Segmentation Analysis of Prompting Data: Comprehensive Study of UNet, DeepLabv3+ and Efficient SAM Models

Kasmira Kurnala
Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, USA
kkurnala3@gatech.edu

Lakshmi Grishma Vennapusa
Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, USA
lvennapusa3@gatech.edu

Janaki Devi Palanki
Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, USA
jpalanki3@gatech.edu

***Abstract*—This paper presents a comprehensive study investigating various segmentation models and techniques, aiming to enhance object segmentation accuracy. Specifically, we focus on employing the UNET and DeepLABV3+ segmentation models to improve object segmentation compared to conventional techniques. Our evaluation encompasses diverse datasets with corresponding masks, enabling a thorough comparison of model performance. Additionally, we introduce an innovative approach aimed at reducing the number of annotations required for predicting the mask by incorporating the efficient SAM. Through this comparative analysis, we assess the efficacy of UNET and DeepLABV3+ in accurately segmenting objects, providing insights into their applicability and potential advancements in the field of computer vision. Furthermore, we propose an efficient model capable of achieving high accuracy while minimizing the number of prompts required for annotation.**

***Index Terms*—UNET, DeepLABV3+, Efficient SAM, SAM[3], tinySAM[4], EfficientSAM[2].**

I. Introduction

Recent years have witnessed significant advancements in computer vision, particularly in image segmentation, which is crucial for applications like medical analysis, autonomous driving, and object recognition. Our research aims to enhance object segmentation accuracy, a fundamental task in computer vision, by investigating the effectiveness of prominent models like UNET, DeepLabv3+, and Efficient SAM compared to conventional techniques. Motivated by the need for robust segmentation methods capable of accurately delineating objects within complex images, we seek to leverage these advanced models to push the boundaries of object segmentation. Through rigorous experimentation on diverse datasets accompanied by ground truth masks, we aim to provide quantitative insights into segmentation performance across different scenarios, contributing to the development of more reliable computer vision systems.

II. Methodology

A. U-Net Architecture

U-Net's architecture[5] is distinguished by its unique combination of a contracting path and an expansive path. The contracting path, composed of encoder layers, is responsible for capturing contextual information and reducing the spatial resolution of the input. Through convolutional operations, these encoder layers create increasingly abstract representations of the input, akin to the feedforward layers in other convolutional neural networks.

On the other hand, the expansive path, comprising decoder layers, decodes the encoded data while maintaining the spatial resolution of the input. These decoder layers upsample the feature maps, allowing for the localization of features. Crucially, skip connections from the contracting path are employed to preserve spatial information lost during the contraction, enhancing the accuracy of feature localization by providing the decoder layers with detailed contextual information. U-Net Architecture representation:

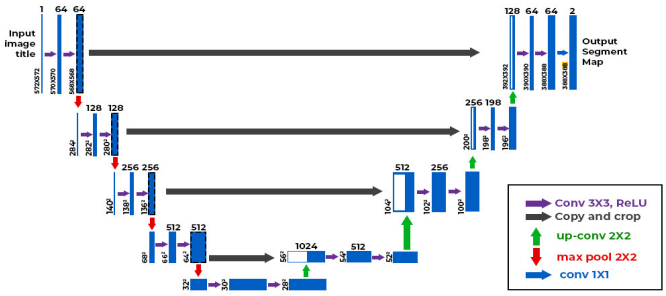

Fig. 1: U-Net Architecture

B. DeepLab V3+ Architecture

DeepLabv3+ [1], introduced in the paper "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," merges the Atrous Spatial Pyramid Pooling (ASSP) from DeepLabv1 with the Encoder-Decoder Architecture from DeepLabv2. Dilated convolutions, featuring a dilation rate parameter 'r', are pivotal. This parameter controls the spacing between kernel points, thus regulating the receptive fields of convolution layers. Atrous convolution enables filters to examine larger input areas without sacrificing spatial

resolution or increasing kernel size. This addresses the spatial resolution reduction inherent in traditional convolutions and max-pooling layers, facilitating more comprehensive feature extraction.

The Mathematical representation of this DeepLab V3 is :

$$\{y[i] = \sum_{k} x[i + r * k]w[k] \quad (1)$$

where x is the input feature map, y be the output and w be the filter and r corresponds to the dilation rate.

