PathSAM: Enhancing Oral Cancer Detection with Advanced Segmentation and Explainability

Suraj Sood[1], Jawad S. Shah[1], Saeed Alqarni[1,2], Yugyung Lee, PhD[1]
[1]Computer Science, University of Missouri-Kansas City, USA
[2]Computing and Informatics, Saudi Electronic University, Saudi Arabia

Abstract

Building on the success of the Segment Anything Model (SAM) in image segmentation, "PathSAM: SAM for Pathological Images in Oral Cancer Detection" addresses the unique challenges associated with diagnosing oral cancer. Although SAM is versatile, its application to pathological images is hindered by its inherent complexity and variability. PathSAM advances beyond traditional deep-learning methods by delivering superior accuracy and detail in segmenting critical datasets like ORCA and OCDC, as demonstrated through both quantitative and qualitative evaluations. The integration of Large Language Models (LLMs) further enhances PathSAM by providing clear, interpretable segmentation results, facilitating accurate tumor identification, and improving communication between patients and healthcare providers. This innovation positions PathSAM as a valuable tool in medical diagnostics.

Introduction

The Segment Anything Model (SAM)[1] has significantly impacted the field of image segmentation, showing remarkable versatility across a wide range of tasks. However, its application to the nuanced realm of medical imaging, particularly in analyzing pathological images, has encountered distinct challenges. To capitalize on SAM's potential for medical segmentation, MedSAM[2] was developed, employing an extensive dataset of over 1.57 million image-mask pairs. This dataset contains 50 abdomen CT scans and each scan contains an annotation mask with 13 organs. This vast dataset aimed to enhance model accuracy, robustness, and adaptability for medical imaging applications. Despite these enhancements, MedSAM faced limitations in accurately delineating the complex spread patterns of cancer cells within pathological images, a critical aspect for precise cancer diagnosis and effective treatment planning[2].

This paper introduces 'PathSAM: SAM for Pathological Images in Oral Cancer Detection,' a model specifically trained on pathological images, offering a distinct advantage over general models like MedSAM, which is not trained on pathological data. PathSAM also distinguishes itself by implementing a novel centroid-based prompt approach, which enhances the model's capability for accurate tumor detection using minimal labeled data, effectively addressing the unique challenges posed by the variable and intricate nature of cancer cell distributions in pathological samples.

PathSAM's efficacy has been validated using two benchmark datasets for oral cancer segmentation, ORCA[3] and OCDC[4]. These datasets were instrumental in demonstrating PathSAM's ability to meet and exceed the performance standards set by existing state-of-the-art models. The model's superior performance is evidenced by its quantitative and qualitative metrics on these datasets.

A critical advancement of PathSAM is its integration with Large Language Models (LLMs), significantly advancing the model's explainability. This integration enables PathSAM to provide explanations of its segmentation results, marrying technical precision with interpretative clarity. Such a feature is invaluable for clinicians, offering a more comprehensive understanding of the imaging outcomes, thus facilitating enhanced cancer research and more informed patient care.

The contributions of this paper extend beyond the technical improvements of PathSAM. By introducing a specialized segmentation solution for oral cancer characterized by a centroid-based approach, PathSAM addresses a vital need within medical diagnostics. Its exemplary performance on crucial datasets and its efforts in enhancing explainability through LLMs establish PathSAM as an innovative and effective tool in oral cancer diagnostics, contributing significantly to the advancement of medical imaging and the evolution of patient care methodologies.

Background

Recent research in medical image segmentation has focused on developing universally applicable models that can be effectively applied across various tasks, enhancing versatility and consistency. However, foundational models like SAM[1] have struggled to meet the unique demands of medical imaging due to substantial differences from conventional images. Addressing these challenges is central to the development of our model, PathSAM.

While SAM represents a notable advancement in segmentation foundation models, its strategy of utilizing points or bounding boxes to define target areas of segmentation closely parallels the techniques employed in traditional interactive segmentation methods, where user input is often required to guide the segmentation process. However, despite its improved generalization capabilities, SAM's performance is less effective in medical scenarios characterized by ambiguous boundaries or low contrast, revealing a significant gap in the adaptability of current models to the varied and complex landscape of medical imaging tasks.[2,5].

The deployment of MedSAM represents a crucial enhancement, refining SAM's approach to medical image segmentation. By fine-tuning on an extensive dataset, MedSAM has demonstrated superior performance compared to traditional segmentation and specialized models designed for specific imaging modalities[6,7]. Its comprehensive evaluations underscore its potential to transform the approach to medical image segmentation, offering a more versatile and adaptable solution.

