Meghana Manvitha Venna* and Rohini Palanisamy

Enhancing Ovarian Tumor Segmentation with Attention-Driven Adaptation of MedSAM

https://doi.org/10.1515/cdbme-2025-0140

Abstract:
Undetected fluid-filled ovarian cysts can lead to severe complications like rupture and hemorrhage, requiring emergency medical interventions. Ultrasound is a commonly used imaging modality for detecting cysts and tumors. However, it presents challenges with weak contrast, blurred boundaries, and noise affecting the diagnostic reliability. Segment Anything Model (SAM) is a powerful framework for image segmentation, but struggles in medical imaging tasks due to its lack of domain-specific knowledge. MedSAM extends SAM to clinical data and demonstrates superior capabilities across various adnexal regions and imaging modalities. It lacks specificity for ultrasound images and underperforms in ovarian tumor segmentation. This limitation is addressed by efficient adaptation of MedSAM using attention-based adapter modules. Instead of fine-tuning the entire MedSAM, incorporating an adapter module is proven to be an efficient training strategy. The proposed approach optimizes model training using an attention based adapter, guided by a binary cross-entropy loss function with logits to ensure numerical stability. The adapted model achieves a dice score of 70%, demonstrating robust segmentation capabilities. With its ability to produce sharp boundaries and reliably segment regions of interest, this work highlights the potential of an attention-driven discriminative learning framework for ovarian tumor detection, contributing towards automated and effective computer-aided diagnosis of ovarian cancer.

Keywords: Ovarian Cancer, Cyst, Ultrasound, MedSAM, Adaptation, Attention

1 Introduction

Ovarian cancer is the eighth most common cancer among women [1], accounting for over one-third of newly diagnosed cancer cases worldwide [2]. It is also one of the deadliest gynecological malignancies, frequently diagnosed at an advanced stage due to a lack of early-stage symptoms and effective screening methods.

Ovarian cancer originates in the ovaries, forming tumors or cystic masses within the pelvic region. Early identification, detection and segmentation of ovarian masses such as cysts and tumors are crucial for accurate diagnosis and better treatment. Among the medical imaging modalities, ultrasound (US) is one of the most commonly used reliable methods for evaluating adnexal masses, due to its non-ionizing nature, accessibility, and cost-effectiveness [3] [4] . Ultrasound imaging is particularly predominant in gynecological diagnostics, which makes it an essential tool in ovarian cancer detection.

Ultrasound imaging does present with unique challenges compared to other modalities. Low image resolution, noise and artifacts can lead to blurry ovarian mass edges, making it difficult for both medical professionals and computer vision-based imaging software to accurately identify tumors and cysts. These challenges can result in error-prone analysis or significant delays in detecting regions of interest within the ultrasound scans.

To address these limitations, there is a growing need for highly robust data-driven models that can improve accuracy and reduce latency in ultrasound-based ovarian cancer screening. Deep learning architectures have shown remarkable success in the medical imaging domain [5]. Convolutional neural networks (CNN) are used to extract spatial features while transformer-based models are used to understand long-range dependencies within images [6] [7]. They have been shown to improve the accuracy of tumor detection and segmentation in ultrasound imaging.

Segment Anything Model (SAM) [8] is a widely recognized benchmark in natural image segmentation, illustrating versatility and efficiency. However, its direct application to medical imaging tasks has been shown to be suboptimal [9] [10]. This performance gap arises due to the significant domain shift between natural and medical images as the latter requires specialized contextual understanding. To overcome this limitation and extend the robustness of SAM to medical image segmentation, different studies have explored adaptation strategies that leverage large-scale pretraining on natural images. MedSAM [11] has emerged as a promising framework, achieving significant improvement across various medical imaging modalities. Despite its effectiveness, MedSAM struggles with ovarian ultrasound imaging, mainly due to its

***Corresponding author: Meghana Manvitha Venna,** IIITDM, Kancheepuram, Chennai, Tamil Nadu, 600127, India, e-mail: meghana07d@gmail.com
Rohini Palanisamy, IIITDM, Kancheepuram, Chennai, Tamil Nadu, 600127, India, e-mail: rohinip@iiitdm.ac.in