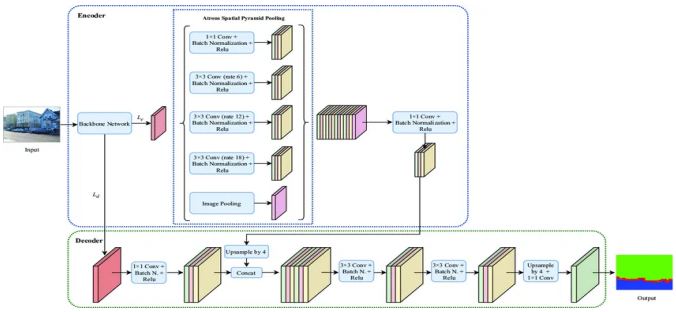

Fig. 2: DeepLab V3+ Architecture

C. Efficient Segment Anything Model

Efficient-SAM[2] is a lightweight iteration of the Segment Anything Model (SAM), prioritizing high performance in segmentation tasks while minimizing computational costs. It outperforms models like MobileSAM and SAM-MAE-Ti in zero-shot transfer, particularly in segmenting objects from single points or boxes. In instance segmentation tasks, it surpasses FastSAM and MobileSAM on COCO and LVIS datasets, demonstrating efficiency with fewer parameters. Qualitative evaluations confirm its strong segmentation capabilities across various prompts, offering competitive performance compared to the original SAM but with reduced complexity. Moreover, Efficient-SAM excels in salient instance segmentation, making it beneficial for applications such as assisting individuals with hand impairments.

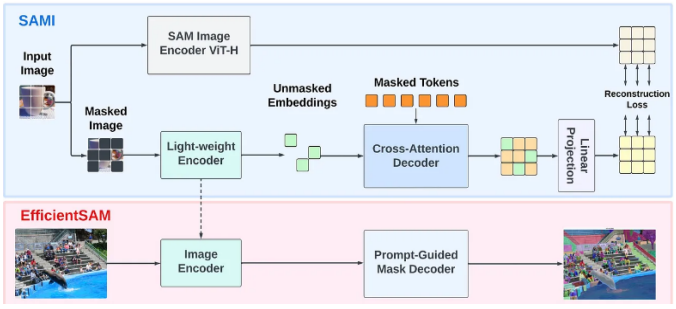

Fig. 3: Efficient SAM Architecture

Efficient SAM-Tiny and Efficient SAM-Small are variations of SAM optimized for computational efficiency. SAM-Tiny prioritizes minimal memory and computational footprint, suitable for edge devices or real-time processing. SAM-Small offers a balance between efficiency and performance, catering to applications with more flexibility in computational resources, such as large-scale inference or resource-constrained devices like tablets.

The impact of finetuning steps on EfficientSAMs was explored, revealing decent performance even at 0.1 epoch and significant gains at 1 epoch. EfficientSAM-S achieved a final performance of 76.9 mIoU, only 1.5 mIoU lower than SAM. This underscores the adaptability and robustness of EfficientSAMs across various finetuning steps, further enhancing their applicability in real-world scenarios.

III. Experimental Analysis

A. Exploratory Data Analysis (EDA)

We have setup the data for the experiments by the following steps.

1. We have analysed the pixel intensity and nonzero proportions.
2. Normalized the data as required and performed the data augmentation for the images.
3. Performed other required preprocessing steps to clean the data.

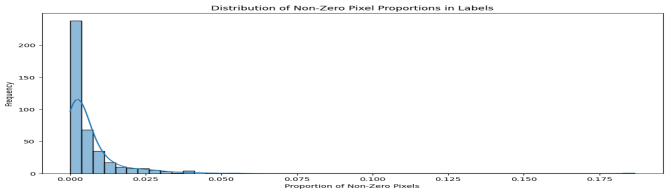

Fig. 4: Distribution of non-zero pixel proportions in labels

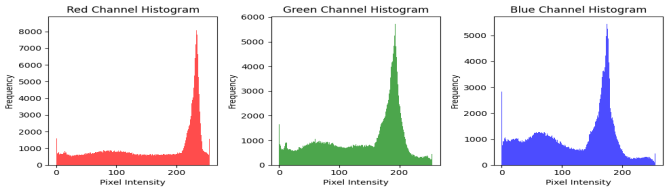

Fig. 5: Pixel Intensity

B. U-Net

U-Net typically yields superior results in datasets involving semantic segmentation of structures with well-defined boundaries, such as medical images and certain object recognition tasks. Hence, it tends to perform well in datasets like Skin(Fig-8) and average results in datasets like Cow, Cat, Stop sign(in some scenarios- Fig-6), and Polyp, where the objects have distinct shapes and textures conducive to U-Net's architecture. Conversely, U-Net might yield worse results in datasets like Clock, Bird, Bus, Baseball Bat, and Tie, and in some scenarios of Stop sign(Fig-7) where the objects may have irregular shapes, occlusions, or varying backgrounds. Additionally, datasets like Skin and Salt Dome, which may have complex textures or ambiguous boundaries, might also present challenges for U-Net, potentially resulting in average or below-average performance.

