University of Europe for Applied Sciences

Semantic Segmentation using FCN | U-net | DeepLabV3+

VEHICULAR TECHNOLOGY -

12/7/2025

INTRODUCTION

PROJECT

Semantic segmentation of urban driving scenes using the CamVid dataset with 32 pixel-wise annotated classes.

MODELS EVALUATED

FCN (ResNet-50), U-Net with attention, and DeepLabv3+.

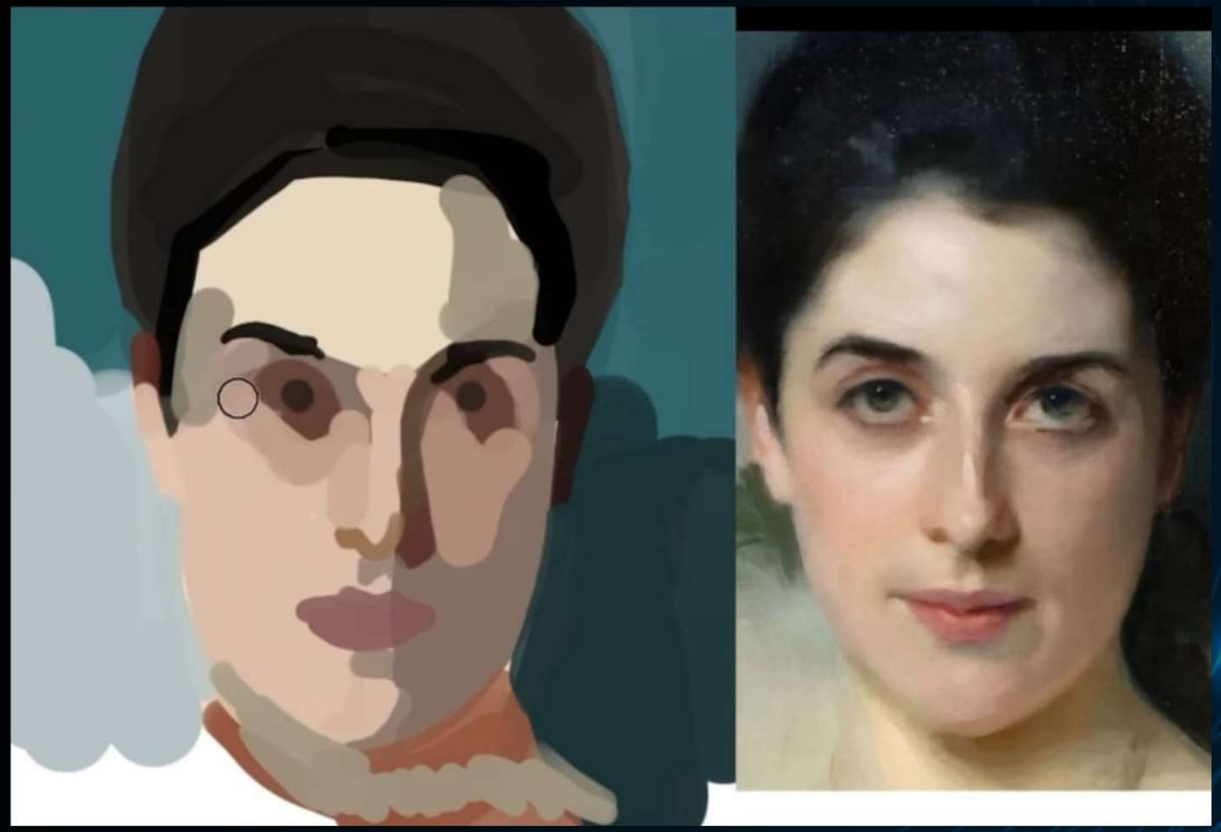
METHODOLOGY

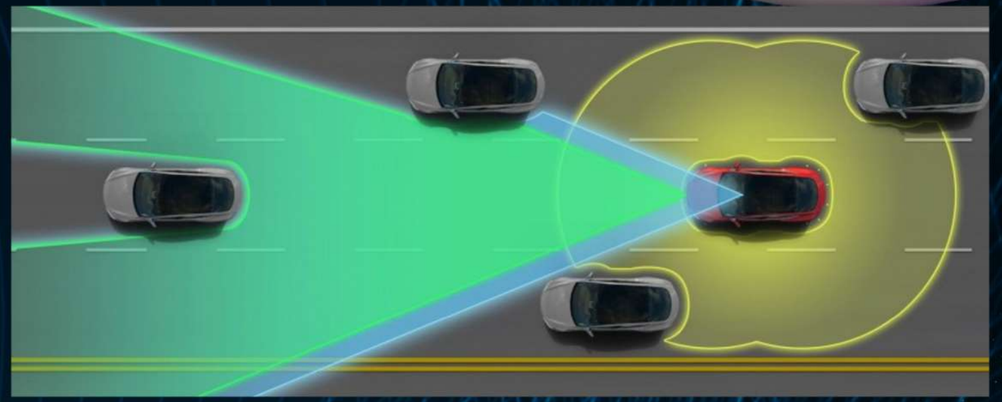
All models trained and evaluated on identical dataset splits with consistent preprocessing and augmentation.

