Medical Image Segmentation Based on the CLIP and SAM Base Models

Jingyi Yu

{yv105086@163.com}

Computer Engineering Department, TAIYUAN INSTITUTE OF TECHNOLOGY, Taiyuan, Shanxi, China

Abstract. Medical image segmentation is a key technology for achieving precise medical care. Traditional methods rely on a large amount of labeled data and have limited generalization capabilities. Visual foundation models represented by CLIP and SAM have obtained general capabilities through large-scale pre-training, providing a new paradigm for medical segmentation. CLIP achieves semantic-guided segmentation through image-text alignment, while SAM achieves general segmentation through prompt interaction. Both can effectively reduce the reliance on labeled data. This paper systematically reviews the applications of these two types of models in medical segmentation: first, analyzes the necessity of their applications; second, separately summarizes the technical routes and adaptation progress of semantic-guided methods based on CLIP and prompt-interaction methods based on SAM; finally, discusses their advantages and challenges. The review shows that these models significantly improve data efficiency and cross-domain generalization capabilities, but still need further exploration in medical specificity adaptation and fine-grained segmentation.

Keywords: Base model, Medical image segmentation, CLIP, SAM, Multimodal fusion.

1 Introduction

Segmentation of medical images is a core operation in the extraction of essential structures in a medical image that is essential in the diagnosis of the disease and long-term treatment strategies. The latest methods employed in the mainstream rely on deep learning, although each of them is often limited by two key factors: one is the large volume of manually labeled data required to produce results, which is expensive to produce; the other is the lack of generalization of the models, which is why they are hardly applicable to new devices, institutions, or new pathological changes. Such restrictions have frustrated the extensive use of artificial intelligence in medical practice.

In order to address these restrictions, the base model paradigm provides solutions. The large model that is already trained on a large amount of data to gain general capabilities is known as a base model and can easily be transferred and adapted to different tasks. The general

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365106

knowledge and high transferability gained during pre-training are the core values of base models compared to traditional models that are trained to accomplish specific tasks.

Compared to the medical image analysis, base models differ significantly in their application in the given field. The paper concentrates on visual or visual-language-based frameworks manifested by CLIP and SAM instead of conventional medical segmentation frameworks or alternative forms of base frameworks. This category will be determined by three main characteristics: First, the models are achieved through a large-scale pre-training on non-medical general visual data; Second, they feature high zero-shot or few-shot transfer abilities; and Lastly, the models are natural to interact with the users by prompt means. By means of image-text comparison learning, CLIP establishes visual and language semantic mapping, thus being able to comprehend text cues and retrieve their visual associations; SAM attentively requires no task-specific practice and is able to discover segmentation abilities in a prompt-based task.

The difference between the base model in the medical sector and the traditional model is that the traditional model embraces training fresh to do particular medical tasks or tune on the small medical data, whereas CLIP/SAM-like base models harbor rich general visual priors and kick-start the prior knowledge by adapting to the medical task. Such a difference provides unique benefits of base models in the context of data efficiency, generalization, and interactivity, but introduces significant new challenges in the medical domain adaptation.

This article aims to review systematically the progress achieved so far in terms of applications of CLIP and SAM in medical image segmentation, study their technical principles, adaptations, strengths and weaknesses, and future perspectives. The article will present the issue of medical image segmentation and opportunities offered by base models, then also discuss the medical segmentation one based on CLIP and SAM separately, and then conclude with an overview of current issues and development trends. The proposed review should give the researchers an effective technical framework, as well as facilitate the effective use of visual-based models in the medical field.

2 Current Status of the Dataset

One of the core challenges in medical image segmentation tasks lies in the acquisition and annotation of datasets. The current medical image segmentation datasets are mainly classified into the following categories (Table 1).

Single-modal datasets, such as CT, MRI or ultrasound, which are common imaging data, are usually labeled for a single organ or tumor. Common datasets include LUNA16, which is an open dataset focused on the detection of pulmonary nodules in lung CT images, and its data comes from multiple institutions. Although it has greatly promoted related research, the scale is relatively limited, and there are certain differences in the acquisition parameters and labeling standards of images from different sources, which can easily lead to performance limitations of the model when generalizing to new data. Other examples include LiTS, which is the CT segmentation dataset for liver tumors.

Multimodal data sets are groups of imaging sequences of the same anatomical location used to give complementary data. The general example would be the BraTS dataset on the brain tumor

segmentation challenge that combines different MRI modalities, including T1, T2, T1 enhancement, and FLAIR images of an individual. This multimodal fusion plays a vital role in the complete representation of various tissue areas of brain tumors (including edema, enhancing tumors, and necrotic cores) facilitating the formation of more robust segmentation models, but its data annotation is more complex as well.

