

AI-Based Analysis of Abdominal Ultrasound Images to Support Medical Diagnosis in Emergency Departments

Zahra HAMED^a, Lorenzo BRIGATO^a, Ethan DACK^{a,d}, Mira SCHÜTZ^{a,b}, Beat LEHMANN^b, Aristomenis EXADAKTYLOS^b, Stavroula MOUGIAKAKOU^{a,b,1}, and Gert KRUMMREY^{b,c}

^a*ARTORG Center for Biomedical Engineering Research, University of Bern*

^b*Department of Emergency Medicine, Bern University Hospital*

^c*Institute for Medical Informatics (IMI), Bern University of Applied Sciences (BFH)*

^d*Graduate School for Cellular and Biomedical Sciences, University of Bern*

ORCID ID: Zahra Hamed <https://orcid.org/0000-0001-5279-500X>, Lorenzo Brigato <https://orcid.org/0000-0002-0872-069X>, Mira Schütz <https://orcid.org/0009-0008-8806-344X>, Beat Lehmann <https://orcid.org/0000-0003-3218-2470>, Aristomenis Exadaktylos <https://orcid.org/0000-0002-2705-5170>, Stavroula Mougiakakou <https://orcid.org/0000-0002-6355-9982>, Gert Krummrey <https://orcid.org/0000-0002-8397-2336>

Abstract. The goal of segmentation in abdominal imaging for emergency medicine is to accurately identify and delineate organs, as well as to detect and localize pathological areas. This precision is critical for rapid, informed decision-making in acute care scenarios. Vision foundation models, such as Segment Anything Model (SAM), have demonstrated remarkable results on many different segmentation tasks, but they perform poorly on medical images because of the scarcity of medical datasets. They lack robust generalizability across diverse medical imaging modalities, and they need to be fine-tuned specifically for medical images, as these images considerably differ from natural images. This study aims to investigate the application of a foundation segmentation model to ultrasound (US) images of the abdomen. We employed SAMed to segment and classify all organs and free fluid present in each US image. A dataset comprising 286 US images, corresponding segmentation masks, and organ-level labels was collected from the Bern University Hospital Inselspital. Due to the relatively small size of our dataset, we pre-trained SAMed on a larger public US dataset to fine-tune it for US imaging. We then applied this fine-tuned SAMed on the Inselspital dataset to generate multi-class masks and assessed its performance against ground truth annotations using standard evaluation metrics. The results demonstrated that the fine-tuned SAMed can identify and classify multiple organs, though challenging cases, such as free fluid segmentation, reveal opportunities for improvement. Furthermore, transfer learning proved to be a reliable solution for managing small datasets, a key obstacle in the medical imaging realm.

Keywords. SAM, Segmentation, Classification, Ultrasound images

1. Introduction

Ultrasound imaging is a portable, cost-efficient, and non-invasive diagnostic tool used in medical settings, including emergency departments [1]. While effective, accurate

¹ Corresponding Author: Stavroula Mougiakakou, stavroula.mougiakakou@unibe.ch

interpretation relies heavily on physician expertise. Artificial Intelligence (AI)-driven real-time segmentation and classification of US images can support diagnostic accuracy and efficiency [2].

In general, medical image semantic segmentation attempts to find anatomical structures or pathological conditions within medical images and assigns a certain class to each pixel in the image [3]. Accurate segmentation and classification of the desired region in medical images can support precise diagnosis, disease assessment, and tailored treatment plans [2]. Deep learning models, such as nnU-Net [4], have achieved outstanding results on medical imaging segmentation tasks. However, these models require extensive amounts of annotated data for training. Generating such large datasets needs skilled clinicians to manually annotate each image at the pixel level, which is time-consuming and labor-intensive [3]. Recently, the advent of large-scale foundation models has mitigated the dependency on large, annotated datasets for training. Since these models were trained on extensive and diverse datasets, they can provide zero-shot or few-shot generalization capabilities [5].

While foundation models like SAM [6] have shown impressive achievements in computer vision, there are still certain limitations to their use in medical images [2]. This problem arises because these models were trained on non-medical datasets. Therefore, it limits their ability to recognize complex human anatomy structures in medical images such as US [7]. Furthermore, they fail to associate the segmented regions with relevant and meaningful semantic categories [8].

