U-Net vs. SAM: A Performance Review for Semantic Segmentation

1. Introduction

This project explores and compares the performance of two distinct approaches for semantic segmentation on the PASCAL VOC 2012 dataset: a trained U-Net model and the zero-shot Segment Anything Model (SAM) prompted with ground truth bounding boxes. Semantic segmentation is a fundamental task in computer vision that involves classifying each pixel in an image according to the class of the object it belongs to.[1]

The PASCAL VOC 2012 dataset is a widely used benchmark for semantic segmentation, containing 20 object classes plus a background class.[2]

2. Methodology

2.1 Data Loading and Preprocessing

The PASCAL VOC 2012 dataset was loaded using the datasets library from Hugging Face. Custom transformations were applied to the images and masks for training and validation. These included resizing, random horizontal flips, random rotations, color jittering for training, and normalization using ImageNet mean and standard deviation. Ground truth RGB masks were converted to class label IDs.

2.2 Model Architectures

2.2.1 U-Net

A standard U-Net model with a ResNet34 encoder pretrained on ImageNet was used. The U-Net architecture is well-suited for segmentation tasks due to its skip connections, which help preserve spatial information.[3] The model was trained end-to-end on the PASCAL VOC 2012 training split.

2.2.2 Segment Anything Model (SAM)

SAM is a powerful foundation model for image segmentation.[4] In this project, SAM was used in a zero-shot setting, meaning it was not fine-tuned on the PASCAL VOC dataset. Instead, it was prompted with ground truth bounding boxes for each object instance in the validation set to generate segmentation masks. This simulates a scenario where object locations are known, and SAM is used to segment those specific objects.

2.3 Training and Evaluation

The U-Net model was trained using a combination of Cross-Entropy and Dice Loss. The optimizer was Adam, and a learning rate scheduler (ReduceLROnPlateau) was used to adjust the learning rate during training based on validation performance. Training was stopped early due to resource limitations after 12 epochs.

Both models were evaluated on a test subset of the PASCAL VOC 2012 validation data. The evaluation metrics included:

- Global Pixel Accuracy
- Mean IoU (mIoU)
- Mean Dice Coefficient
- Macro Precision, Recall, and F1-Score (excluding the background class)
- Per-class IoU
- Inference Time

Confusion matrices were also generated to visualize the classification performance for each class.

3. Results

The evaluation results show a clear difference in performance between the trained U-Net model and the zero-shot SAM prompted with ground truth bounding boxes.