PathSAM builds upon these advancements by incorporating Large Language Models (LLMs), specifically utilizing GPT-4, to enrich the model's explainability. This integration serves not just as a technical improvement but also fulfills a critical requirement in medical imaging: rendering complex segmentation results understandable and actionable in clinical settings[8,9]. PathSAM moves beyond the limitations of prior models by achieving accurate segmentation with minimal spatial data, coupled with text-based explanations of the results. This dual capability is paramount in challenging medical scenarios such as oral cancer, where precision and interpretability are essential for informed diagnosis and strategic treatment planning.

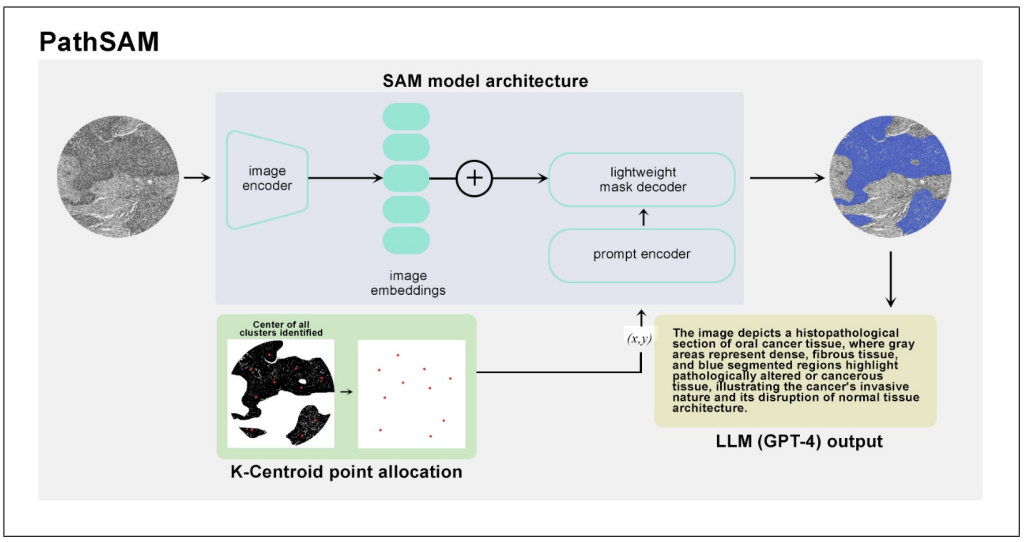

Figure 1. Overview of the PathSAM architecture for oral cancer detection. The process begins with a histopathological slide of cancerous tissue, which is encoded into image embeddings. These embeddings are processed through a lightweight mask decoder, guided by a prompt encoder, to segment the tissue. The K-Centroid point allocation identifies key clusters within the tissue, enhancing segmentation precision. The final output, combined with an LLM (GPT-4), provides detailed descriptions, improving the model's explainability. This architecture demonstrates PathSAM's precision in identifying critical areas, distinguishing it from other models like MedSAM.

Methods

This section highlights two essential enhancements integrated into the SAM model, resulting in the creation of PathSAM—a specialized tool designed for the segmentation and analysis of pathological images in oral cancer

diagnosis. PathSAM focuses on accuracy, precision, and an understanding of the characteristics involved in oral cancer imaging.

SAM Enhancement with K-Centroid Allocation

We enhance the Segment Anything Model (SAM) by integrating a k-centroid allocation algorithm, which significantly improves the identification of Regions of Interest (ROIs) using a select number of critical points (k) instead of relying on a full mask. SAM's ability to accept points as prompts is augmented by this approach, allowing us to use multiple points effectively to target specific regions. K-means clustering is utilized to identify these points or clusters, ensuring they represent the most relevant areas for segmentation. This method reduces the need for extensive manual input, optimizing both accuracy and efficiency in the segmentation process by strategically selecting the best areas for point prompts.

SAM works by taking an input medical image I and producing a segmented output S. It's adaptable to different types of medical imaging M and tasks T, such as prompts like bounding boxes or points, due to its deep learning architecture, which learns how to map input images to segmented outputs based on specific tasks and imaging modalities. Thus, it is learning a mapping $f : I \times T \times M \rightarrow S$, allowing for context and modality-specific segmentation.

$$S = SAM(I; \theta_T, \theta_M) \tag{1}$$

where θ_T and θ_M are the parameters adapting SAM to specific tasks and modalities.