SAMUS [9] is built upon SAM and is tailored for US image segmentation. They introduced an auto prompt generator to SAM, which enables the model to automatically generate prompt embeddings instead of relying on a manual prompt encoder. Additionally, a convolutional neural network (CNN) branch and cross-branch attention are included to overcome the problem of insufficient local features in vision transformers. As SAMUS did not change the segmentation head of SAM, it can only detect one target region in US images and cannot recognize all organs or abnormalities within a single US image.

The present research constitutes a proof-of-concept application in the US imaging domain using a fine-tuned version of SAMed [8], a foundation model for medical image segmentation. SAMed is developed to produce multiple masks for distinct anatomical structures or abnormal regions in different medical image modalities. In our research, we introduce a fine-tuned version of the SAMed to identify and classify abdominal organs in US images, with a focus on detecting free fluid, a key indicator of malignancy or infection. The model is accompanied by a dataset of 286 US images with corresponding segmentation masks from Emergency Department of the Bern University Hospital, supplemented with 11,223 US images from publicly open-access datasets.

2. Methods

In this work, we employed SAMed to segment and classify different organs and free fluids in US images. SAMed is developed based on SAM and exploits its exceptional image segmentation capabilities while being tailored specifically for medical image segmentation. To adapt SAM for semantic segmentation tasks in the medical domain, several modifications were made to the original architecture to enhance its performance on such specialized data, as we describe in the following.

SAM is designed as an ambiguity-aware model capable of predicting multiple masks for a single prompt when the input prompt is ambiguous. However, SAMed modifies the segmentation head of SAM to tailor its output for semantic segmentation tasks. Rather than generating multiple masks in cases of ambiguity, SAMed produces multiple masks corresponding to different classes. It combines k distinct semantic masks and applies softmax and argmax operations along the channel dimension. The final output is a mask where each pixel is assigned to a distinct class.

Notably, SAMed does not require prompts during inference to perform automatic segmentation, which provides advantages for automatic medical diagnosis. Medical images often contain multiple regions of interest, thereby a prompt-free model can automatically detect all regions without defining specific prompts for each target region. In SAMed, all prompts are removed, and the model updates learnable embeddings during training to offer more efficient and flexible segmentation. In our dataset, the number of organs varies across different images. SAMed allows the model to automatically recognize all organs present in each image without the need for determining separate prompts for each individual organ.

SAMed makes use of a technique known as Low-Rank Adaptation (LoRA) to efficiently reduce computational resource requirements [10]. LoRA addresses the challenges associated with fine-tuning large models such as SAM. Retraining all parameters in such models would be computationally expensive. However, LoRA mitigates this issue by freezing the pre-trained model weights and introducing trainable low-rank decomposition matrices. In SAMed, LoRA is implemented in the projection layers of query, key, or value to update the attention scores.

3. Experiments

3.1. Dataset

We collected 286 US images of abdominal organs from Bern University Hospital Inselspital (InselUS), with the hospital's consent and the necessary ethical approvals (KEK Bern 2024-00953). All images were de-identified prior to analysis. Each image contains one or more abdominal organs, resulting in a dataset comprising 105 cases of the kidney, 131 of the liver, 32 of the gallbladder, 74 of the spleen, 40 of the bladder, 47 of the pancreas, and 50 cases of free fluid accumulation in the abdominal cavity (ascites, hemorrhage). The annotation and labeling of the InselUS were performed using a semi-automatic SAM-based tool specifically developed in-house for this research project (available at https://github.com/hastih/SAM_based-Tool/tree/main). Additionally, we collected 11223 US images from open-access datasets, including 1601 of the nerve neck, 7774 of the thyroid, 630 of the breast, 732 of the liver, and 486 of the kidney. Both datasets show significant class imbalance.

3.2. Implementation Details

To increase the dataset size, we applied several data augmentation techniques, including image rotation by random multiples of 90° , followed by horizontal/vertical flips along the horizontal or vertical axis. Additionally, random rotation with angle variation between -20° and 20° was applied. All images were resized to 512×512 pixels to maintain consistent model input sizes.