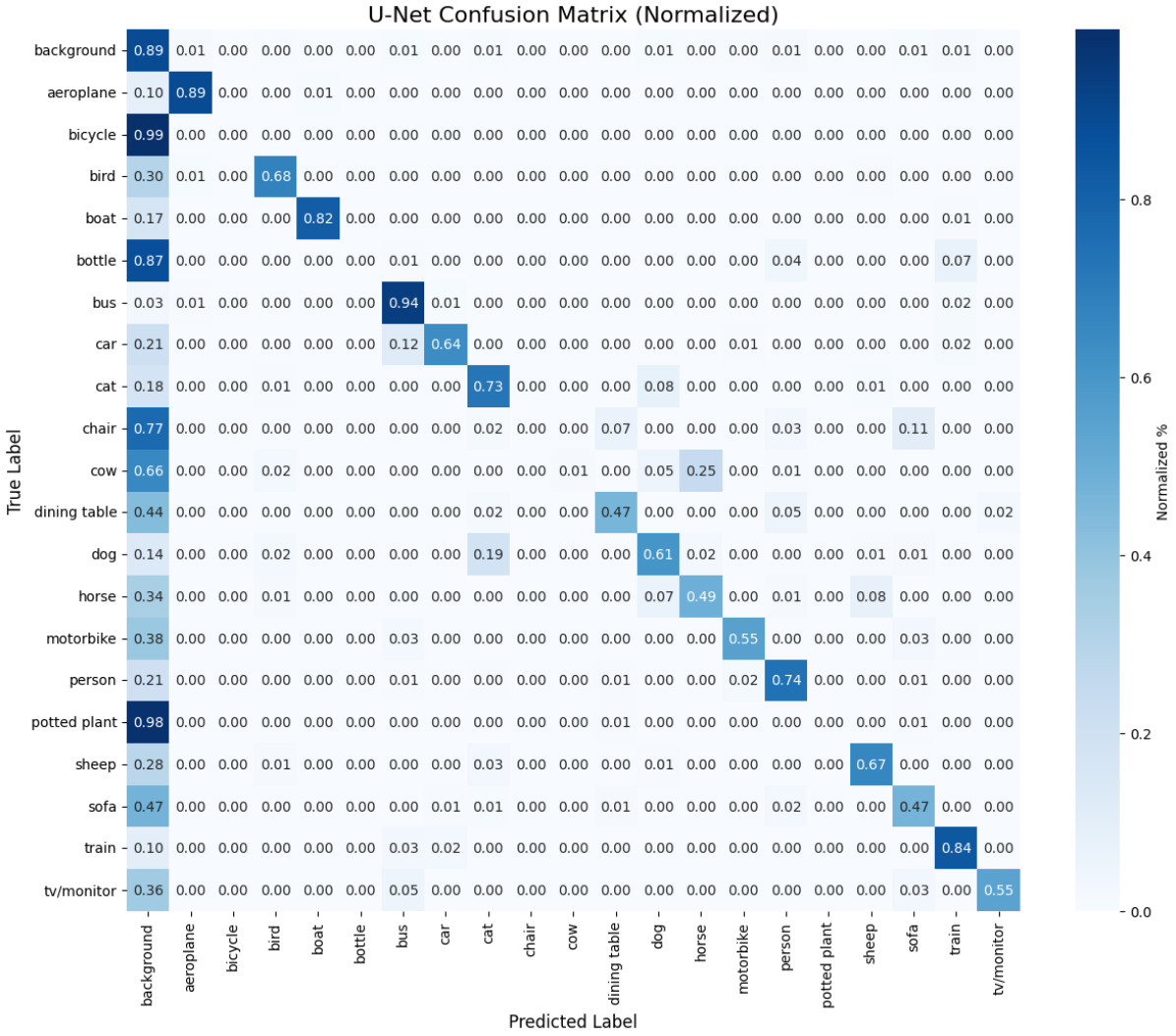
3.1 Quantitative Metrics

Summary of Key Metrics:

