmacodynamic substances in salvia miltiorrhiza for prevention and treatment of hyperlipidemia and coronary heart disease based on lipidomics technology and network pharmacology analysis," *Biomed. Pharmacotherapy*, vol. 141, Sep. 2021, Art. no. 111846.

[4] A. Quarteroni, T. Lassila, S. Rossi, and R. Ruiz-Baier, "Integrated heart—Coupling multiscale and multiphysics models for the simulation of the cardiac function," *Comput. Methods Appl. Mech. Eng.*, vol. 314, pp. 345–407, Feb. 2017.

[5] M. A. Shoaib, J. H. Chuah, R. Ali, K. Hasikin, A. Khalil, Y. C. Hum, Y. K. Tee, S. Dhanalakshmi, and K. W. Lai, "An overview of deep learning methods for left ventricle segmentation," *Comput. Intell. Neurosci.*, vol. 2023, pp. 1–26, Jan. 2023.

[6] C. Petitjean et al., "Right ventricle segmentation from cardiac MRI: A collation study," *Med. Image Anal.*, vol. 19, no. 1, pp. 187–202, Jan. 2015.

[7] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi, "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging," *Magn. Reson. Mater. Phys., Biol. Med.*, vol. 29, no. 2, pp. 155–195, Apr. 2016.

[8] D. Mahapatra and J. M. Buhmann, "Automatic cardiac RV segmentation using semantic information with graph cuts," in *Proc. IEEE 10th Int. Symp. Biomed. Imag.*, Apr. 2013, pp. 1106–1109.

[9] E. Smistad and A. Østvik '2D left ventricle segmentation using deep learning," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Sep. 2017, pp. 1–4.

[10] K. Yamamoto, M. M. Redfield, and R. A. Nishimura, "Analysis of left ventricular diastolic function," *Heart*, vol. 75, no. 6, pp. 27–35, Jun. 1996.

[11] S. F. Nagueh, O. A. Smiseth, C. P. Appleton, B. F. Byrd, H. Dokainish, T. Edvardsen, F. A. Flachskampf, T. C. Gillebert, A. L. Klein, P. Lancellotti, P. Marino, J. K. Oh, B. A. Popescu, and A. D. Waggoner, "Recommendations for the evaluation of left ventricular diastolic function by echocardiography: An update from the American society of echocardiography and the European association of cardiovascular imaging," *J. Amer. Soc. Echocardiography*, vol. 29, no. 4, pp. 277–314, Apr. 2016.

[12] Y. Lan and R. Jin, "Automatic segmentation of the left ventricle from cardiac MRI using deep learning and double snake model," *IEEE Access*, vol. 7, pp. 128641–128650, 2019.

[13] M. Alkhodari, H. F. Jelinek, S. Saleem, L. J. Hadjileontiadis, and A. H. Khandoker, "Revisiting left ventricular ejection fraction levels: A circadian heart rate variability-based approach," *IEEE Access*, vol. 9, pp. 130111–130126, 2021.

[14] P. L. M. Kerkhof, P. M. van de Ven, B. Yoo, R. A. Peace, G. R. Heyndrickx, and N. Handly, "Ejection fraction as related to basic components in the left and right ventricular volume domains," *Int. J. Cardiology*, vol. 255, pp. 105–110, Mar. 2018.

[15] R. A. Nishimura, C. M. Otto, R. O. Bonow, B. A. Carabello, J. P. ErwinIII, R. A. Guyton, P. T. O'Gara, C. E. Ruiz, N. J. Skubas, P. Sorajja, T. M. SundtIII, and J. D. Thomas, "2014 AHA/ACC guideline for the management of patients with valvular heart disease: Executive summary: A report of the American college of cardiology/American heart association task force on practice guidelines," *Circulation*, vol. 129, no. 23, pp. 2440–2492, 2014.

[16] W. C. Little and T. R. Downes, "Clinical evaluation of left ventricular diastolic performance," *Prog. Cardiovascular Diseases*, vol. 32, no. 4, pp. 273–290, 1990.

[17] F. Liu, K. Wang, D. Liu, X. Yang, and J. Tian, "Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101873.

[18] Y. Ali, F. Janabi-Sharifi, and S. Beheshti, "Echocardiographic image segmentation using deep Res-U network," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102248.

[19] A. Amer, X. Ye, and F. Janan, "ResDUnet: A deep learning-based left ventricle segmentation method for echocardiography," *IEEE Access*, vol. 9, pp. 159755–159763, 2021.

[20] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023, *arXiv:2304.02643*.

[21] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: An experimental study," *Med. Image Anal.*, vol. 89, Oct. 2023, Art. no. 102918.

[22] J. F. Pombo, B. L. Troy, and R. O. Russell, "Left ventricular volumes and ejection fraction by echocardiography," *Circulation*, vol. 43, no. 4, pp. 480–490, Apr. 1971.

[23] J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, and Y. Jin, "Medical SAM adapter: Adapting segment anything model for medical image segmentation," 2023, *arXiv:2304.12620*.

[24] X. Hu, X. Xu, and Y. Shi, "How to efficiently adapt large segmentation model(SAM) to medical images," 2023, *arXiv:2306.13731*.

[25] K. Zhang and D. Liu, "Customized segment anything model for medical image segmentation," 2023, *arXiv:2304.13785*.

[26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[27] A. Augustin, J. Yi, T. Clausen, and W. Townsley, "A study of LoRa: Long range & low power networks for the Internet of Things," *Sensors*, vol. 16, no. 9, p. 1466, Sep. 2016.

[28] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[29] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[30] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 8792–8802.

[31] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, J. D'hooge, L. Lovstakken, and O. Bernard, "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2198–2210, Sep. 2019.

[32] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945.

[33] J. Henrikson, "Completeness and total boundedness of the Hausdorff metric," *MIT Undergraduate J. Math.*, vol. 1, nos. 69–80, p. 10, 1999.

[34] M. Hu, Y. Li, and X. Yang, "SkinSAM: Empowering skin cancer segmentation with segment anything model," 2023, *arXiv:2304.13973*.

YAQI ZHU is currently pursuing the master's degree with Ningbo University, China. Her research interest includes segmentation of cardiac ultrasound images and videos.

CHANGCHUN XIONG is currently pursuing the master's degree with Ningbo University, China. His research interest includes biomedical signal processing.

HENG ZHAO is currently pursuing the master's degree with Ningbo University, China. His research interests include deep learning and medical image denoising.

YUDONG YAO (Fellow, IEEE) is currently a Professor with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, USA. His research interests include deep learning and medical imaging processing. He is a fellow of American Institute for Medical and Biological Engineering (AIMBE).