Adapting SAM2 Model from Natural Images for COVID-19 Lung CT Scan Lesion Segmentation X-Ray Images

Mahraj Afsan Olive, MD Zafrul Hasan, MD Abdul Aziz, Mahedi Hasan

Computer Science & Engineering, American International University-Bangladesh,

Dhaka, Bangladesh

Abstract:

Lung CT imaging, due to its high diagnostic accuracy and ability to detect subtle abnormalities, has become a widely used tool for COVID-19 detection and monitoring. Accurate lung lesion segmentation is crucial for effective disease assessment and treatment planning, enabling clinicians to quickly and precisely evaluate the severity of COVID-19 in patients. This dataset aims to enhance COVID-19 detection and segmentation in medical imaging. We propose a segmentation method that utilizes a U-Net architecture, optimized with LoRA adapters for efficient fine-tuning. We incorporate attention gate decoders to further improve segmentation performance, which enhance feature selection and refine segmentation accuracy. The combination of U-Net, LoRA, and attention gates enables the model to learn effectively from the dataset while minimizing computational resources. Our approach significantly improves segmentation accuracy for COVID-19 lesions, demonstrating the potential of this dataset and method for advancing automated COVID-19 detection in clinical applications.

Literature Review:

In Xiong, X., Wu, Z., Tan, S., Li, W., Tang, F., Chen, Y., Li, S., Ma, J. and Li, G., 2024[1] proposed SAM2-UNet, a segmentation model that uses SAM2's Hiera backbone as an encoder and a U-Net-style decoder for various segmentation tasks. The model uses the Hiera backbone from SAM2 to extract multi-scale features, enhancing segmentation performance. To efficiently fine-tune SAM2 without modifying its large number of parameters, adapters are inserted into the encoder, enabling parameter-efficient learning. Additionally, Receptive Field Blocks (RFBs) are incorporated to refine the extracted features and reduce the computational complexity. The decoder follows the classic U-Net structure, replacing SAM2's original two-way transformer-based mask decoder. The model is trained using weighted IoU loss and binary cross-entropy (BCE) loss, ensuring improved segmentation accuracy. SAM2-UNet is evaluated on 18 datasets spanning five tasks, including camouflaged object detection, salient object detection, marine animal segmentation, mirror detection, and polyp segmentation. The results show that SAM2-UNet outperforms specialized state-of-the-art segmentation models, proving that SAM2 is a powerful encoder for both natural and medical image segmentation.

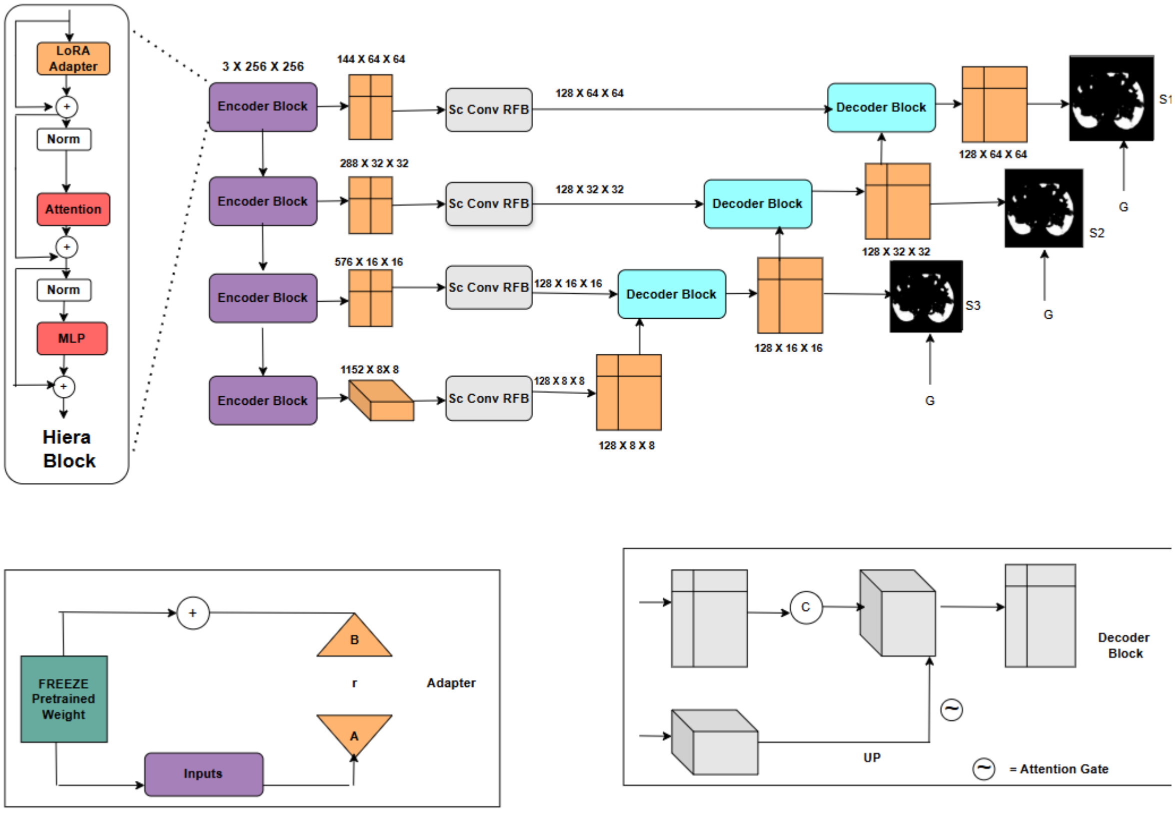
In Li, Z., Tang, W., Gao, S., Wang, Y. and Wang, S., 2024[2] focused on adapting SAM2 for tooth segmentation in dental panoramic X-ray images, addressing challenges such as low contrast, noise, overlapping anatomical structures, and limited datasets. Since medical images significantly differ from natural images, the direct application of SAM2 does not yield optimal results. To overcome this, adapter modules are used to fine-tune the pre-trained SAM2 model, and ScConv modules are introduced to reduce feature redundancy and enhance multi-scale feature extraction. Additionally, a gated attention mechanism is implemented in the skip connections to improve focus on relevant structures. Given the high computational cost of deep models, the paper also introduces LightUNet, a lighter version of SAM2 trained using knowledge distillation, where the fine-tuned SAM2 model serves as the teacher network. This technique retains high segmentation accuracy while reducing computational costs, making the model more efficient for real-time applications. Experimental results on the UFBA-UESC dataset show that LightUNet achieves superior performance compared to the traditional UNet model while using only 1.6% of its parameters and requiring 24% of the inference time, making it ideal for deployment on edge devices in dental diagnostics.