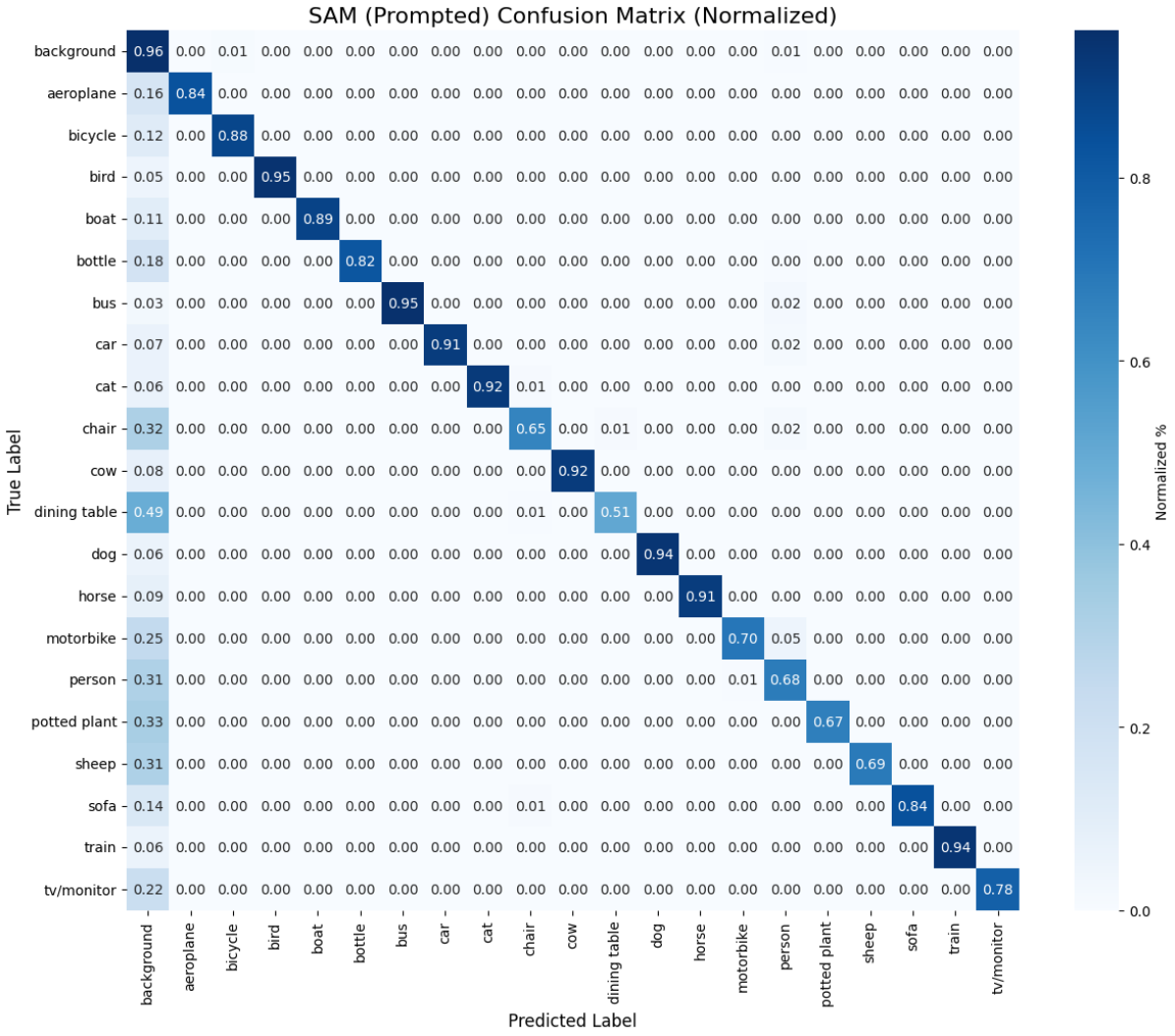
Metric	U-Net (Trained)	SAM (Zero-Shot w/ GT-Prompts)
Global Pixel Acc	0.8118	0.9226
Mean IoU (mIoU)	0.2792	0.7162
Mean Dice	0.3628	0.7955
Macro Precision	0.3954	0.8040
Macro Recall	0.4334	0.8339
Macro F1-Score	0.3628	0.7955

