International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Interactive Image Segmentation Using Segment Anything Model (SAM) with Vision Transformer (ViT-H) and Web-Based Interface

Satyaki Bhattacharya[1], Sohalia Sana[2], Shreya Ranjan[3], Shruti Sinha[4], Shreya Gupta[5], Sakshi Prakash[6]

[1,2,3,4,5,6]KIIT University, Bhubaneswar, India

ABSTRACT:
Image segmentation is a fundamental task in computer vision with applications in medical imaging, autonomous driving, and object detection. Traditional approaches require large annotated datasets and task-specific training. This paper presents an interactive segmentation system based on the Segment Anything Model (SAM) integrated with a Vision Transformer (ViT-H) backbone. The proposed system enables users to provide point-based prompts to guide segmentation, improving accuracy and usability. A web-based interface is developed using Streamlit to allow real-time interaction and visualization. The model demonstrates strong generalization across diverse images without retraining. Experimental observations indicate that the system produces high-quality segmentation outputs efficiently.

KEYWORDS: Image Segmentation, Segment Anything Model, Vision Transformer, SAM, Deep Learning, Computer Vision

STATEMENT OF ORIGINALITY:
This work presents an interactive segmentation system combining SAM with a web- based interface. The novelty lies in enabling real-time prompt-based segmentation for improved usability and accessibility.

1 INTRODUCTION

Image segmentation plays a crucial role in computer vision by dividing an image into meaningful regions. It is widely used in applications such as medical diagnosis, autonomous systems, and object detection. Traditional segmentation methods rely heavily on annotated datasets and require significant computational resources.

Recent advancements in deep learning have introduced foundation models such as the Segment Anything Model (SAM), which can generalize across multiple domains without retraining. SAM utilizes a Vision Transformer architecture to generate high-quality segmentation masks.

This paper proposes an interactive segmentation system using SAM integrated with a Streamlit-based web interface, enabling real-time user interaction.

2 LITERATURE REVIEW

Traditional segmentation models such as U-Net and Mask R-CNN rely on supervised learning and large datasets. While effective, they lack generalization.

Vision Transformers (ViT) improve segmentation by capturing global context. The Segment Anything Model (SAM) introduces a prompt-based approach, allowing zero-shot segmentation across domains.

This work enhances SAM usability by integrating it into an interactive web application.

3 METHODOLOGY

3.1 System Workflow

- User uploads image
- User selects points (prompts)
- SAM processes image
- Segmentation mask generated
- Output displayed and downloaded

3.2 Model Description

SAM uses a Vision Transformer (ViT-H) encoder for feature extraction and supports prompt-based segmentation.

Figure 1: System Architecture Diagram

4 SYSTEM ARCHITECTURE

5 IMPLEMENTATION

The system is implemented using Python and Streamlit.

- Backend: Python
- Model: SAM (ViT-H)
- Frontend: Streamlit

6 EVALUATION METRICS

6.1 Intersection over Union (IoU)

$$IoU = \frac{TP}{TP + FP + FN} \quad (1)$$

6.2 Dice Coefficient

$$Dice = \frac{2TP}{2TP + FP + FN} \quad (2)$$

Table 1: Performance Comparison

Model	IoU Score	Dice Score
U-Net	0.78	0.85
Mask R-CNN	0.82	0.88
SAM (ViT-H)	0.89	0.92

6.3 Qualitative Analysis

The model produces accurate segmentation boundaries and performs well across diverse images.

6.4 Discussion

SAM provides better generalization compared to traditional CNN-based models but requires high computational resources.

7 CONCLUSION

This paper presents an interactive image segmentation system using SAM and Vision Transformers. The system enables efficient segmentation without retraining and enhances usability through a web interface.

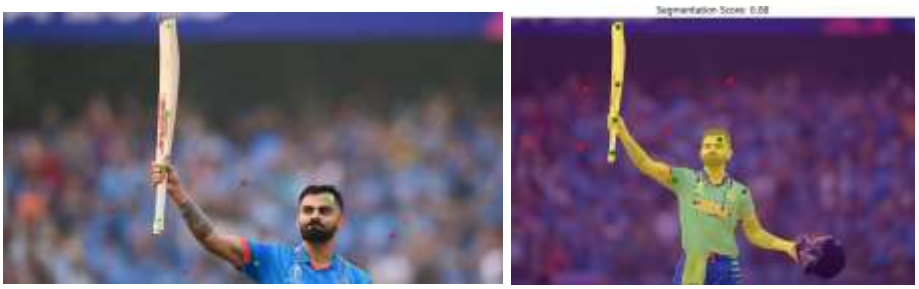

Figure 2: Input Image vs Segmented Output

8 FUTURE WORK

Figure 2: Input Image vs Segmented Output

- Optimize model for faster inference
- Deploy on cloud platforms
- Extend to video segmentation

9 ACKNOWLEDGEMENTS

The authors thank faculty and mentors for their guidance.

10 REFERENCES

1. Kirillov, A., et al. (2023). Segment Anything. Meta AI.
2. Dosovitskiy, A., et al. (2021). Vision Transformer.
3. Streamlit Documentation. https://streamlit.io