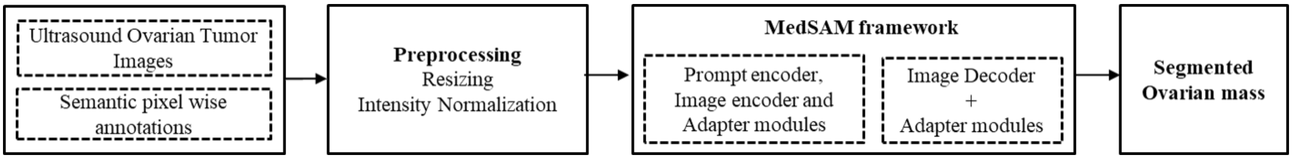
Ultrasound Ovarian Tumor Images
Semantic pixel wise annotations
Preprocessing
Resizing
Intensity Normalization
MedSAM framework
Prompt encoder, Image encoder and Adapter modules
Image Decoder + Adapter modules
Segmented Ovarian mass

Fig. 1: Proposed work flow schematic

lack of modality-specific adaptation. In this context, this study proposes an efficient attention-based adaptation of MedSAM for ovarian cancer detection. The aim is to enhance the segmentation performance for ovarian masses, including cysts and tumors, through improving boundary sharpness and segmentation accuracy.

2 Methodology

The workflow of the proposed methodology is shown in the Figure 1

2.1 Dataset

In this study, Ovarian tumor images are obtained from publicly available Multi-Modality Ovarian Tumor Ultrasound (MMOTU) dataset [12] dataset. 1469 standard 2D ultrasound images are considered with pixel-wise semantic annotations and global categorical labels for segmentation and classification tasks. These images are preprocessed for intensity normalization and resized to the network resolution ensuring efficient training and inference.

2.2 Model

SAM model architecture contains three primary components namely, image encoder, prompt encoder and mask decoder. Pre-trained SAM uses ViT as image encoder with global attention blocks to extract hierarchical image features. The prompt encoder processes spatial cues, including points and bounding boxes, and generates positional embeddings. Mask decoder is a transformer-style decoder block that has two-way cross-attention to facilitate the interaction between Image embeddings and prompt embeddings.

To enhance the adaptability of SAM for ovarian ultrasound segmentation, adapter modules are added into the architecture by MedSAM framework using a bottleneck design strategy. Each adapter consists of a down-projection layer to reduce dimensionality, followed by a non-linear activation function (ReLU) and an up-projection layer to restore dimensionality. These adapters are integrated into the SAM architecture as follows: two adapters per ViT block in the image encoder and three adapters per ViT block in the mask decoder branch. This design ensures efficient fine-tuning without overfitting in addition to preserving the feature extraction capabilities of the SAM model.

The dataset is split into a 70-30 ratio for training and testing. The model is trained for 30 epochs using the Adam optimizer function with a learning rate of 1e-4. Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss) function is used as the loss criterion, while validation is performed using the DICE score to evaluate segmentation performance.

2.2.1 Loss function

The loss function used for model training is BCEWithLogitsLoss. Instead of applying a sigmoid activation followed by binary cross-entropy loss as given in eqn (1) and (2) separately, BCEWithLogitsLoss integrates both steps into a single function as shown in eqn (3), to ensure numerical stability. Instability is caused in binary cross-entropy loss due to extreme probability values. When probabilities reach 0 or 1, logarithm-based operations can cause precision issues in floating-point representations, potentially causing either overflow or underflow. Similarly, using a standalone sigmoid function introduces an exponent in the denominator, which can also lead to numerical instability. To avoid this, BCEWithLogitsLoss stabilize computations by preventing the extreme values from dominating the calculations. This improves numerical precision, and leads to more reliable tumor segmentation in ultrasound images.

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{1}$$

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right] \tag{2}$$

$$\mathcal{L}_{BCELogits} = \frac{1}{N}\sum_{i=1}^{N}\Big[\log(1+e^{-z}) + z(1-y_i)\Big] \tag{3}$$

Here, N represents the total number of samples, y_i is the true label (either 0 or 1) for the i-th sample and $\hat{y}_i$ represents the predicted probability of the i-th sample.

3 Results

Figure 2 presents representative ovarian ultrasound images and their corresponding segmented ovarian mass, where the presence of a cyst or tumor is indicated in the center. This visualization highlights the reflective properties of the sonar waves, which are important in distinguishing the tissue types present in the image. It is also essential to note the presence of speckle noise, which is a common artifact in ultrasound images that can complicate accurate segmentation. Ultrasound device markings may also introduce inconsistencies or noise, and the model must be capable of distinguishing between relevant features and the noise present in the image.

The training and validation performance metrics of the model is illustrated in Figure 3. The training loss exhibits a steady decline over the epochs, indicating the ability of the model to learn ovarian tumor-specific features. Likewise, the decreasing validation loss demonstrates the capability of the model to generalize to unseen data. The image encoder designed in this study captures patch-wise long-term dependencies across the image, thereby identifying global ovarian mass boundaries while preserving fine-grained features. This ability to recognize the global tumor boundaries is critical for generating an accurate mask. By pooling information from all tumor pixels, the encoder facilitates a more comprehensive understanding of the shape and position of the tumor. The decoder leverages this knowledge for precise segmentation of the tumor pixels from the image. The adapted MedSAM model for ovarian ultrasound effectively captures discriminative features of cysts and tumors, enabling precise segmentation of ovarian masses.