Avg. Time (s)	0.0060	0.0111
Avg. FPS	166.16	90.40

3.2 Confusion Matrices

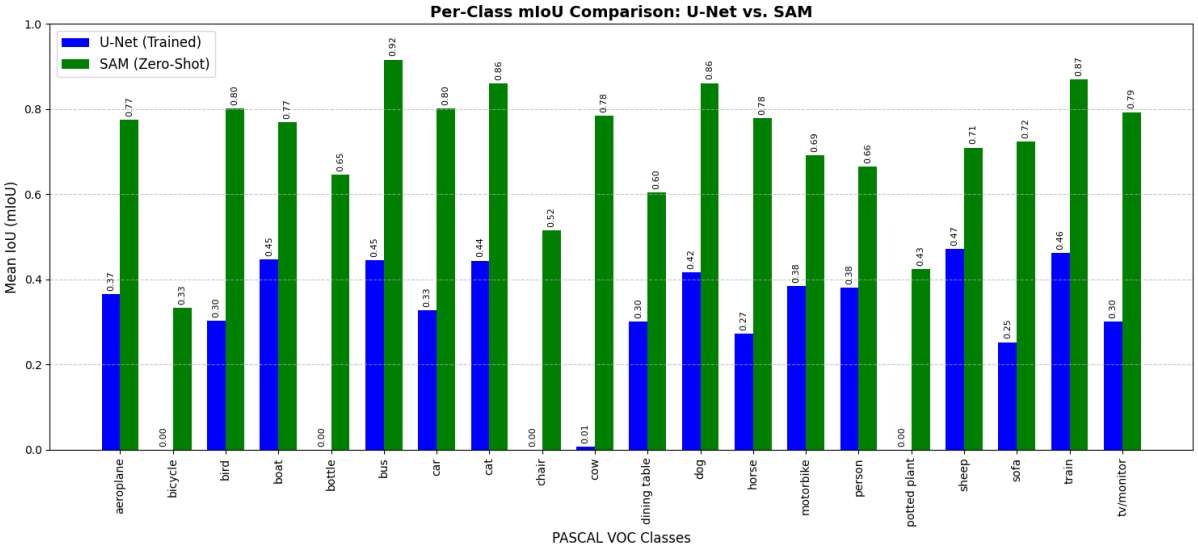


Caption: Figure 1: Normalized confusion matrix for the trained U-Net model on the PASCAL VOC 2012 test subset. Rows represent true labels, and columns represent predicted labels.