Development of the PathSAM Framework

The development of the PathSAM framework represents an innovative leap in medical image segmentation for oral cancer detection, integrating advanced methodologies for segmentation and understanding. PathSAM distinguishes itself by employing a centroid-based K-Means and SAM strategy for point-prompt training.

Centroid-based K-Means and SAM Approach:

1. **Preprocessing and Point Allocation:** The process starts by analyzing binary masks to identify key areas, using morphological operations to smooth and clarify regions, reducing noise.
2. **Labeling and Region Analysis:** After preprocessing, the image is segmented into distinct regions by labeling contiguous pixel groups. In denser regions, K-means clustering is used to find centroids as focal points, while in sparser areas, all pixel coordinates are used.
3. **Centroid Refinement:** The centroids identified through K-means clustering are refined using an Euclidean distance adjustment, ensuring they align accurately with pixel positions, enhancing precision.
4. **Aggregation and Training:** The refined centroids from each mask are compiled to represent the critical areas. These points, along with the original images, are used in PathSAM's training, focusing on key areas for segmentation using point-prompt methods. These refined points, which are instrumental during the training phase, are also used during inference, ensuring that the model leverages these prompts to achieve accurate segmentation in new, unseen data."
5. **Implementing LLMs for Explainability:** By using GPT-4, PathSAM generates explanations for the segmented regions, providing insights into the outcomes and improving the model's usefulness in clinical settings.

K-Centroid Allocation Using K-Means

To improve ROI precision, we first clean up the image masks using morphological operations, then apply K-Means clustering to select the centroids. Choosing the right centroids (k) is key to accurately representing different areas within the ROIs, with Euclidean distance adjustments ensuring the centroids align precisely with the intended segmentation targets.

The k-centroid allocation algorithm enhances the Segment Anything Model (SAM) with key benefits over traditional segmentation techniques:

Efficiency and Reduced Manual Input

- Optimizes ROI identification by using a small, strategically chosen set of critical points (k).
- Reduces the need for extensive manual annotation, making the process more efficient.

Accuracy and Precision

- Ensures precise representation of distinct ROI areas through k-means clustering and Euclidean distance adjustments.
- This method improves the accuracy of identifying key areas critical for diagnosis and treatment planning.

Balanced ROI Representation

- While other clustering methods like DBSCAN excel in handling noise and identifying clusters of varying shapes and sizes, k-means clustering provides a balanced approach to hitting all ROIs in a pathological image.
- K-means efficiently manages the unique morphological shapes in oral cancer imaging, ensuring each ROI is accurately represented without being overshadowed by noisier data.
- This balanced representation enhances the generalizability and accuracy of the model in identifying and segmenting complex cancerous structures.

Algorithm Definition and Hypothesis Testing: We start with a collection of image masks $M = \{m_1, m_2, \ldots, m_n\}$. Each mask m_i is preprocessed to enhance feature detection. We hypothesize that by using K-Means clustering, adjusted by Euclidean distance, we can isolate and accurately represent the key points within the ROIs.

Algorithm 1 K-Centroid Allocation for ROI Expansion in SAM Using K-Means and Euclidean Distance

Require: Image masks $M = \{m_1, m_2, \ldots, m_n\}$, Number of centroids k
Ensure: Coordinates within ROIs

```
Initialize an empty list C for coordinates.
for each mask m_i ∈ M do
    Apply morphological operations to m_i to reduce noise.
    Label connected regions in the mask.
    Initialize an empty list for coordinates in m_i.
    for each region in m_i do
        Extract all pixel coordinates within the region.
        if pixels in the region ≥ k then
            Apply K-Means clustering to form k clusters.
            for each cluster centroid do
                Refine the centroid to the nearest pixel using Euclidean distance.
            end for
            Add refined centroids to the list for m_i.
        else
            Add all pixel coordinates from the region to the list.
        end if
    end for
    Add coordinates from m_i to the global list C.
end for
Optionally, visualize each mask with its refined coordinates.
Return the list C of coordinates.
```

Our analysis confirms that the K-Means clustering method, augmented by Euclidean distance refinement, achieves precise segmentation with minimal manual intervention. Comparative studies have validated the efficiency and accuracy of our k-centroid allocation strategy in segmenting oral cancer tissue, demonstrating significant improvements over traditional annotation techniques.