EVALUATION METRICS

Performance compared using mean Intersection over Union (mIoU).

CONCEPT

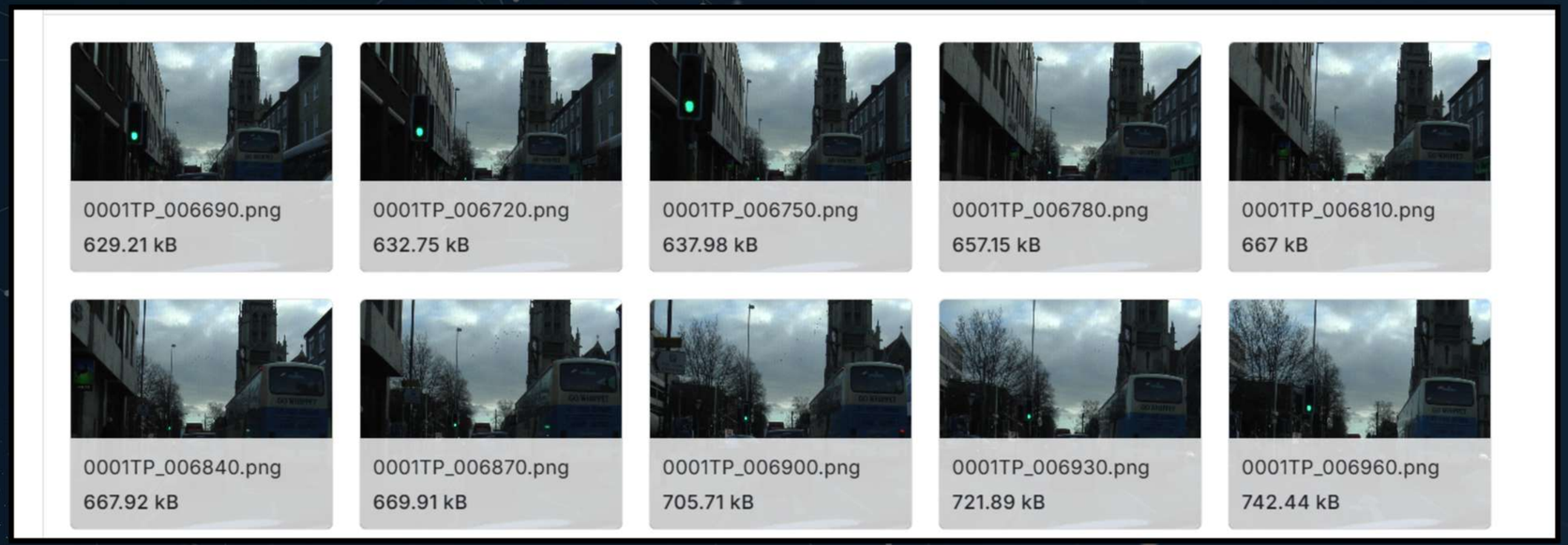




DATASET DESCRIPTION

- CamVid is a widely used semantic segmentation dataset with pixel-level labels for 32 classes in urban driving scenes, captured from videos taken by a moving vehicle.

- The dataset includes training, validation, and test image folders with their corresponding label masks, plus a class_dict.csv file mapping RGB values to class labels for all 32 classes.



Dataset Handling & Preprocessing

Custom Dataset

A CamVidDataset PyTorch class loads images, converts masks to class indices, applies augmentations, and integrates with DataLoader for efficient batch processing.

Data Augmentation

Training samples are augmented (resize, crop, flip, rotation, blur, color jitter, normalization) to boost robustness and prevent overfitting; validation/test sets use only resizing and normalization.

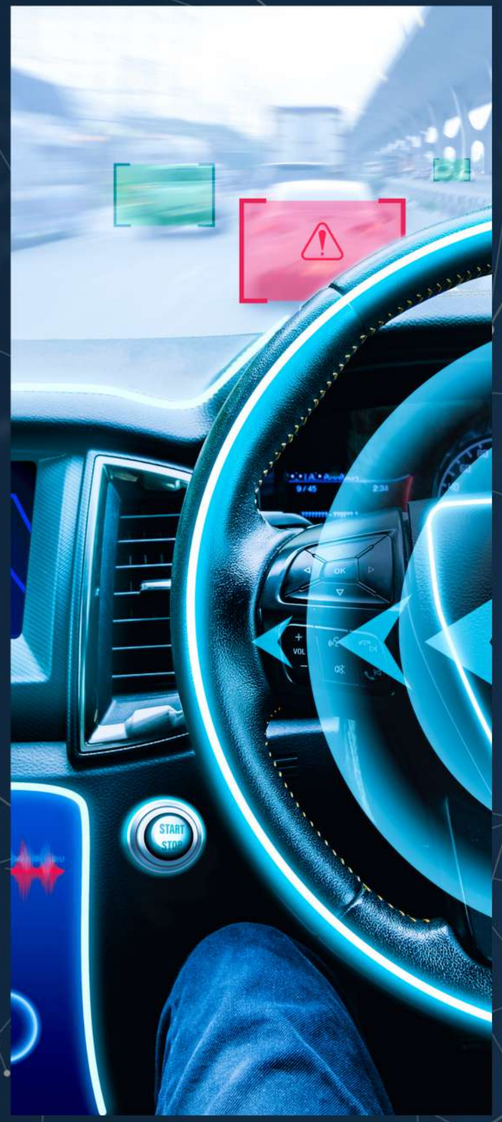
Visualization Function

Converts and displays images and masks side by side, confirming correct dataset loading and label mapping.

