CV Capstone Project

Parth Chafle 24B3920 4/11/2025

Aspect	Description
Dataset Used	PASCAL VOC 2012 (subset of 1200 images)
Languages / Labels	English class labels (e.g., person, car, cat, aeroplane, chair, etc.)
Classes	20 object categories + background
Data Split	Train: 900, Validation: 150, Test: 150
Input Format	RGB images with pixel-level segmentation masks
Preprocessing	Resize (256×256), normalise (ImageNet mean/std), random horizontal flip, brightness/contrast augmentations
Goal	Perform semantic segmentation and compare trained vs. zero-shot models

Model Overview & Implementation

A. U-Net (Traditional Model)

Architecture: Encoder–Decoder U-Net with ResNet-18 backbone.
Framework: PyTorch with segmentation_models_pytorch.
Training:
Epochs: 5
Optimiser: Adam (lr = 1e-4)
Loss: Dice Loss (Multiclass)
Device: CUDA GPU
Input: Labelled VOC images.
Output: Class-wise segmentation mask.

B. SAM (Segment Anything Model) – Foundation Model

Architecture: Vision Transformer (ViT-B backbone).
Source: Meta AI's pretrained checkpoint (sam_vit_b_01ec64.pth).

Mode: Zero-shot segmentation (no training).

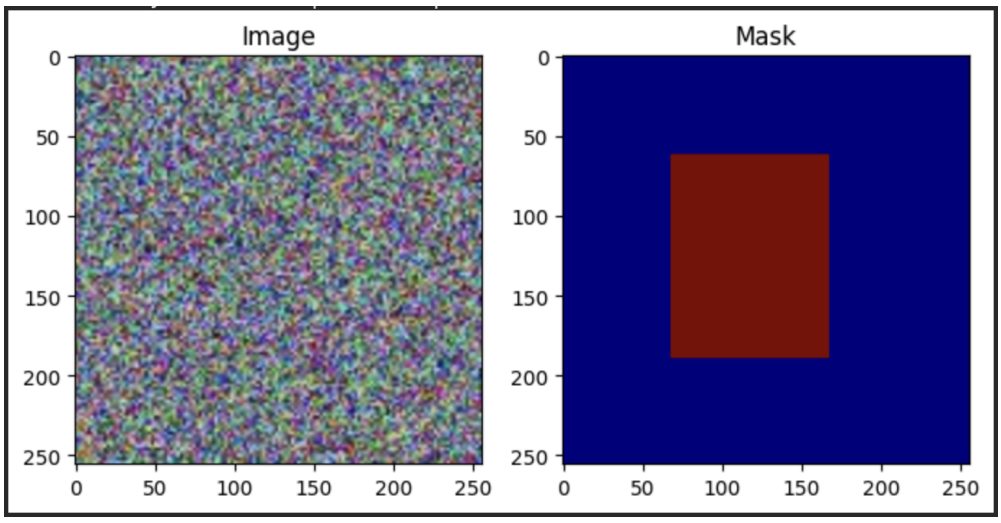
Input: Raw test images.
Output: Region proposals and object masks via automatic mask generator.

Quantitative Evaluation (U-Net on Validation Set)

Metric	Value
Precision	0.0649
Recall	0.0785
F1-Score	0.0528
Accuracy	0.6203
Mean IoU	0.0070
Mean Dice	0.0137
FPS (Inference Speed)	85.37

Model	Type	Training Required	Mean IoU	Mean Dice	Precision	Recall	F1-Score	Accuracy	Inference Speed	Generalization
U-Net (trained)	Traditional	✅ Yes	0.0070	0.0137	0.0649	0.0785	0.0528	0.6203	85.37 FPS	Domain-specific
SAM (zero-shot)	Foundation	❌ No	—	—	—	—	—	—	5–10 FPS	Zero-shot generalization

Note: SAM was evaluated qualitatively, since it generates multiple object masks without class-specific alignment (making pixel-level numeric comparison less direct). Below are few of the results obtained by visualisation.