Enhancing PathSAM with LLMs for Advanced Explainability

The PathSAM (Pathology Segmentation Model for Oral Cancer) involves integrating it with Large Language Models (LLMs), a step that significantly improves the explainability of oral cancer image segmentation.

This integration benefits clinicians by providing not just segmented images but also explanations of the segmentation results. Unlike traditional models that require clinicians to interpret complex visual outputs independently, PathSAM's use of LLMs bridges the gap between technical results and clinical understanding. This enhancement allows for clearer communication of the findings, supports more informed decision-making, and improves patient communication by translating technical segmentation results into easily understandable language. By doing so, PathSAM not only offers state-of-the-art segmentation accuracy but also makes these results more accessible and actionable in real-world clinical settings.

LLMs, which are at the forefront of natural language processing advancements, work by mapping textual input (T) to descriptive output (D). This can be represented as $LLM : T \rightarrow D$. These models are trained on large collections of text data (C) and are very skilled at processing and generating complex language patterns.

$$D = LLM(T; \phi_C) \tag{2}$$

In this formula, ϕ_C signifies the parameters learned from the training corpus C, equipping the LLM to generate contextually accurate outputs based on T.

This integration is mathematically represented as:

$$E(S(I), L(S(I))) = \begin{cases} \text{Segmentation Output: } S(I), \\ \text{Textual Explanation: } L(S(I)) \end{cases} \tag{3}$$

In this framework, the LLM interprets the segmentation output $S(I)$ and translates it into a explanation $L(S(I))$. This fusion of visual data and textual analysis deepens the understanding of the segmentation results.

By combining PathSAM's segmentation output with LLM-generated textual explanations, each input image I receives a dual-layered elucidation. The segmentation result $S(I)$ is enhanced by its linguistic counterpart $L(S(I))$, represented as $E(S(I), L(S(I)))$. This dual explanation covers both the visual segmentation $S(I)$ and its textual narrative $L(S(I))$.

The integration of LLM-driven explanations into PathSAM greatly expands its utility. Moving beyond conventional visual tools, it offers an analysis of segmentation outcomes. LLMs' textual narratives bring additional insights, rendering the segmentation results more interpretable and actionable, especially in clinical contexts. This advanced explainability is vital for informed clinical decision-making, improved patient communication, and educational endeavors, effectively connecting detailed segmentation results to tangible medical applications.

PathSAM Modeling

1. **Data Preparation and Preprocessing**
   - *Binary Mask Processing:* Simplified the task into binary classification to distinguish non-affected tissue ('0') from affected areas ('1'), streamlining the analysis process. This step helps the model focus specifically on identifying the presence or absence of disease, making the analysis more straightforward and accurate.
   - *Image Formatting:* Converted images to `uint8` and grayscale to enhance memory efficiency, processing speed, and library compatibility. Converting images to `uint8` reduces the amount of data the model has to process, while grayscale ensures that the images are in a simple format that is easier to analyze.
   - *Image Normalization:* Applied min-max normalization to scale pixel values between 0 and 1 for consistent model training. Normalizing the pixel values ensures that the data fed into the model is consistent, which helps improve the model's performance and speed up the training process.
   - *Resizing for SAM:* Standardized image size to 256 x 256 pixels for compatibility with the Segment Anything Model (SAM), ensuring uniform and accurate processing. Resizing the images ensures they all meet the input size requirements of the SAM model, which helps maintain consistency and accuracy during processing.
2. **Data Splitting Structure**

- *Split Ratio:* Adopted a 70:20:10 distribution for Training, Validation, and Testing sets to balance training depth with adequate evaluation resources. We ensured robust and generalized datasets for training, validation, and testing by carefully shuffling the entire dataset before splitting. This approach allowed us to create well-distributed sets that effectively support model training and tuning while providing a reliable benchmark for evaluating the model's performance. The random shuffling ensured that each set is representative of the overall dataset, enhancing the model's ability to generalize across unseen data.
- *Dataset Details:* The ORCA Dataset, consisting of patches from pathological oral cancer images, was divided into 39 training, 11 validation, and 6 testing images. Similarly, the OCDC Dataset was split into 308 training, 88 validation, and 45 testing images. These datasets are well-balanced between affected and non-affected tissues, owing to the unique, erratic, and spread-out shapes of the affected regions. This careful distribution ensures a robust dataset across different phases, providing a strong foundation for evaluating PathSAM's performance in diverse and challenging imaging scenarios.