Importance

Ensures high-quality data input for training and verifies the accuracy of color-to-class mapping in segmentation tasks.

OVERALL PIPELINE



1. Model Selection

- Choose architecture: FCN, U-Net (ResNet-50), or DeepLabv3+ (all with 32 output classes)
- Set identical training and evaluation conditions

2. Data Preparation

- Apply GroupNorm (32 groups) for stable training with small batch sizes
- Use combined Dice (0.7) + Jaccard (0.3) loss for balanced segmentation

3. Optimization & Training

- Train with Adam optimizer
- Monitor performance with EarlyStopping (patience = 7, delta = 0.001) to prevent overfitting

4. Performance Metrics

- Evaluate and compare models using mean Intersection over Union (mIoU).

5. Reporting

- Summarize and visualize results (mIoU scores, qualitative mask comparisons).
- Highlight key findings and compare model performance.

MODEL OVERVIEW

FCN

- FCN uses a ResNet-50 encoder and a decoder with transposed convolutions for upsampling.
- Maintains spatial resolution for pixel-wise predictions.
- Well-suited for segmenting large road elements in images.

U-NET

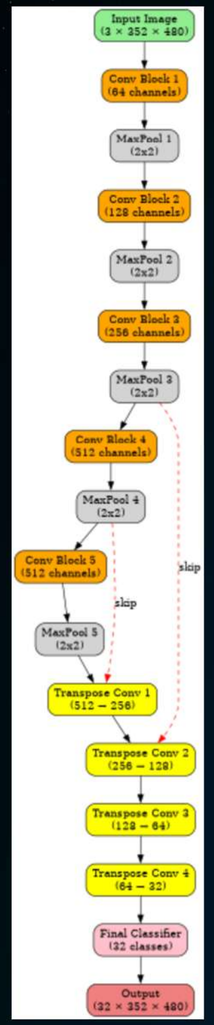
- U-Net adds attention blocks to each skip connection, focusing on important features.
- Uses a ResNet-50 encoder and a decoder with bilinear upsampling and convolutions.
- Excels at segmenting small objects and capturing fine details.

DEEPLABV3+

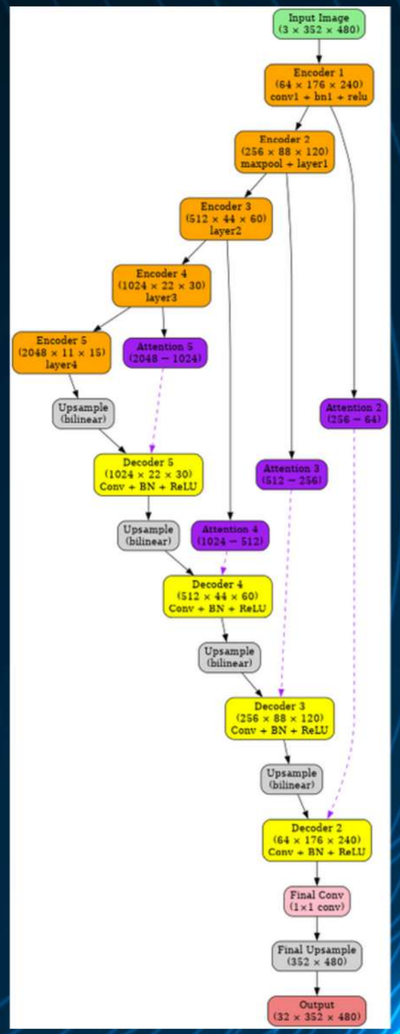
- DeepLabv3+ uses a ResNet-50 backbone and atrous convolutions to capture multi-scale context while preserving spatial resolution.
- The final layer is adapted to match the number of segmentation classes.
- Effective for capturing complex scene details in urban environments.

