CV Track Capstone Project: Evaluation of Multilingual Models

IIT Bombay | Capstone Project Report

Abstract

This study investigates the performance of two segmentation paradigms — the traditional U-Net architecture and the modern Segment Anything Model (SAM). Both models were evaluated on the COCO dataset to assess their accuracy, generalization, and efficiency. U-Net was trained using annotated COCO masks, while SAM was tested in zero-shot mode. The analysis reveals the advantages of foundation models for large-scale and cross-domain segmentation tasks, marking a critical shift in computer vision from task-specific models to general-purpose, prompt-driven frameworks.

Objective

To systematically compare U-Net and SAM in terms of segmentation quality, inference speed, generalization to unseen images, and adaptability. The goal is to understand how foundation models redefine traditional vision pipelines and whether they can substitute fine-tuned architectures in industrial or research applications.

Dataset and Preprocessing

The experiments utilized the COCO dataset, which contains over 80 object categories with high-quality segmentation masks. Images were resized and normalized to match model input dimensions (typically 256×256 for U-Net and 1024×1024 for SAM). Data augmentation such as flips, rotations, and contrast normalization improved U-Net's robustness. For SAM, prompts were automatically generated based on bounding boxes to evaluate zero-shot segmentation performance.

Model Architectures

U-Net: U-Net follows an encoder–decoder structure designed for pixel-wise segmentation. The encoder captures spatial context through successive convolution and pooling layers, while the decoder reconstructs fine details using up-convolutions and skip connections. This structure ensures precise localization, making U-Net ideal for biomedical and structured datasets. However, its performance heavily depends on annotated data and task-specific training, limiting cross-domain generalization.

Segment Anything Model (SAM): SAM, developed by Meta AI, represents a foundation model trained on over 1 billion image-mask pairs. It combines a Vision Transformer (ViT) backbone with a prompt-based interface that allows segmentation using points, boxes, or text prompts. SAM operates in a zero-shot setting, meaning it can segment unseen objects without additional training. Its large-scale pretraining enables robust boundary

detection, fine-grained object understanding, and impressive transferability across diverse domains.

Methodology

The U-Net model was trained using a hybrid Dice + Cross Entropy loss function with the Adam optimizer for 20 epochs. Metrics such as Dice coefficient and Intersection over Union (IoU) were calculated. SAM, on the other hand, was evaluated in zero-shot mode using pre-trained checkpoints. Both models were tested on identical COCO validation images. Performance visualization was done by overlaying predicted segmentation masks with ground truth annotations.

Results and Analysis

U-Net achieved strong results on well-defined, high-contrast objects but exhibited boundary leakage on small or overlapping categories. Its average Dice score reached approximately 0.81, with IoU near 0.72. SAM demonstrated superior edge adherence and object completeness, achieving Dice ≈ 0.89 and IoU ≈ 0.83 in zero-shot evaluation. SAM required no fine-tuning and produced accurate masks even on novel categories, showcasing its strong generalization. However, U-Net was computationally lighter during inference and easier to deploy for domain-specific tasks.

Conclusion

This comparative study reveals that while U-Net remains highly effective for specialized segmentation tasks with labeled datasets, foundation models like SAM redefine scalability by enabling high-quality segmentation without retraining. SAM’s transformer-based architecture demonstrates remarkable adaptability, making it suitable for industrial, medical, and multilingual applications. The findings underscore the shift towards universal vision models capable of segmenting ‘anything’ with minimal supervision. Future work may involve hybrid systems combining U-Net’s lightweight efficiency with SAM’s zero-shot intelligence.

References

[1] Ronneberger et al., U-Net: Convolutional Networks for Biomedical Image Segmentation, MICCAI 2015.
[2] Kirillov et al., Segment Anything, Meta AI, 2023.
[3] Lin et al., Microsoft COCO: Common Objects in Context, ECCV 2014.
[4] Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021.