3. **Model Parameters and Storage Size**
   - *Parameters:* Both ORCA and OCDC dataset models encompass 93,735,472 parameters, showcasing their complexity and detail-capturing capacity.
   - *Model Size:* The storage size for each model is 375.1 MB, highlighting the significant resources needed for their management.
   - *Training Time:* PathSAM was trained on an NVIDIA A6000 GPU, taking approximately 45 minutes per session.
4. **Model Configuration and Training Regimen**
   - *Learning Rate and Optimizer:* Utilized a learning rate of 0.00001 with the Adam optimizer for precise model weight adjustments.
   - *SAM - Segment Anything Model:* Fine-tuned using base weights from the Hugging Face model library, enhancing performance through pre-trained models.
   - *Loss Function:* Employed Dice Loss for its effectiveness in segmentation tasks and handling class imbalance, which is particularly beneficial in scenarios with irregularly shaped regions.
   - *Epochs and Early Stopping:* Trained for 50 epochs with an early stopping mechanism set at a patience of 5 epochs to prevent overfitting and conserve resources.
5. **Integration of SAM with GPT-4 for Enhanced Explainability**
   - Merges SAM's segmentation capabilities with GPT-4's natural language processing for oral cancer detection.
   - Generates narrative interpretations of segmentation outcomes, providing clinicians with textual alongside visual data.
   - Enhances diagnostic accuracy and supports informed treatment planning.
   - Improves transparency in AI diagnostics, making complex models more accessible and reliable for clinical use.
   - Advances oral cancer diagnostics and bridges AI technology with real-world healthcare applications.

Evaluation Metrics for Medical Image Segmentation

The performance of medical image segmentation models is commonly evaluated using three key metrics: Accuracy, Dice Similarity Coefficient, and Mean Intersection over Union (mIoU). These metrics quantitatively assess how well a segmentation model identifies and delineates regions of interest in medical images.

Accuracy (Acc%): Accuracy measures the overall correctness of the segmentation model. It is calculated as the ratio of correctly predicted pixels (true positives and true negatives) to the total pixels in the image:

$$\text{Accuracy (Acc\%)} = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Number of Pixels}} \times 100 \tag{4}$$

A high accuracy means the model is good at correctly identifying both the regions of interest and the background. For instance, in medical imaging, ensures that both healthy tissue (non-interest areas) and affected areas (regions of interest) are accurately classified.

Dice Similarity Coefficient (Dice%): The Dice coefficient measures the similarity between the predicted segmentation and the actual (ground truth) segmentation. It is particularly useful for understanding how well the model overlaps with the true regions of interest. It is calculated as:

$$\text{Dice} = \frac{2 \times |\text{True Positives (TP)}|}{2 \times |\text{True Positives (TP)}| + |\text{False Positives (FP)}| + |\text{False Negatives (FN)}|} \times 100 \tag{5}$$

The Dice coefficient ranges from 0 to 1 (or 0% to 100%), where 1 (or 100%) means perfect overlap between the predicted and actual segments. This metric is especially important in medical imaging as it ensures the precision of the segmentation in identifying the exact area of interest, such as a tumor.

Mean Intersection over Union (mIoU%): The Mean Intersection over Union calculates the average ratio of the overlap to the union of the predicted and actual segments across all classes. It gives a broader view of the model's performance by considering all classes within the image:

$$\text{mIoU\%} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{|\text{True Positives (TP)}_i|}{|\text{True Positives (TP)}_i| + |\text{False Positives (FP)}_i| + |\text{False Negatives (FN)}_i|} \right) \times 100 \tag{6}$$

Where N is the number of classes, the IoU score ranges from 0 to 1 (or 0% to 100%), with 1 (or 100%) indicating perfect agreement between predicted and actual segments. This metric is essential for assessing the model's performance across different tissues or regions in medical images.

These metrics provide a thorough evaluation of a model's performance in medical image segmentation, covering its overall accuracy, ability to precisely segment specific structures, and adaptability to different conditions or types of images. This comprehensive assessment ensures that the model can be trusted for clinical applications, providing reliable and accurate results for medical professionals.

Baseline Models for Oral Cancer Segmentation

This section reviews the progression of key medical image segmentation models that have paved the way for the introduction of PathSAM, a model specifically developed for the precise detection of oral cancer. Highlighting the evolution of this field, we discuss several foundational models that have significantly influenced segmentation techniques through their innovative architectures, thereby enhancing the accuracy, efficiency, and complexity management in segmentation tasks.