MODELS ARCHITECTURES

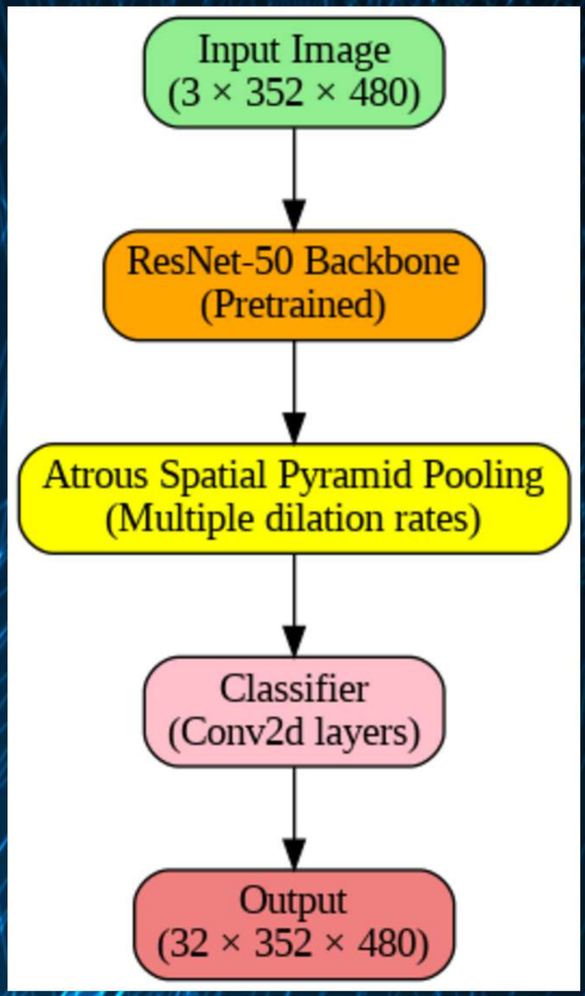
FCN



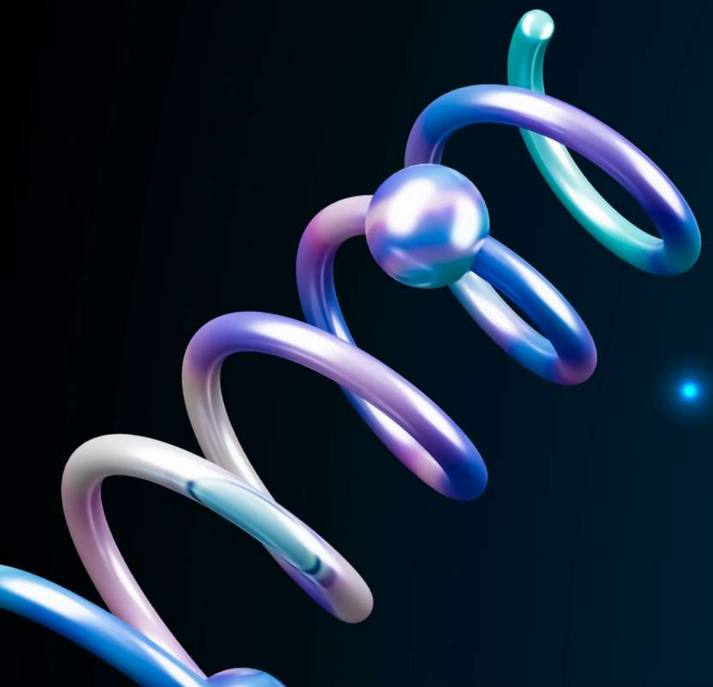
UNET



DEEPLABV3+



Results & Model Comparison



FCN (Fully Convolutional Network):

- Smooth convergence, no overfitting; validation accuracy ~0.88.
- Performs well on large classes (road, sky, buildings) but struggles with small objects and fine boundaries.
- Produces smoother masks, leading to slightly lower mIoU and pixel accuracy.

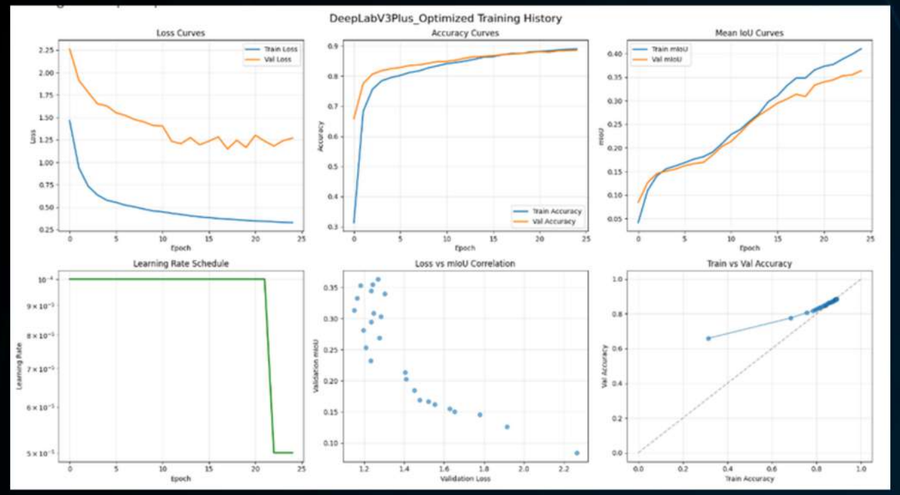
U-Net:

- Best overall performance; validation accuracy above 0.91, mIoU = 0.4572, pixel accuracy = 0.8817.
- Excels at both large and small object segmentation, with sharp boundaries and minimal class bleeding.
- Benefits from skip connections for detailed, precise masks.