Since SAMed was originally trained on computer tomography (CT) scans of the abdominal multi-organ segmentation dataset (Synapse), we adapted its weights via fine-tuning on US images. Because the InselUS was limited in size for fine-tuning SAMed, we employed a transfer learning approach. Hence, initially, we trained SAMed on the collected public US data. Following fine-tuning, the model was then applied to the InselUS dataset. Transfer learning ensured that the model was better suited to segmenting the abdominal organs and free fluids in US images, as we will show in the next section.

We optimized the model with AdamW and a learning rate of 0.0004. To mitigate severe class imbalance, we employed the weighted Dice and cross-entropy losses. The model is evaluated on 20% held-out samples using the popular metrics of Intersection over Union (IOU), Dice score, Hausdorff distance 95% (HD95), Accuracy, Precision, and Sensitivity.

4. Results and Discussion

In Table 1, we summarize the semantic segmentation results obtained from the fine-tuned SAMed model on the InselUS dataset, presenting the performance for each organ as well as for free fluid.

Table 1. Segmentation and classification results per class and averaged over all classes.

Class	Dice	HD95	Sensitivity	Precision	IOU	Accuracy
Bladder	0.6271	19.3365	0.6105	0.7267	0.5511	0.7691
Gallbladder	0.4759	18.4852	0.4171	0.6795	0.3859	0.7256
Kidney	0.4907	31.9442	0.5308	0.5098	0.4105	0.6642
Liver	0.4924	41.5934	0.4702	0.5563	0.4178	0.6291
Pancreas	0.1115	37.8118	0.1402	0.2119	0.0696	0.3501
Spleen	0.3744	41.8523	0.3675	0.4836	0.3086	0.5581
Free Fluid	0.4066	67.6334	0.4345	0.5057	0.2980	0.7695
All	0.4255	36.9510	0.4244	0.5248	0.3488	0.6380

The highest value for HD95 is observed in the segmentation of free fluid. A high Hausdorff distance indicates discrepancies between the predicted and ground truth segmentation boundaries. This is primarily due to the complex and variable shape of free fluid, which changes depending on the location of its accumulation. In contrast, the smallest HD95 values are recorded for the bladder and gallbladder at 19.3365 and 18.4852, respectively. The boundaries of these organs are much clearer, with more distinct pixel contrast compared to other organs.



Figure 1: Visualization of SAMed’s predicted segmentation masks compared to the ground truth. The central column shows the ground truth masks for the liver and kidney, while the right column displays the corresponding predicted masks generated by SAMed.

In Table 2, we summarize the results for all classes, before and after fine-tuning SAMed on the US images. As observed, the metrics improved after fine-tuning, which demonstrates the effectiveness of transfer learning in tackling small dataset problems. The results revealed clear but modest improvements after fine-tuning. A possible reason is that we applied LoRA with a rank of 4 in SAMed. Consequently, we updated only a small subset of the model’s parameters rather than fine-tuning the entire set of parameters. Despite fine-tuning some of the model’s parameters, transfer learning still improved the model’s performance and demonstrated its potential to overcome the problem of small datasets.

Table 2. Segmentation and classification results per class and averaged over all classes.

Metric	SAMed After Fine-Tuning	SAMed Before Fine-Tuning
Dice	42.55%	40.02%
HD95	36.95%	37.82%
Sensitivity	42.42%	40.40%
Precision	52.48%	50.15%
IOU	34.88%	32.59%
Accuracy	63.80%	62.06%

The results reveal that SAMed struggles with different medical imaging modalities, and must be fine-tuned to perform well on a specific image modality. However, inadequate medical datasets, particularly US images, make it difficult to properly fine-tune these models. As the InselUS dataset was small, we boosted the model’s accuracy to 63.80% through transfer learning. We plan to enlarge the dataset and include a greater variety of training samples in order to further enhance the model’s performance.