Dice coefficient is a widely used similarity metric in image segmentation tasks to quantify the overlap between the predicted segmentation mask and ground truth. It ranges from 0 to 1, where 0 indicates no overlap and 1 implies perfect overlap with the annotation. It offers a robust measure of pixel wise agreement in binary mask evaluations. Intersection of Union (IoU), also called Jaccard index, measures the amount of overlap between the predicted segmentation mask and the ground truth mask. While DICE measures the similarity between the prediction and ground truth masks, IoU provides the precise overlap relative to the combined area.

Variation of IoU and Dice coefficient score across epochs is plotted in Figure 4. The graphs show consistent improvement across epochs on the validation dataset, suggesting that the predicted segmentation masks progressively achieve better alignment with ground truth annotations. By the end of training, the model attains a maximum Dice score of about 0.7, with a minimum loss of 0.25 at a learning rate of 1e-4.

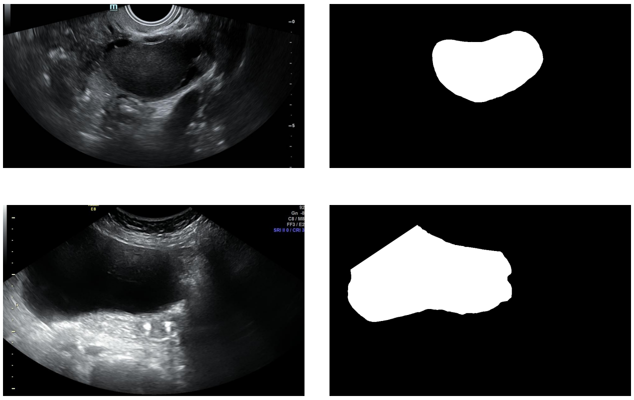

Fig. 2: Ovarian ultrasound tumor images and corresponding ground truth masks

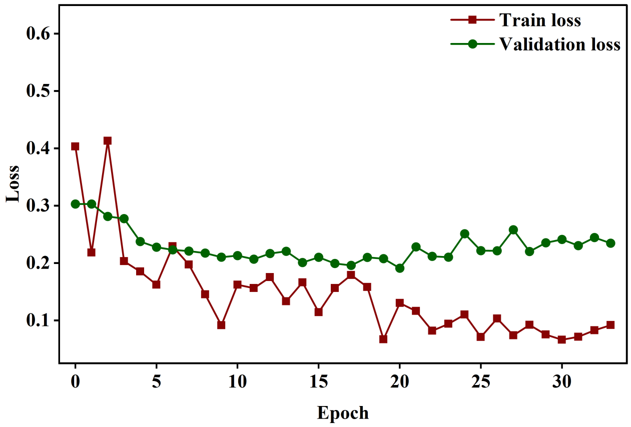

Fig. 3: Variation of Train and Validation loss across epoch

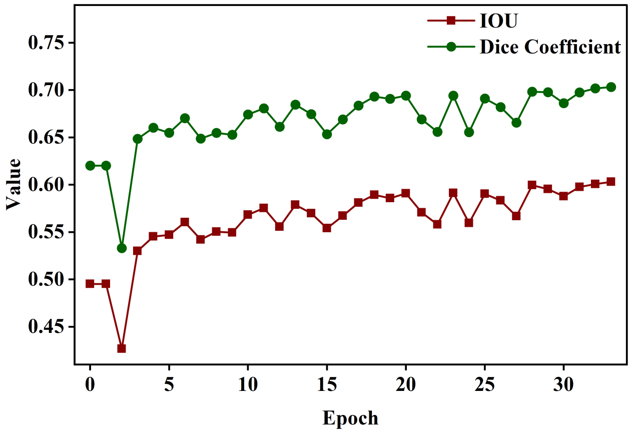

Fig. 4: Variation of IoU and Dice coefficient during training

The qualitative performance of the model is demonstrated in Figure 5, where representative test images, confidence maps, final thresholded segmentation masks, and ground truth images are presented. The confidence maps highlight pixel-wise probability estimates for tumor regions while preserving structural details such as image texture, wall boundaries, and