- *U-Net*[4,10] is recognized for laying the foundational architecture in medical image segmentation. Its encoder-decoder design set a new benchmark, inspiring a series of advancements in the field.
- *U-Net + RGB*[10] extends the capabilities of the original U-Net framework by integrating RGB color space data. This addition seeks to bolster segmentation accuracy by leveraging the additional image detail provided by color information.
- *U-Net++*[11] enhances the U-Net architecture further, featuring nested, skip connections. This design facilitates more nuanced feature integration across different resolution scales, contributing to improved segmentation outcomes.
- *U^2-Net*[12] introduces a deep nested U-structure to extract highly detailed features, thereby addressing the demands of complex segmentation challenges with its elaborate architecture.
- *Att_U-Net*[13] advances the U-Net model by embedding attention mechanisms. This adaptation shifts the model's focus towards the most relevant features within an image, aiming to refine segmentation precision.
- *OCU-Net*[14] represents a model that explicitly targets oral cancer segmentation. By adopting and adapting the U-Net architecture to incorporate attention mechanisms, OCU-Net zeroes in on the unique challenges of oral cancer detection.

While traditional methods have advanced medical image segmentation, they lack the ability to incorporate minimal spatial prompts like points into their models. SAM, however, can integrate these prompts due to its extensive and diverse training data, enabling effective generalization across different imagery. We've further enhanced SAM by using a k-means clustering approach to accurately identify points, ensuring precise

Region of Interest (ROI) identification and improving segmentation outcomes. This makes SAM, particularly PathSAM, a more versatile and advanced tool in medical image segmentation than traditional methods.

In addition to these models, two notable developments are specialized versions of SAM: MedSAM, which is tailored for medical imaging, and the Segment Anything Model (SAM) for Digital Pathology.

- *MedSAM*[2] signifies a leap in universal medical image segmentation, developed over an extensive dataset to address a broad spectrum of medical imaging tasks, including various cancer types and imaging modalities. Its approach aims for wide applicability and enhanced accuracy across different medical segmentation challenges.
- *SAM for Digital Pathology*[15] targets explicitly the unique demands of digital pathology, including tumor detection. It leverages a zero-shot learning approach to segment whole slide images without prior domain-specific training, showcasing potential in areas where labeled data are scarce.

These advancements highlight a significant evolution in medical image segmentation, leading to specialized models like PathSAM tailored for specific conditions like oral cancer.

Results and Evaluation

In Figure 2, the performance metrics of the model are illustrated across various cluster configurations, revealing: (a) **Accuracy:** Training and validation accuracy peak at 10 centroid points, indicating an optimal balance in the model's capacity to generalize. (b) **Loss:** The model's efficiency is underscored by the lowest loss values at a configuration of 10 clusters, mirroring the optimal accuracy results. (c) **Testing Metrics:** The accuracy, precision, sensitivity, and specificity stability at 10 clusters suggest robust performance and model reliability. (d) **Training Time and Convergence:** The model consistently reached its best performance quickly when using 10 clusters, which means it learned effectively without wasting time or resources. This balance made the 10-cluster configuration the most efficient for training. The consistent superiority of the 10-centroid configuration across various metrics underscores its effectiveness as the optimal choice for the model's performance compared to the other number of centroids tested. This effectiveness is maintained during inference, where the same centroid-based points guide the model in accurately segmenting new images.

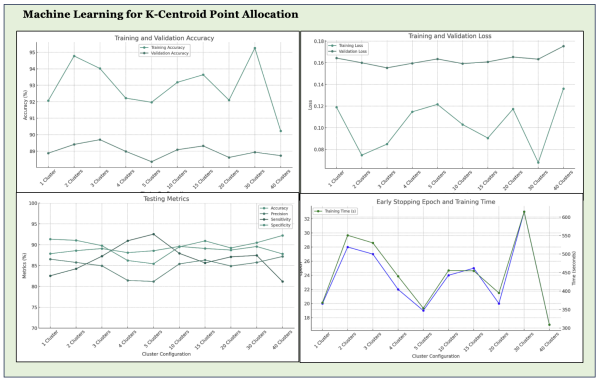

Figure 2. Evaluation of PathSAM's centroid learning performance, highlighting the optimal configuration at 10 centroid points for Training and Validation Accuracy, Loss, Testing Metrics, and the relationship between Early Stopping Epoch and Training Time.