In Zhu, J., Qi, Y. and Wu, J., 2024 [3] introduced a novel approach by treating medical image segmentation as a video object tracking problem. Unlike traditional deep learning models that struggle with generalization and handling both 2D and 3D data, MedSAM-2 employs a self-sorting memory bank mechanism to dynamically select informative embeddings based on confidence scores and dissimilarity, ensuring that only relevant features contribute to segmentation. This mechanism significantly enhances 3D medical image segmentation and enables One-Prompt Segmentation, where a single prompt can segment multiple 2D images that are not temporally related. The MedSAM-2 pipeline builds upon SAM2's memory-enhanced segmentation framework while improving generalization across diverse medical imaging datasets. It is tested on 14 different benchmarks, covering 25 tasks such as white blood cell segmentation, optic cup detection, kidney tumors, coronary arteries, and cerebral arteries, and outperforms existing state-of-the-art models in both 2D and 3D segmentation tasks. By leveraging context-aware video pretraining and self-sorting memory banks, MedSAM-2 provides a unified framework for handling both ordered and unordered medical images, making it a robust and scalable solution for medical diagnostics.

All three papers build on SAM2 for segmentation tasks, using fine-tuning techniques like adapters to adapt SAM2 to different domains. They integrate multi-scale feature extraction, with SAM2-UNet using Receptive Field Blocks (RFBs) and the Tooth Segmentation model using ScConv modules. Additionally, all three incorporate U-Net-style decoders for improved segmentation accuracy. However, SAM2-UNet focuses on general segmentation, while MedSAM-2 and the Tooth Segmentation model target medical imaging. MedSAM-2 introduces a self-sorting memory bank, treating 3D medical images as videos, enabling One-Prompt Segmentation, which neither of the others have. The Tooth Segmentation model prioritizes efficiency, using knowledge distillation to create LightUNet, making it suitable for edge devices, unlike SAM2-UNet and MedSAM-2, which focus more on accuracy. Overall, SAM2-UNet is a general solution, MedSAM-2 excels in handling medical images as videos, and the Tooth Segmentation model optimizes for efficiency and deployment. Their differences highlight unique optimizations tailored to their applications rather than direct competition.