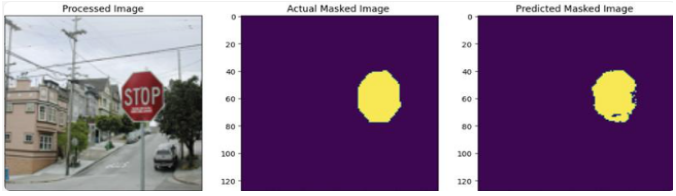

Fig. 6: U-Net gave Average Mask Prediction for some Stop Sign images

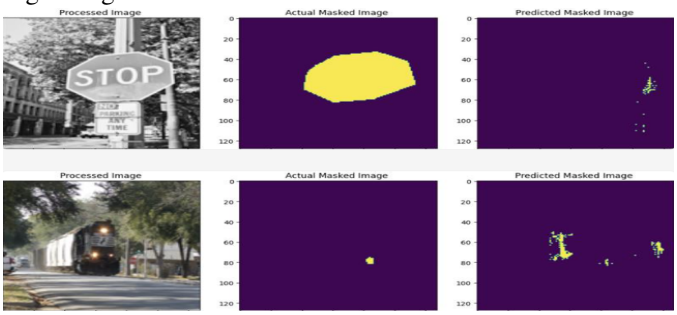

Fig. 7: U-Net gave Worst Mask Prediction for some Stop Sign images

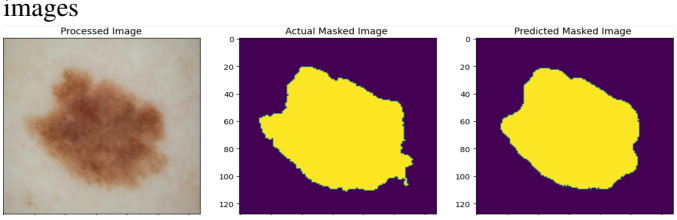

Fig. 8: U-Net gave best predicted Mask for Skin

C. DeepLabV3+

We have integrated DeepLabV3+ with ResNet50v2, which is added as the backbone network in the DeepLabV3+ architecture. We chose this architecture with an idea to leverage the powerful feature extraction capabilities of ResNet50v2 and sophisticated semantic segmentation architecture of DeepLabv3+ in our model.

We have trained and tested this model with provided 16 datasets and prompting results. Out of the analsysis we made, it is observed that though the model has the benefits of two efficient architectures, it still had some discrepancies in predicting the masks for few datasets. The observations are as follows:

1. This model performed extremely well on the datasets such as stop sign, breast(Fig-9) where the images are distinct in features from the background. We have also identified that this prediction can be improved if we had a better preprocessed data.

2. There are instances where model performed well only on few images in the dataset. In datasets like cow(Fig-10), cat, bus and bird, it was able to predict better mask only in scenarios of clean environment where the image is distinguishable from the background.

3. For other datasets, like base ball it was not able to predict well(Fig-11). Here the images are very small and they align with other objects in the image or match with the features of other objects in the image.