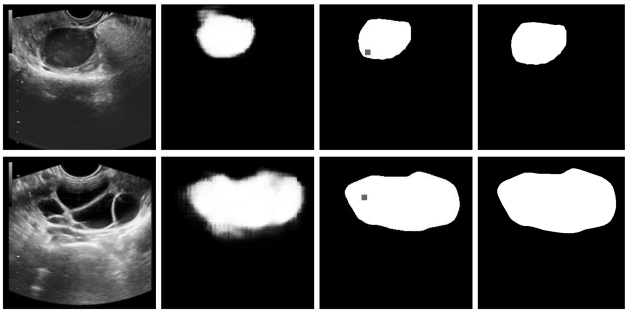

Fig. 5: Representative segmented images using trained MedSAM model (a) Input image, (b) Confidence map, (c) Prediction mask, (d) Ground truth

hierarchical features. Thresholding is subsequently applied to generate the final binary segmentation masks. Visual interpretation of the obtained segmentation mask demonstrates strong similarity with the corresponding ground truth masks, making this approach reliable for use in the automated segmentation of ovarian tumors.

4 Conclusion

This study investigates the adaptation of MedSAM for ovarian tumor and cyst segmentation. The introduction of the adapter modules in SAM framework enabled efficient attention-based adaptation to ultrasound images and ovarian tumor data. The gradual decline of validation loss indicates that the adapter module facilitates optimal feature learning in fine tuning of MedSAM to medical images. Furthermore, Dice coefficient of about 70% indicates reliable segmentation accuracy and strong overlap between the predicted and ground truth masks. These results demonstrate that this approach effectively segments ovarian masses with improved precision and robustness and has the potential for integration into automated ovarian cancer diagnosis, offering a reliable tool for accurate and low-latency tumor detection in clinical settings.

References

[1] World Cancer Research Fund. *Ovarian cancer statistics.* Available at: https://www.wcrf.org/preventing-cancer/cancer-statistics/ovarian-cancer-statistics/.

[2] Park IS, Kim SI, Han Y, Yoo J, Seol A, Jo H, et al. Risk of female-specific cancers according to obesity and menopausal status in 2.7 million Korean women: Similar trends between Korean and Western women. Lancet Reg Health West Pac 2021;11:100146. doi:10.1016/j.lanwpc.2021.100146.

[3] Avola, Danilo, Luigi Cinque, Alessio Fagioli, Gianluca Foresti, and Alessio Mecca. "Ultrasound Medical Imaging Techniques: A Survey." *ACM Computing Surveys*, vol. 54, no. 3, Apr. 2022, article 67, pp. 1-38. Association for Computing Machinery, New York, NY, USA. Available at: https://doi.org/10.1145/3447243.

[4] Shantharam R, Palanisamy R. Analysis of Hyperparameter tuned UNet++ Deep model for Delineation of Ultrasound Ovarian Tumors. In: 2023 IEEE 7th Conference on Information and Communication Technology (CICT); 2023 Dec; Jabalpur, India. p. 1–5. Available from: https://doi.org/10.1109/CICT59886.2023.10455397.

[5] Zhou SK, Greenspan H, Davatzikos C, Duncan JS, van Ginneken B, Madabhushi A, Prince JL, Rueckert D, Summers RM. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE Inst Electr Electron Eng. 2021 May;109(5):820-838. doi: 10.1109/JPROC.2021.3054390.

[6] Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929, 2020.

[7] Shamshad, Fahad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. "Transformers in medical imaging: A survey." *Medical Image Analysis*, vol. 88, 2023, 102802. ISSN 1361-8415. Available at: https://doi.org/10.1016/j.media.2023.102802.

[8] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollar P, Girshick R. Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct. p. 4015-4026.

[9] He S, Bao R, Li J, Grant PE, Ou Y. Accuracy of Segment-Anything Model (SAM) in medical image segmentation tasks. CoRR. 2023;abs/2304.09324. Available from: https://doi.org/10.48550/arXiv.2304.09324.

[10] Huang Y, Yang X, Liu L, Zhou H, Chang A, Zhou X, et al. Segment anything model for medical images? Med Image Anal. 2024;92:103061. Available from: https://doi.org/10.1016/j.media.2023.103061.

[11] Wu, Junde, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. "Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation." arXiv preprint arXiv:2304.12620, 2023.

[12] Zhao, Qi, Shuchang Lyu, Wenpei Bai, Linghan Cai, Binghao Liu, Guangliang Cheng, Meijing Wu, Xiubo Sang, Min Yang, and Lijiang Chen. "MMOTU: A Multi-Modality Ovarian Tumor Ultrasound Image Dataset for Unsupervised Cross-Domain Semantic Segmentation." arXiv preprint arXiv:2207.06799, 2023. Available at: https://arxiv.org/abs/2207.06799.