Synthetic datasets are intended to overcome the lack of quality large-scale data with real labels. In other research works, attempts have been made to implement generative medical image datasets, like generative adversarial networks, using generative methods. An example that can be given is the publicly available synthetic brain MRI dataset SynthRAD, which imitates brain images in various disease pathological conditions with the help of generative adversarial networks and offers the exact pixel-level labels, which is frequently utilized to test the achievements of segmentation models in case real-labeled data is absent. The use of such synthetic data is useful to enlarge the scale and diversity of the training set within a given context, e.g., during model pre-training or data augmentation, however, the domain gap between synthetic images and real clinical images must be addressed.

Although the existing datasets provide valuable resources for the application of base models, the scale and diversity of these datasets are still insufficient overall. Especially in the medical field, the highly specialized annotation work requiring professional knowledge incurs extremely high costs, which makes the construction of large-scale and fully annotated medical image datasets a huge challenge in reality.

Table 1. Classification and Characteristics of Common Medical Image Segmentation Data Sets.

Modal	Dataset name	Characteristics
Single-modal dataset	LUNA16 (Pulmonary CT Dataset)	For the annotation of a single organ (such as a pulmonary nodule), the data comes from multiple institutions, but the scale is limited, and there are differences in imaging and annotation standards, which affect the generalization of the model.
Multimodal dataset	BraTS (Brain Tumor Segmentation Challenge Dataset)	By integrating multiple MRI sequences such as T1, T2, T1 enhancement, and FLAIR, the tumor structure can be comprehensively depicted, which is conducive to training robust models, but the annotation process is complex.
Synthetic dataset	SynthRAD	By using generative adversarial networks to synthesize brain MRI images with precise labels, it is aimed to expand the training set and verify the generalization ability of the model. However, there are domain differences compared to the real images.

3 Medical Image Segmentation Based on the Base Model

The research on base models originated from the combination of natural language processing and visual tasks. With the continuous evolution of the model structure, it has now become one of the hotspots in the field of deep learning. The following are the main technical routes of base models in medical image segmentation at present.

3.1 Medical image segmentation based on CLIP

Contrastive Language-Image Pretraining (CLIP) is a visual-language base model trained through contrastive learning on a large-scale set of image-text pairs. It enables visual feature and natural language semantic alignment in a common space and, hence, gives the model the zero-shot property of visual conceptual understanding by textual prompts. The usage of medical text knowledge (descriptions in imaging reports, anatomical terms) as high-level semantic priors to inform and constrain the segmentation of medical images is also converted to this feature of CLIP in the medical image segmentation field. Its central strength is in the fact that it allows integrating constrained pixel-level annotation data with enormous and readily available weakly supervised image-text indicators, which offer an alternative manner in obtaining stronger and generalized segmentation when limited labeled data are obtainable in a medical context [1].

Simple fine-tuning of the general CLIP model to the medical sector is also a simple technique of modification. An example is that a study has optimized the image encoder of CLIP on the multi-organ CT segmentation data, allowing the segmentation network of the former to obtain features that can better discriminate the various anatomical structures, which in turn led to a better performance of the latter segmentation network [2]. Subsequent work gave specific suggestions in order to further integrate the medical domain knowledge. As an illustration, Zhang et al. provided contrastive learning pre-training using large quantities of chest X-rays and matching diagnostic findings, and the CLIP model fails to comprehend and react to clinical text prompts, including lung nodules and heart enlargement, with considerable success in improving the specificity of localization and segmentation of chest diseases [3]. One more paper was dedicated to retinal fundus image segmentation. With the help of building a high-quality image-diagnostic reports pair, a specialized visual-language model was trained and was capable of using the semantic information in the text description to enhance the recognition of small targets in the telescopic task of diabetic retinopathy lesions segmentation significantly [4].

3.2 Medical image segmentation based on SAM

The SAM can be considered a universal visual foundation model that is made to segment everything. Its fundamental innovation is the embrace of the interaction paradigm of promptness. The trained model can produce high-quality segmentation masks in real time with simple visual cues, including points and bounding boxes, given by users after they have been trained on a massive amount of segmentation data. In the case of medical image segmentation, two major benefits of SAM design include first, the high zero-shot generalization capability of the design, which can prompt the model to first complete the initial segmentation without re-training on a specific medical data set, which greatly lowers the application threshold; Secondly, its interactivity in that clinical experts are allowed to take part in the segmentation process in the most intuitive manner, which is a combination of professional judgment and the computational efficiency of the design, which is capable of performing successful human-machine collaborative annotation and fine correction [5].