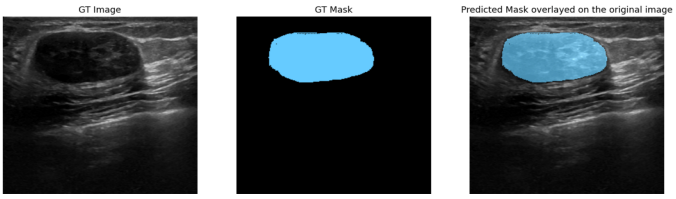

Fig. 9: DeepLabV3 gave best predicted Mask for Breast

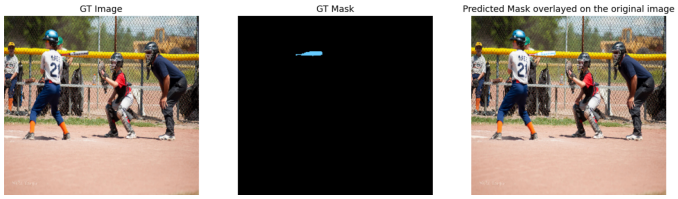

Fig. 10: DeepLabV3 gave worst predicted Mask for Baseball Bat

Fig. 11: DeepLabV3 gave moderate predicted Mask for Cow

D. Efficient SAM

Analysis of Mean IOU values across various models reveals that the Small Efficient SAM consistently achieves IOUs close to 0.9 across all datasets, indicating its superior capability in accurately segmenting objects/images compared to the Tiny Efficient SAM and student-generated masks. Additionally, the quality of prompts provided by students significantly influences segmentation accuracy, with fewer points yielding better results. Specifically, in the Breast dataset, prompts from student 3 (st3) demonstrated superior performance. For the Bus dataset, student 1 (st1) prompts were more effective. In the Cow dataset, student 2 (st2) prompts outperformed others. Similarly, student 3 (st3) prompts excelled in the Dog dataset, while in the Dolphin below dataset, student 3 (st3) prompts were superior. For the Dolphin above dataset, student 2 (st2) prompts stood out. Lastly, in the Skin dataset, student 1 (st1) prompts yielded better segmentation results.

Observations from Efficient SAM-Tiny and Efficient SAM-Small are that

1. Efficient SAM-Tiny and SAM-Small achieve higher mIOU scores than manual prompting data with fewer prompts, showcasing efficient segmentation performance.

2. We have compared the mIOU of students for a particular dataset against the mIOU achieved from the manual prompts

and drawn an inference on which student has done the better prompting.

We can also observe that the segmentation using Efficient SAM on the clock dataset is better than that on U-Net(Fig-15 and Fig-16).

Fig. 12: On Cow dataset, Efficient SAM-Small performed better than Efficient SAM-Tiny

Fig. 13: Bus Dataset , Efficient SAM-Small performed better than Efficient SAM-Tiny

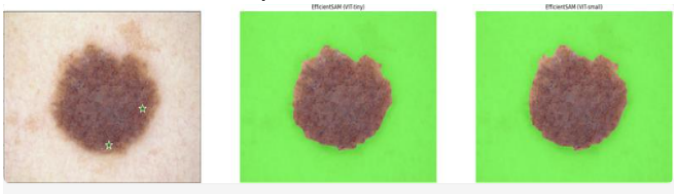

Fig. 14: Skin Dataset, Both Models performed equally

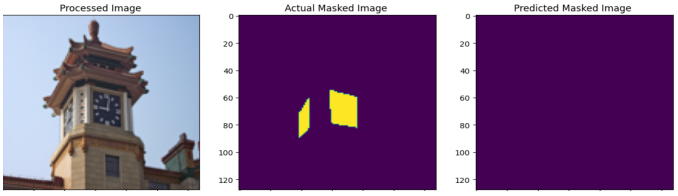

Fig. 15: U-Net gave worst predicted Mask for Clock

Fig. 16: Efficient SAM has performed better than U-Net

Fig. 17: On Clock Dataset, comparision of performance of U-Net and Efficient SAM model

TABLE I: Means IOU's of SAM, Tiny Efficient SAM and Small Efficient SAM

DatasetsDatasetsDatasetspt¡ -Datasetspt¿	st1			st2			st3		
	μ	TE-SAM	TS-SAM	μ	TE-SAM	TS-SAM	μ	TE-SAM	TS-SAM
Breast	0.7148	0.8540	0.9268	0.7581	0.8702	0.9362	0.7651	0.8385	0.9362
Bus	0.7717	0.7598	0.9662	0.8005	0.774	0.9514	0.8096	0.6602	0.9498
Cow	0.8367	0.8090	0.92024	0.8394	0.83944	0.9614	0	0	0
Dog	0.9036	0.8244	0.9209	0.8557	0.8540	0.9268	0.8628	0.6961	0.9243
Dolphin below	0.841	0.9769	0.99943	0.8119	0.9737	0.9868	0.8024	0.9789	0.9956
Dolphin Above	0.7185	0.9840	0.9812	0.7859	0.9781	0.9901	0.7569	0.9727	0.9924
Skin	0.5824	0.7949	0.962749	0.7172	0.8886	0.9528	0.4977	0.8470	0.9528