5. Conclusion and Discussion

In this work, we demonstrated the practical application of foundation models on US images, focusing on the segmentation and classification of abdominal organs and free fluids to support medical diagnosis in the emergency department. SAM’s strong performance in general imaging tasks does not translate to medical imaging, where it suffers from substantial degradation and poor generalization without proper fine-tuning on medical images. Consequently, we fine-tuned SAMed on US images to improve its performance in segmenting US data. Additionally, we handled the challenge of limited labeled data by employing transfer learning. SAMed was initially trained on a larger public US dataset and then fine-tuned on our smaller InselUS dataset containing abdominal images. A comparison of the results before and after fine-tuning SAMed on US images highlights the improvements resulting from the use of transfer learning in addressing the issue of small datasets, which is a critical challenge in the medical domain.

While US images are highly effective, their accurate interpretation and identifying organ structures and abnormal regions depend on the physician's expertise. Therefore, developing a reliable AI-powered system for segmenting abdominal US images can aid in correct organs recognition and abnormalities spotting. Our fine-tuned SAMed model on US images can recognize free fluids in these images, which can be a sign of malignancy or infection. Moreover, it can find and classify organs to assist sonographers in evaluating organ condition and size to support disease diagnosis.

While further refinement is needed, our model has the potential to support junior physicians in acute diagnostic settings by providing precise and timely segmentation and classification of medical images. By accurately detecting organs and identifying pathological areas, it can effectively cope with the lack of extensive human expertise and enhance diagnostic efficiency to ensure patients receive appropriate care.

6. Future Works

Future work will focus on enhancing the performance and robustness of our model by expanding the dataset and incorporating images from other emergency departments and diverse clinical settings. This aims to improve the model's accuracy and precision, and in parallel to increase its generalizability across various use cases. Additionally, we plan to evaluate the model's performance in real-world clinical environments to explore its potential for real-time deployment of this model to make sure that it can provide timely and reliable assistance to medical professionals during patient examinations. Moreover, integrating explainability tools into the framework is expected to enable clinicians to better understand the model's decision-making process. This integration will promote transparency, foster trust and enhance the reliability in computer-aided diagnosis in clinical practice.

References

- [1] Gilanie G, Saher A, Batool SN, Khursheed A, Shafique H, Perveen S, Latif A, Sajid M, Saeed M. Digital image processing for ultrasound images: A comprehensive review. *Int J Innov Creat Change*. 2021;15(3)
- [2] Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun*. 2024 Jan;15(1):1234-1238. doi: 10.1038/s41467-024-44824-z.
- [3] Shi Y, Ma J, Yang J, Wang S, Zhang Y. Beyond pixel-wise supervision for medical image segmentation: From traditional models to foundation models. *arXiv*. 2024 Apr 24. Available from: <https://arxiv.org/abs/2404.13239>.
- [4] Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021 Feb;18(2):203-211. doi: 10.1038/s41592-020-01008-z.
- [5] Xie B, Tang H, Duan B, Cai D, Yan Y. MaskSAM: Towards auto-prompt SAM with mask classification for medical image segmentation. *arXiv*. 2024 Mar 14. Available from: <https://arxiv.org/abs/2403.14103>.
- [6] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y, Dollár P, Girshick R. Segment anything. *arXiv*. 2023 Apr 5. Available from: <https://arxiv.org/abs/2304.02643>.
- [7] Huang Y, Yang X, Liu L, Zhou H, Chang A, Zhou X, Chen R, Yu J, Chen J, Chen C, Liu S, Chi H, Hu X, Yue K, Li L, Grau V, Fan D-P, Dong F, Ni D. Segment anything model for medical images. *Med Image Anal*. 2024 Mar;92:103061. doi: 10.1016/j.media.2023.103061.
- [8] Zhang K, Liu D. Customized Segment Anything Model for Medical Image Segmentation. *arXiv*. 2023 Apr 28. Available from: <https://arxiv.org/abs/2304.13785>.
- [9] Lin X, Xiang Y, Yu L, Yan Z. Beyond adapting SAM: Towards end-to-end ultrasound image segmentation via auto prompting. *arXiv*. 2024 Sep 13. Available from: <https://arxiv.org/abs/2309.06824>.
- [10] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. LoRA: Low-rank adaptation of large language models. *arXiv*. 2021 Jun 17. Available from: <https://arxiv.org/abs/2106.09685>