The difference in texture, contrast, and morphology of targets between medical images and natural images is so large that the direct application of the original SAM is likely not produce the necessary clinical accuracy. Thus, much research was devoted to the project of SAM

adaptation in the medical sphere. The first to undergo supervised fine-tuning on scale and diversity (millions of medical image-mask pairs) was MedSAM proposed by Wu et al. and showed a strong performance improvement over the original SAM and current medical-specific architectures on 11 2D and 3D medical image segmentation tasks [6]. To reduce the workload of the interactions between clinicians, research has suggested uncertainty-sensitive guidance schemes, which are clever enough to find areas of uncertainty in model forecasting and subsequently propose the best next point of click action, therefore significantly lowering the amount of interactions that the user needs in order to obtain the desired segmentation outcome [7]. Reflectively, a systematic experimental study was able to evaluate the zero-shot skill of the original SAM across several modalities (CT, MRI, X-ray) and tasks (organ, tumor segmentation) to find out that it has a significant generalization skill on medical images and limitations when it comes to specific structures of the anatomy (e.g., fine blood vessels, blurred lesion boundaries), which serves as a guideline in further targeted optimization [8]. More advanced discovery is centered on the fusion of SAM with particular clinical operations, including the creation of interactive segmentation instruments particularly in the examination of dermatoscopy pictures to facilitate the screening and characterization of skin lesions [9] or the potential of SAM as a real-time exploration in intraoperative ultrasound images in order to provide dynamic anatomical structure guidance in surgery operations [10].

3.3 Technological development trends

With the continuous development of deep learning technology, the technical route of base models in medical image segmentation has also been constantly evolving. Currently, prompt learning (Prompt Learning) has become an important research direction. It guides the model to complete the segmentation task through various methods such as text, points, and boxes. In addition, multimodal fusion technology is gradually emerging. By combining medical images with auxiliary information such as clinical reports and medical records, the base model can better understand the semantic information of diseases and improve the segmentation accuracy.

4 Comparative Analysis and Discussion

CLIP and SAM represent two complementary technical paradigms in medical image segmentation. CLIP is based on visual-language alignment and achieves semantic guidance through text prompts. Its advantage lies in being able to utilize rich medical prior knowledge for zero-shot reasoning, providing a good interpretability foundation for the model. However, its segmentation accuracy is usually limited by the accuracy of the text description, and it has shortcomings in the fine delineation of complex anatomical structures. In contrast, SAM is based on pure visual prompt interaction and has strong general segmentation capabilities, capable of achieving precise boundary segmentation through simple prompts such as points and boxes, with high interaction efficiency. However, its main limitation is the unstable zero-shot generalization ability for special targets in the medical field, and the model's decision-making process lacks explicit semantic correlation.

This is with a trend of integration of the two in accordance with the current technological trend. It will be assumed that by building a multimodal team structure, combining the strengths of semantic comprehension and linking location precision, such as applying CLIP to text report de-scaffolding to create semantic focus points, then refined segmentation and

boundary repair with SAM will be feasible. The integration paradigm offers a realistic way of realizing a smarter and dependable auxiliary segmentation system. Though to accomplish this objective, a number of critical questions remain unanswered, such as how well the cross-modal representations can successfully be aligned, how well the fusion architecture can be optimised in terms of its computational efficiency, and how the fusion architecture has the Generalisation power in real settings in clinical practice. These issues will have a direct impact on the practical applicability of the base model in medical image analysis with the resolution of the challenges.

5 Conclusions

The base model provides a new technical approach for medical image segmentation tasks, especially in cases of scarce data and complex tasks, demonstrating great potential. However, the application of the base model still faces some challenges, such as the difference between medical data and general data distribution, the accuracy and consistency of annotations, etc. In the future, the research on the base model in medical segmentation still needs to address these challenges and explore more efficient training strategies and data augmentation methods. With the advancement of technology, the base model is expected to play a more important role in medical image segmentation and promote the development of precision medicine.

References

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning transferable visual models from natural language supervision. International Conference on Machine Learning. pp. 8748–8763 (2021)

[2] Zhou, Y., Chen, H., Li, Y., Shen, D.: Prior-aware neural network for partially-supervised multi-organ segmentation. Medical Image Analysis. pp. 101996 (2019)

[3] Zhang, Y., Jiang, H., Miura, Y., Langlotz, C. P.: Contrastive learning of medical visual representations from paired images and text. Proceedings of the Conference on Health, Inference, and Learning. pp. 2–25 (2022)

[4] Li, J., Chen, J., Tang, Y., Wang, C.: Vision-language pre-training for boosting scene text detection in the wild. Pattern Recognition. pp. 108866 (2022)

[5] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al.: Segment anything. Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)

[6] Wu, J., Ji, L., Zhang, Y., Xu, Y.: MedSAM: Segment anything in medical images. arXiv preprint arXiv:2304.12306 (2024)

[7] Ma, J., Wang, B., Zhang, L.: Click-efficient medical image segmentation with uncertainty-aware guidance. Medical Image Analysis. pp. 102871 (2023)

[8] Mazurowski, M. A., Dong, H., Gu, H., Ren, Y.: Segment anything model for medical image analysis: An experimental study. Medical Image Analysis. pp. 102514 (2023)

[9] Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data. pp. 180161 (2018)

[10] Moghbel, M., Mashohor, S., Saripan, M. I. B.: Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography. Artificial Intelligence Review. pp. 497–537 (2018)