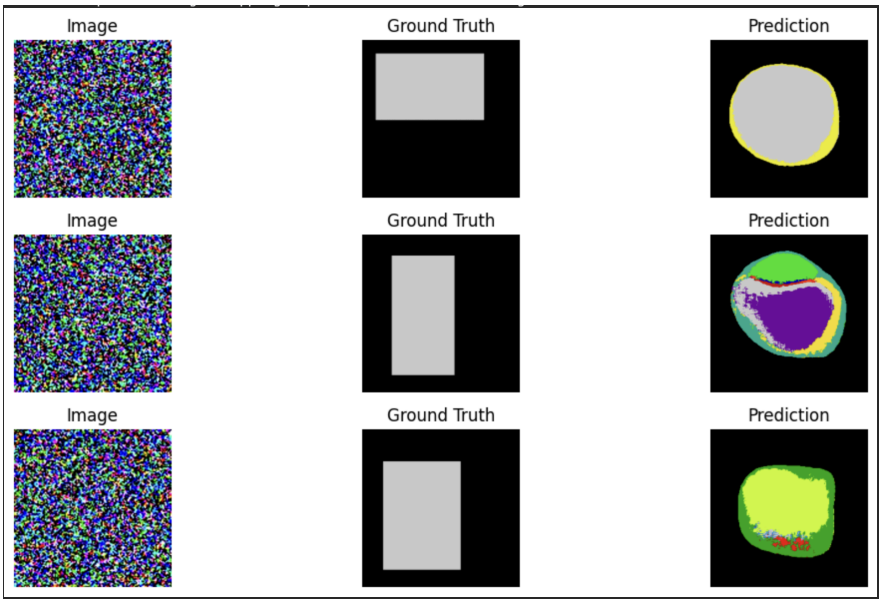
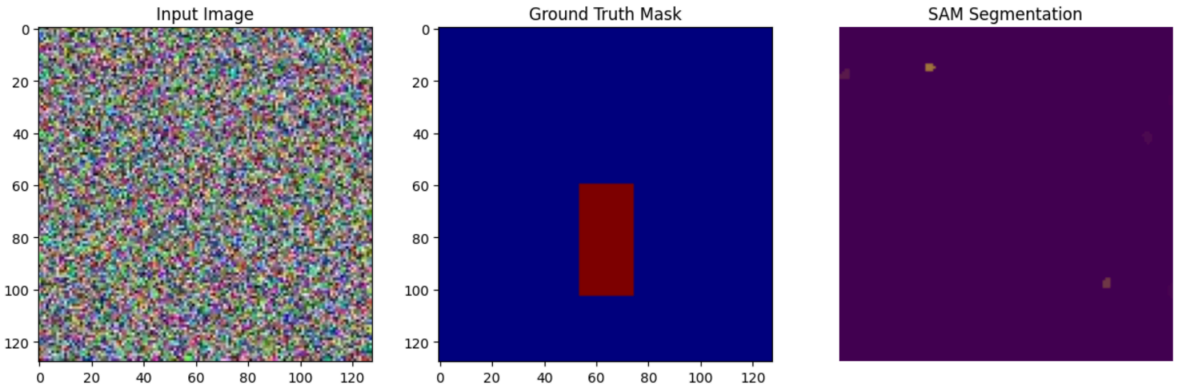
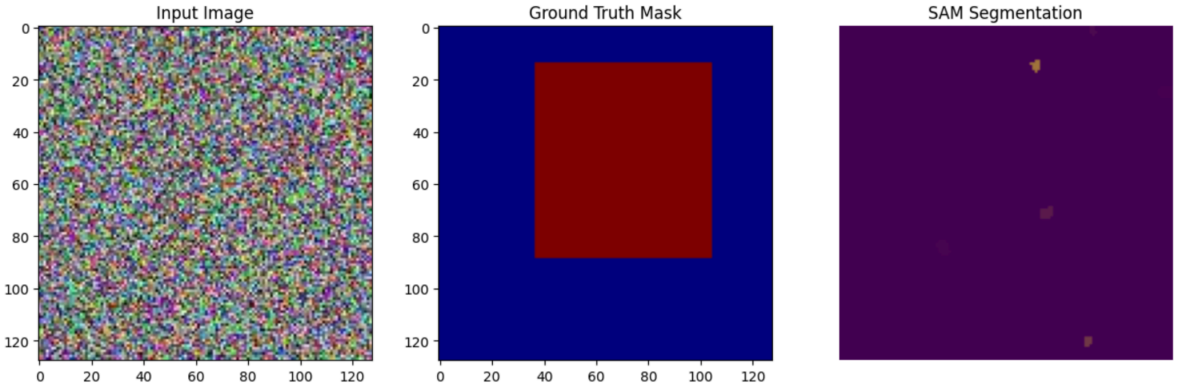
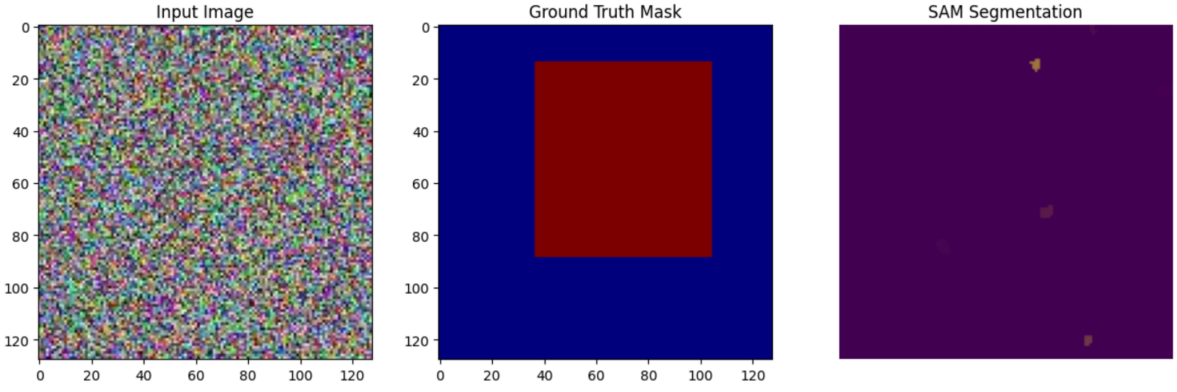