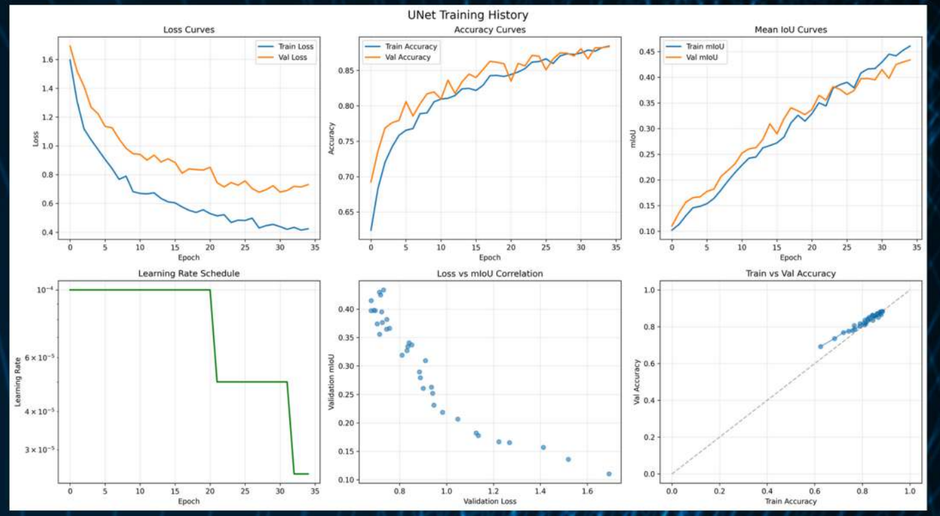
DeepLabV3+:

- Strong training and validation performance (mIoU = 0.3813, pixel accuracy = 0.8637).
- Great at capturing overall scene layout and large classes; improved context awareness.
- Slightly less precise than U-Net on small objects and fine structures.

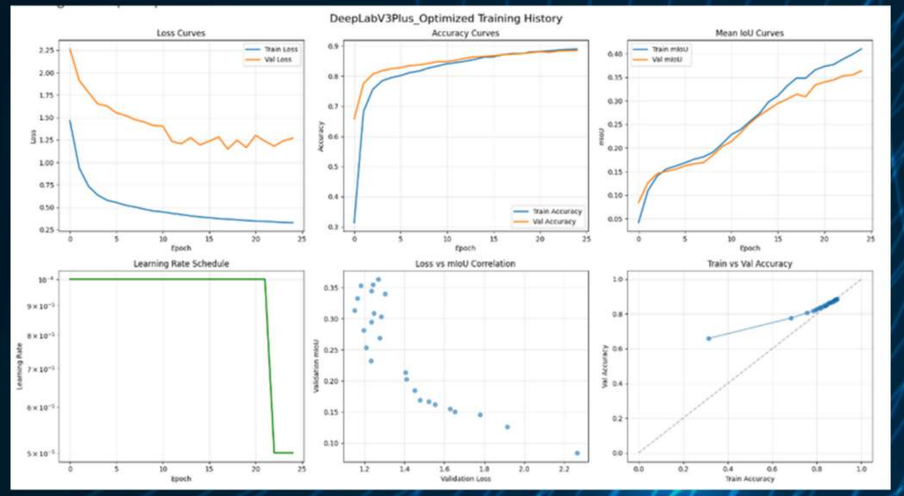
FCN (FULLY CONVOLUTIONAL NETWORK)