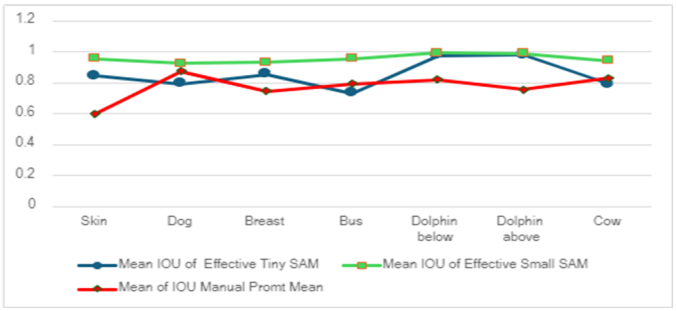

Fig. 18: Mean IOU of Efficient SAM Comparision Graph

IV. Individual Contributions

The authors contributed to this work as follows

Kasmira Kurnala:

1. Conducted exploratory data analysis (EDA) of the dataset, studying its characteristics and distributions.
2. Led the study and analysis of the DeepLabV3+ model architechture, implemented the code ,documented the observations and challenges of the model after performing various experiments while writing code for the integration required.
3. Wrote the Exploratory Data Analysis, Experimental Analysis of DeepLabv3+ Architecure, and a part of Challenges faced and Conclusion of the paper.

Janaki Devi Palanki:

1. Led the investigation and analysis of the UNet model, including a detailed study of its architecture, parameters, and applications. Conducted experiments to assess the performance of the UNet model on the dataset.
2. Implemented the code for UNET and observed the various outcomes and challenges while writing code for the integration required for the UNet architecture.
3. Wrote the Abstract, UNET and DeepLabv3+ part of Introduction, Experimental Analysis of UNet Architecture.

Lakshmi Grishma Vennapusa:

1. Investigated the Efficient Segment Anything Model (SAM), conducting a comprehensive study on its architecture, design parameters, and integration with existing segmentation models.
2. Implemented SAM within the above architectures and conducted experiments to evaluate its performance and accuracy and debugged the issues observed in this process and wrote required code for this integration of data with the model.
3. Wrote the Efficient-SAM part of Introduction, Experimental Analysis of Efficient-SAM model, a part of Challenges faced and Conclusion

Combined all the individual observations and together we analysed and concluded the results and observations.

V. Conclusion

In conclusion, this paper provides a comparative study of both DeepLabV3+, UNET architectures and incorporation of Efficient SAM to study the efficient usage of manual prompts.

Through extensive experimentation and analysis across diverse datasets, we observed varying performance trends between DeepLabV3+ and UNET, with neither model consistently outperforming the other across all datasets. From our observations, DeepLabV3+ worked well on the datasets containing the objects which are very distinct in features from the background. Unet works well on medical datasets such as breast and on datasets which has defined edges but it didn't work well on images which were very small and has features aligning with other objects in the image. We also observed that integrating the Efficient SAM model leads to performing a comparative analysis to assess the accuracy of prompt data from various sources using Intersection over Union (IOU). This analysis helps identify which version, Efficient SAM-Tiny or Efficient SAM-Small, performs better for the given dataset, and which model is more suitable based on the dataset's characteristics.

VI. Challenges Faced

1.Computational Complexity was the major challenge faced. Although UNET and DeepLabV3+ achieves impressive accuracy , their computational complexity is considerable. Our systems didn't have the expected capability which in turn raised a huge challenge in training the model with all the datasets and prompting results.

2. We had to convert the data to different formats to train the model accordingly which in return reduced the image quality and have given few disturbances in the images. Hence, the performance of these models can degraded in many scenarios where the training data contains errors.

3.Improving the effectiveness of the models was quite challenging for the diverse datasets containing radically different scenarios or less structured environments, such as rural areas or regions with diverse vegetation.

4. Choosing the best mask for training out of all the generated mask was a challenging task.

VII. Source Code

The code for our project can be found at **GitHub Repository**

References

[1] Heryadi et al. "The Effect of Resnet Model as Feature Extractor Network to Performance of DeepLabV3 Model for Semantic Satellite Image Segmentation". In: *2020 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS)*. 2020, pp. 74–77. DOI: 10.1109/AGERS51788.2020.9452768.

[2] Yunyang Xiong et al. "EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything". In: 2023.

[3] Alexander Kirillov et al. "Segment anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

[4] Han Shu et al. "TinySAM: Pushing the Envelope for Efficient Segment Anything Model". In: 2023.

[5] Jasjit S. Suri et al. "UNet Deep Learning Architecture for Segmentation of Vascular and Non-Vascular Images: A Microscopic Look at UNet Components Buffered With Pruning, Explainable Artificial Intelligence, and Bias". In: vol. 11. 2023, pp. 595–645. DOI: 10.1109/ACCESS.2022.3232561.