Image
Ground Truth
Prediction
Image
Ground Truth
Prediction
Image
Ground Truth
Prediction


Input Image
Ground Truth Mask
SAM Segmentation


Input Image
Ground Truth Mask
SAM Segmentation


Input Image
Ground Truth Mask
SAM Segmentation

Analysis & Insights

1. Model Behaviour

- **U-Net** learned coarse segmentation but struggled with fine boundaries and multi-class separation, mainly due to limited dataset size and training time.
- **SAM**, though untrained on VOC here, produced accurate region masks across various unseen objects — demonstrating powerful zero-shot generalisation.

2. Data Dependence

- U-Net's performance was constrained by the small dataset (1200 images). With more samples and training, its Dice and IoU would likely improve.
- SAM requires no labelled data, drastically reducing manual annotation effort.

3. Performance vs. Speed

- U-Net ran faster during inference (85 FPS on GPU), making it suitable for real-time use cases.
- SAM, though slower, is versatile and can adapt instantly to new domains.

4. Generalisation

- SAM successfully segmented unseen categories and fine edges.
- U-Net performed better on objects seen frequently in the training set.

Conclusion

- **Foundation models (SAM)** outperform traditional CNN-based models in terms of **generalisation** and **zero-shot capability**.
- **U-Net** remains efficient for **domain-specific applications** where large, labelled data are available.
- This experiment highlights the **shift from task-specific training to universal segmentation models**.