PathSAM effectively competes with state-of-the-art segmentation models by utilizing minimal spatial prompts, such as precisely placed points, to achieve high accuracy. On the ORCA and OCDC datasets, PathSAM matches or exceeds the performance of traditional models that rely on fully labeled masks. By employing K-means clustering to identify Regions of Interest (ROIs), PathSAM enhances SAM's segmentation accuracy with significantly less manual effort. This approach demonstrates PathSAM's strong capability in oral cancer detection, particularly in scenarios where labeled data is limited, ensuring that high performance is maintained even with minimal annotation.

Table 1 presents a performance comparison of segmentation models on the ORCA and OCDC datasets, focusing on Accuracy, Dice Similarity Coefficient, and Intersection over Union (IoU). While traditional deep segmentation methods, such as U-Net and its variants, demonstrate solid performance, the SAM-based approach exemplified by the proposed PathSAM model stands out for its precision in oral cancer segmentation. This distinction underscores the effectiveness of PathSAM in handling complex segmentation tasks, making it a promising choice for accurate oral cancer detection with minimal labeled spatial data.

Table 1. Average performance on the ORCA and OCDC datasets, comparing models using Accuracy (Acc %), Dice Similarity Coefficient (Dice %), and Intersection over Union (IoU %).

Type	Method	ORCA Dataset			OCDC Dataset		
		Acc	Dice	IoU	Acc	Dice	IoU
Traditional	U-Net [4,10]	67.00	68.00	58.00	97.30	91.00	85.20
	U-Net + RGB [10]	80.00	80.00	66.00	-	-	-
	U-Net++ [11]	86.61	78.07	66.23	97.30	90.90	83.32
	U^2-Net [12]	88.21	81.88	71.07	97.77	92.49	86.02
	Att_U-Net [13]	90.29	84.28	74.91	97.45	91.32	84.03
	OCU-Net [14] (ours)	90.02	84.36	75.04	97.91	92.92	86.77
	OCU-Net^m [14] (ours)	**90.98**	86.14	**77.10**	**98.07**	**93.49**	**87.78**
SAM	PathSAM Centroid-based (ours)	89.59	**86.52**	76.76	93.15	90.46	83.61

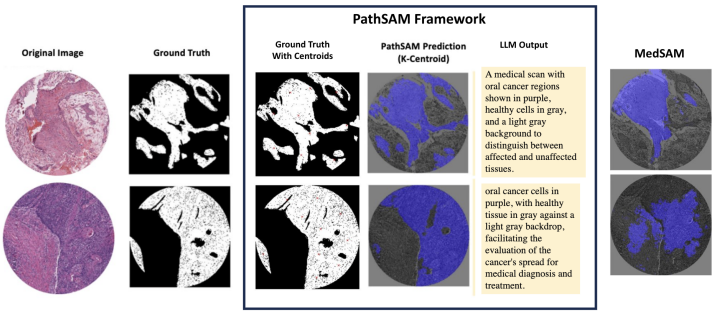

Figure 3. Comparison of segmentation methods for oral cancer detection, featuring original histopathological slides, ground truth annotations, K-Centroid masks, PathSAM's predictions, and MedSAM's predictions. Differences in segmentation quality highlight PathSAM's precision in identifying critical areas and its improved explainability through LLM integration.

Figure 3 shows a comparison of segmentation techniques for oral cancer detection: (a) Original Image: Histopathological slides used for analysis. (b) Ground Truth: Standard reference showing the actual distribution of cancerous and non-cancerous tissues. (c) K-Centroid Mask: The mask derived from the ground truth for precise segmentation. (d) PathSAM Prediction: Segmentation by PathSAM using the K-Centroid approach to identify critical areas. (e) LLM Output: Interpretations from PathSAM using LLMs, distinguishing cancerous regions (gray) from healthy cells (purple). (f) MedSAM: Results from MedSAM for direct comparison with PathSAM. This visual comparison underscores PathSAM's superior segmentation and enhanced explainability, particularly in oral cancer diagnostics.

Discussion

This section addresses the challenges in developing PathSAM, particularly its reliance on annotated data for determining centroids, a key aspect of the Segment Anything Model (SAM). This dependence can limit its use, especially when annotated data are scarce. To overcome this, we are exploring alternative methods for point allocation, reducing PathSAM's reliance on pre-annotated datasets and improving its adaptability to various medical imaging scenarios. One approach involves deep reinforcement learning frameworks to optimize ROI selection with a reward-based system. However, transitioning to these new methods poses challenges in maintaining the accuracy required for clinical diagnoses. Our ongoing work reflects our commitment to enhancing PathSAM's effectiveness in medical imaging.