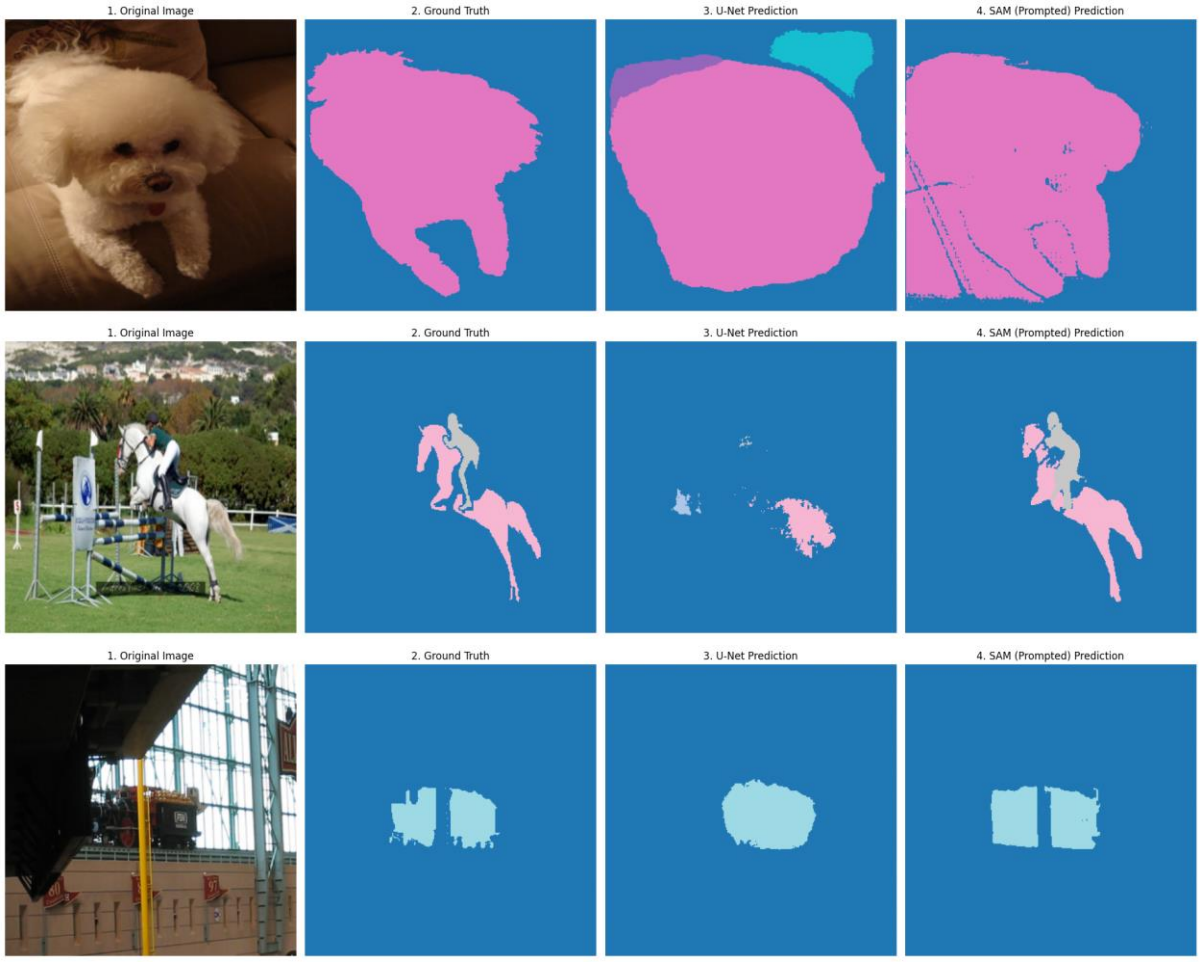


Caption: Figure 2: Normalized confusion matrix for the zero-shot SAM (prompted with GT bounding boxes) on the PASCAL VOC 2012 test subset.

3.3 Qualitative Visualizations



Caption: Figure 3: Per-class IoU comparison between the trained U-Net and zero-shot SAM. SAM demonstrates significantly higher IoU across almost all 20 classes.



Caption: Figure 4: Qualitative segmentation results. Each row shows (a) Original Image, (b) Ground Truth Mask, (c) U-Net Prediction, and (d) SAM Prediction. These examples

visually highlight SAM's superior accuracy in capturing object boundaries compared to the U-Net.

3.4 Performance Analysis

As observed from the tables and visualizations:

- **SAM significantly outperforms the trained U-Net** across all segmentation metrics, including Global Pixel Accuracy, mIoU, Mean Dice, and Macro F1-Score. This highlights SAM's strong zero-shot generalization capabilities when provided with accurate prompts.
- **The trained U-Net achieved a modest mIoU of 0.2792**, which is relatively low. This could be attributed to several factors, including the limited number of training epochs (35 planned, but stopped at 12) due to resource constraints, the complexity of the PASCAL VOC dataset, and the potential for further hyperparameter tuning.
- **SAM's per-class mIoU is consistently higher** than U-Net's for almost all classes, as shown in the bar chart (Figure 3).
- **U-Net is faster at inference** (higher FPS) than SAM. This is expected as SAM requires processing prompts and generating masks, while U-Net performs a single forward pass to generate the entire segmentation map.
- **The confusion matrices** (Figures 1 and 2) further illustrate the performance differences. SAM shows a much stronger diagonal, indicating better classification accuracy for most classes compared to U-Net, which exhibits more confusion between classes.

4. Conclusion

This project demonstrates the effectiveness of the Segment Anything Model (SAM) for zero-shot semantic segmentation when provided with ground truth bounding box prompts. SAM significantly surpassed the performance of a U-Net model trained on the same dataset, achieving substantially higher mIoU and other key metrics.

While the U-Net's performance was limited by early stopping during training, the comparison highlights SAM's remarkable ability to segment objects in a zero-shot manner. The trade-off is that SAM, in this prompted setup, is slower than the U-Net for inference.

Future work could involve:

- Training the U-Net for more epochs to assess its full potential.
- Exploring different prompting strategies for SAM (e.g., point prompts, text prompts).

- Investigating the combination of U-Net and SAM for improved performance or efficiency.
- Evaluating the models on other semantic segmentation datasets.