Method:

Proposed Model:



S2AgScUNet utilizes the Hiera backbone pretrained by SAM2. Compared to the standard ViT encoder used in SAM, Hiera employs a hierarchical structure that enables better multi-scale feature capture, making it more suitable for designing U-shaped networks. Specifically, given an input image I ∈ R3×H×W, where H represents the height and W represents the width

Adapters with LoRA for Efficient Fine-Tuning:

Given the large number of parameters in Hiera (e.g., 214M for Hiera-L), full fine-tuning can be impractical due to memory constraints. To address this, we freeze the base parameters of Hiera and introduce LoRA-based adapters before each multi-scale block. This strategy allows for parameter-efficient fine-tuning. Each adapter consists of a low-rank matrix for efficient learning and a linear layer for down sampling.

Decoder with Attention Gate:

The original mask decoder in SAM2 uses a two-way transformer approach to facilitate feature interaction between the prompt embedding and encoder features. In contrast, inspired by the highly customizable U-shaped structure that has proven effective across various tasks [58,3,2], our decoder adopts the classic U-Net design. It consists of three decoder blocks, each containing two 'Conv-BN-ReLU' combinations, where 'Conv' denotes a 3 × 3 convolution layer and 'BN' represents batch normalization. Additionally, an attention gate is integrated within the decoder blocks to further refine feature selection by allowing relevant features from the encoder to pass through while suppressing irrelevant ones. The output feature from each decoder block is then passed through a 1×1 convolution segmentation head to produce a segmentation result SiS_iSi (where i∈1,2,3i \in 1,2,3i∈1,2,3), which is subsequently upsampled and supervised by the ground truth. truth mask G.

Loss Functions: Binary Cross-Entropy and IoU
Binary Cross-Entropy (BCE) measures the difference between predicted probabilities and actual labels. IoU calculates the overlap between the predicted mask and ground truth. Together, they optimize both pixel-wise accuracy (BCE) and shape consistency (IoU) in segmentation tasks.

Results:

Epoch 50

Dagtasets COVID-19 CT scan lesion segmentation dataset

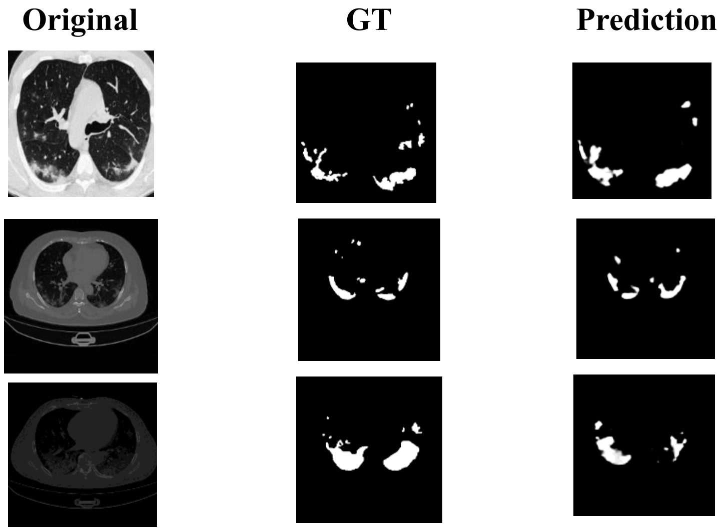

Original **GT** **Prediction**

Fig-1: Comparison of original, ground truth and predicted images.

Segmentation Results

Metric	Value
mIoU	0.489
mDice	0.615

Table-1: Evaluation metrics value.

Conclusion:

The proposed segmentation model enhances COVID-19 lung lesion segmentation by integrating the pretrained SAM2 encoder, LoRA adapters for fine-tuning, and attention gate decoders. Experimental results demonstrate that our approach adapts the standard U-Net model in key metrics, including IoU, achieving enhanced segmentation in medical image datasets. The introduction of SAM2's hierarchical feature extraction significantly boosts the model's ability to capture fine-grained lesion details, especially in small datasets, reducing overfitting and improving generalization. Furthermore, our method optimizes computational efficiency by leveraging parameter-efficient fine-tuning techniques, enabling potential deployment in real-time clinical applications.

Reference:

[1] Xiong, X., Wu, Z., Tan, S., Li, W., Tang, F., Chen, Y., Li, S., Ma, J. and Li, G., 2024. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. arXiv preprint arXiv:2408.08870.

[2] Li, Z., Tang, W., Gao, S., Wang, Y. and Wang, S., 2024. Adapting SAM2 Model from Natural Images for Tooth Segmentation in Dental Panoramic X-Ray Images. Entropy, 26(12), p.1059.

[3] Zhu, J., Qi, Y. and Wu, J., 2024. Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874.

https://www.kaggle.com/datasets/maedemaftouni/covid19-ct-scan-lesion-segmentation-dataset/discussion?sort=undefined