While PathSAM shows excellent results on the ORCA and OCDC datasets, its ability to work well in different medical settings requires further validation. These datasets primarily focus on oral cancer, which means that additional testing is necessary to determine if PathSAM can effectively handle other types of pathological imaging across various cancers. Future research should incorporate a variety of datasets from different cancers and medical imaging methods to thoroughly evaluate PathSAM's flexibility and robustness. Implementing cross-validation techniques will ensure the model's performance remains consistent across diverse clinical environments, preventing it from becoming too tailored to specific datasets. Extensive clinical trials in real-world healthcare settings will also provide valuable insights into PathSAM's practical use and reliability. These steps are essential to establishing PathSAM as a dependable tool for medical image segmentation in various clinical applications.

Ethical Considerations and Real-World Clinical Usage

Ethical considerations play a crucial role in the deployment of AI in healthcare. Ensuring the transparency and explainability of PathSAM's decisions is essential for building trust among clinicians and patients. Moreover, maintaining patient confidentiality and data security is paramount, necessitating strict adherence to data protection regulations. The potential for inherent biases in the training data to affect model performance underscores the need for continuous evaluation and updates. Extensive clinical trials are imperative to validate PathSAM's reliability and effectiveness in real-world healthcare environments, ensuring that it can provide consistent and accurate diagnostic support across diverse clinical settings.

Conclusion

In this study, we introduced PathSAM, a segmentation model designed to enhance oral cancer detection. PathSAM extends the Segment Anything Model (SAM) by incorporating centroid-based K-means clustering for precise ROI identification, reducing manual input and improving segmentation accuracy. It also integrates Large Language Models (LLMs) to explain segmentation results, aiding clinical decision-making. Our results show superior performance on the ORCA and OCDC datasets, with near or higher accuracy and Dice Similarity Coefficient compared to traditional methods, all while using minimal labeled spatial data. PathSAM's precise ROI identification and enhanced explainability make it a valuable tool in medical image segmentation. Future research should focus on expanding its application to other cancers, developing unsupervised learning techniques, and conducting clinical trials to validate its effectiveness in real-world settings, ultimately contributing to improved diagnostics and patient care.

References

1. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment Anything. In: ICCV; 2023. .
2. Ma J, et al. MedSAM: A Foundation Model for Universal Medical Image Segmentation. Nature Methods. 2024;18:203-11.
3. Martino F, Bloisi DD, Pennisi A, Fawakherji M, Ilardi G, Russo D, et al. Deep learning-based pixel-wise lesion segmentation on oral squamous cell carcinoma images. Applied Sciences. 2020;10(22):8285.
4. dos Santos DF, de Faria PR, Travencolo BA, do Nascimento MZ. Automated detection of tumor regions from oral histological whole slide images using fully convolutional neural networks. Biomedical Signal Processing and Control. 2021;69:102921.
5. Antonelli M, et al. The Medical Segmentation Decathlon. Nature Communications. 2022;13:4128.
6. Isensee F, et al. nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation. Nature Methods. 2021;18:203-11.
7. De Fauw J, et al. Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease. Nature Medicine. 2018;24:1342-50.
8. Ouyang D, et al. Video-Based AI for Beat-to-Beat Assessment of Cardiac Function. Nature. 2020;580:252-6.
9. Dosovitskiy A, et al.; OpenReview.net. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations. 2020.
10. Pennisi A, Bloisi DD, Nardi D, Varricchio S, Merolla F. Multi-encoder U-Net for Oral Squamous Cell Carcinoma Image Segmentation. In: 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA). IEEE; 2022. p. 1-6.
11. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer; 2018. p. 3-11.
12. Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M. U2-Net: Going deeper with nested U-structure for salient object detection. Pattern recognition. 2020;106:107404.
13. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:180403999. 2018.
14. Albishri A, Shah SJH, Lee Y, Wang R. OCU-Net: A Novel U-Net Architecture for Enhanced Oral Cancer Segmentation. arXiv preprint arXiv:231002486. 2023.
15. Deng R, Cui C, Liu Q, Yao T, Remedios LW, Bao S, et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. arXiv preprint arXiv:230404155. 2023.