U-NET



DEEPLABV3+

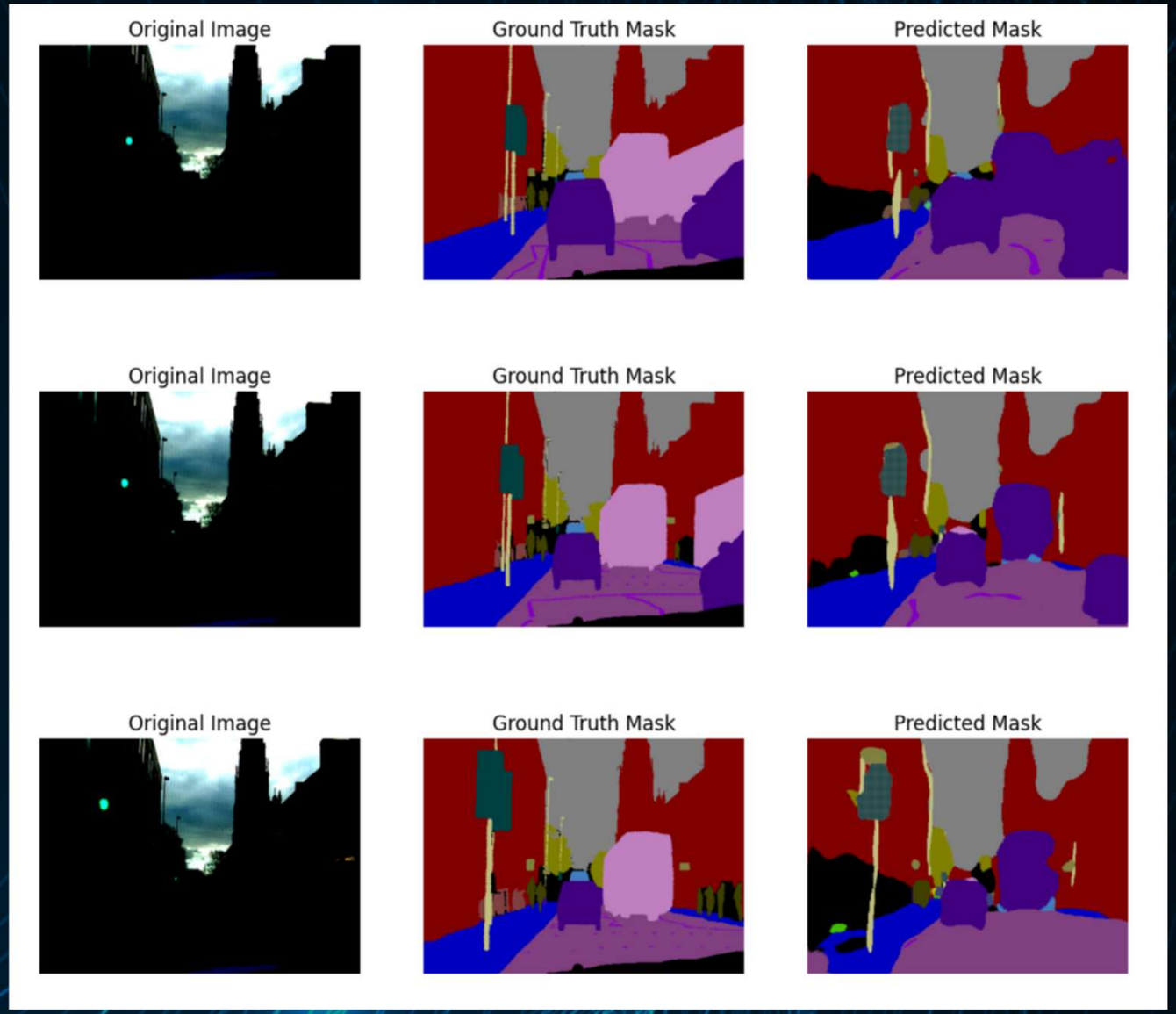


Metric / Observation	FCN	U-Net	DeepLabv3+
Training Loss Trend	Steady decrease, minor overfitting	Steady decrease	Smooth decrease, slight gap
Validation Loss Trend	Slightly higher than training	Higher than training	Slightly diverges late training
Training Accuracy	≈ 80%	≈ 88.4%	≈ 89%
Validation Accuracy	≈ 80%	≈ 88.3%	≈ 89%
Average Pixel Accuracy	7,963	≈ 88%	8,906
Average mIoU (macro)	1,779	4,338	3,635
Visual Prediction Quality	Good structure, some class bleed	Sharp boundaries, fine detail preserved	Strong background separation, better large-object accuracy
Overfitting Signs	Mild	Mild (validation loss higher)	Mild
Strengths	Simple, fast convergence	High accuracy, strong spatial detail retention	Good context understanding
Weaknesses	Lower class-level precision	Large model size, slower training	Slight underperformance in mIoU

VISUAL PREDICTION ANALYSIS

FCN

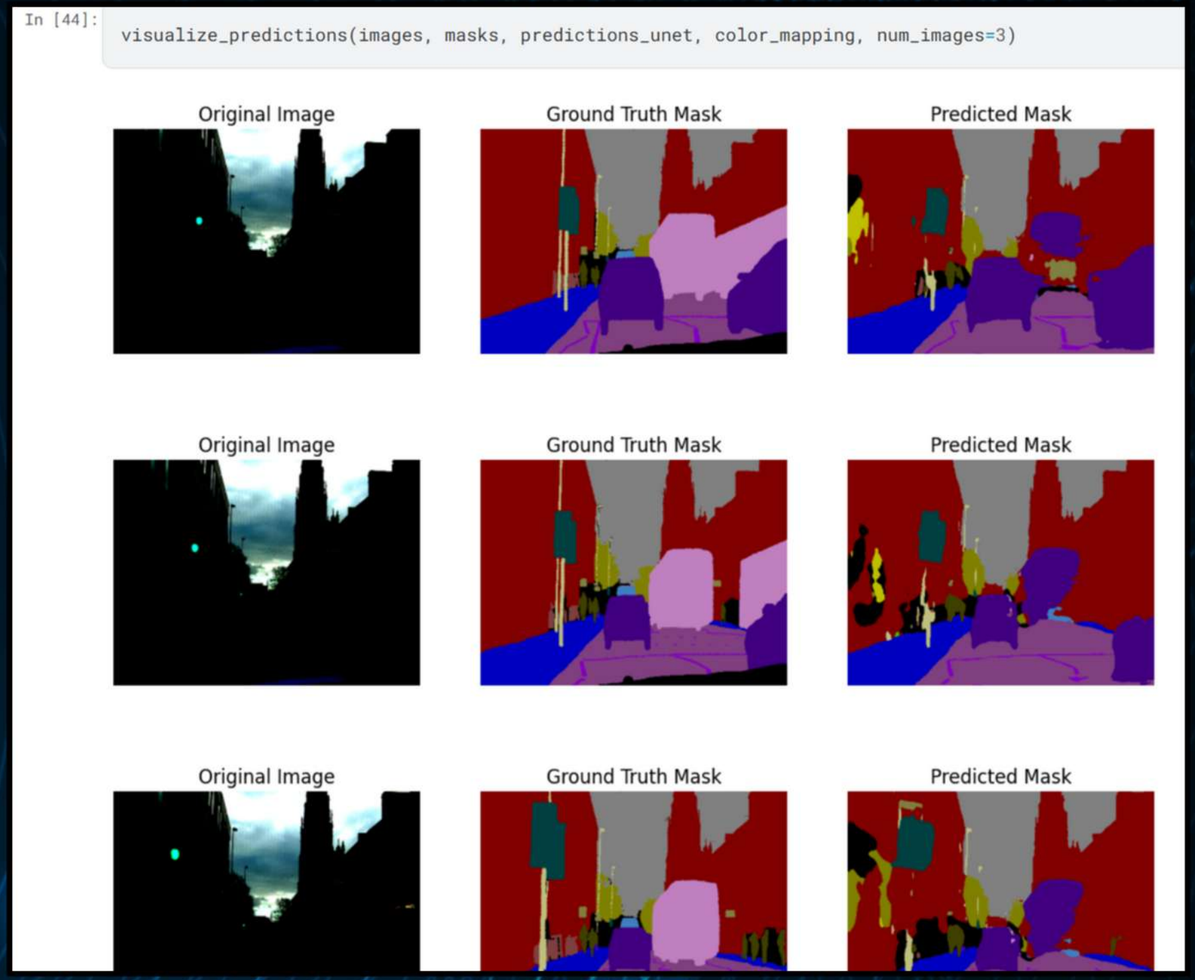
- Captures main scene structure and large classes (road, sky, buildings).
- Smoother predicted masks, but less detail for small objects and fine boundaries.
- Shows some blending between adjacent classes; lower mean IoU and pixel accuracy.



VISUAL PREDICTION ANALYSIS

U-NET

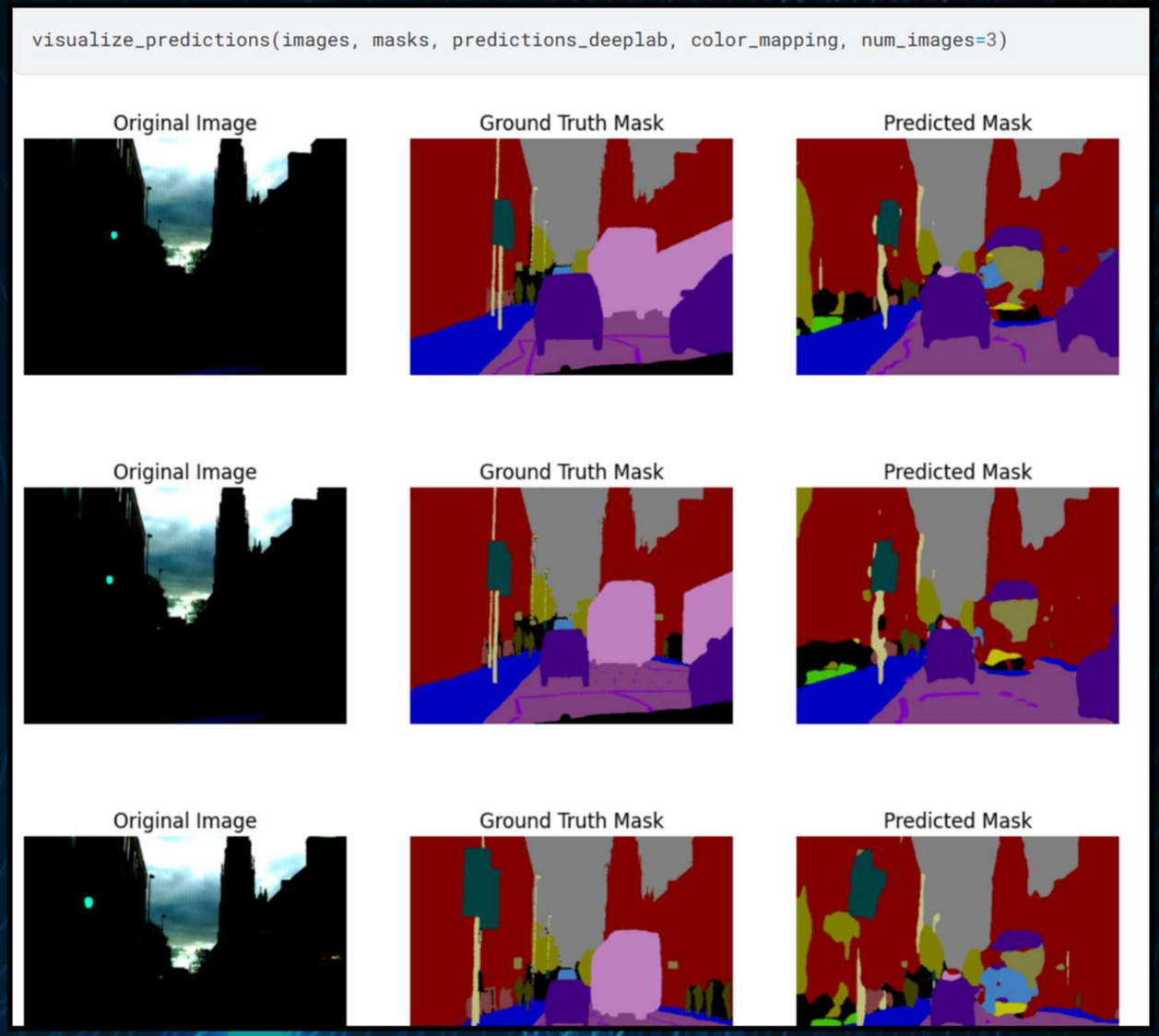
- Most precise segmentation; sharp boundaries and strong alignment with ground truth.
- Excels at both large and small object detection, including thin structures.
- Highest mIoU (0.4572) and pixel accuracy (0.8817).



VISUAL PREDICTION ANALYSIS

DEEPLABV3+

- Good scene layout and consistency for large classes.
- Better contextual awareness than FCN, but less sharp than U-Net on fine details.
- Some fragmentation on small objects; strong overall performance (mIoU 0.3813, pixel accuracy 0.8637)



CHALLENGES FACED

1

CLASS IMBALANCE

- Difficult to segment small/rare objects (e.g., poles, pedestrians, traffic signs).
- Required careful loss weighting with Dice and Jaccard losses.

2

OVERFITTING

- Limited data and frame continuity led to overfitting, especially for deeper models.
- Mitigated with GroupNorm and early stopping.

3

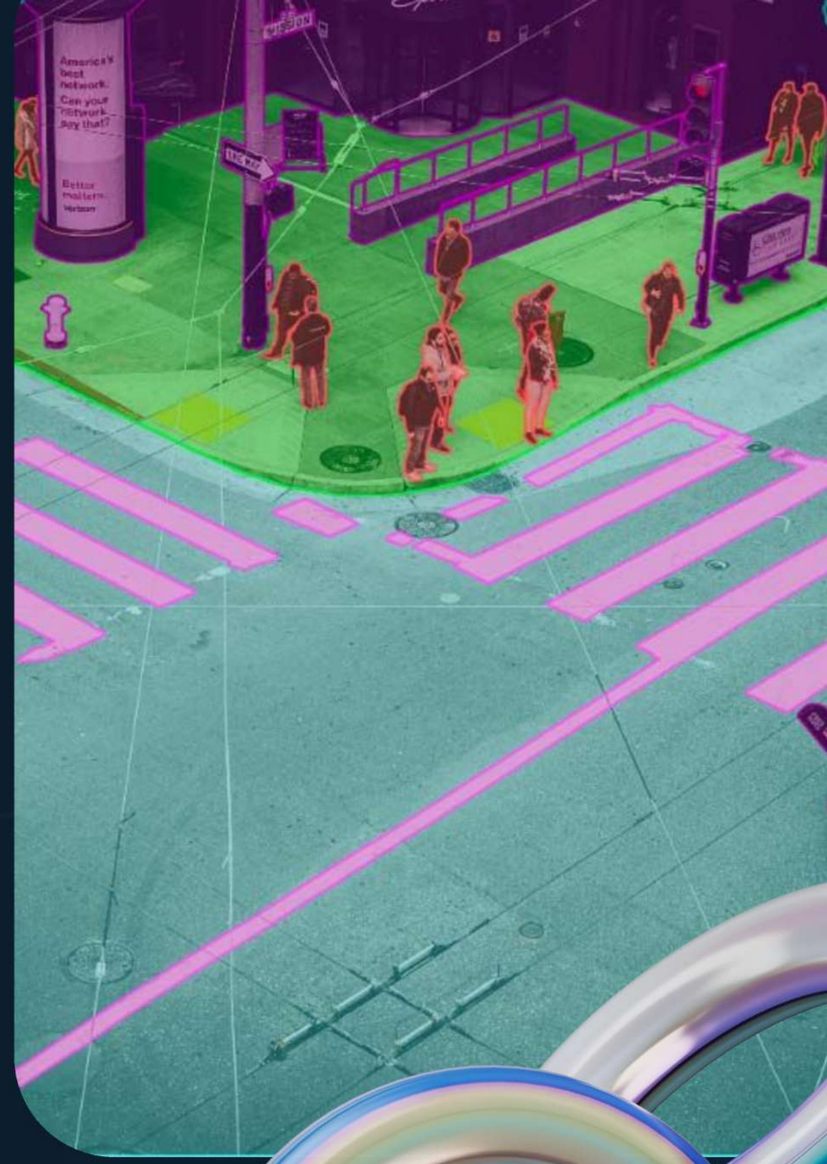
RESOURCE CONSTRAINTS

- Short project timeline limited extensive tuning and experimentation.
- Limited GPU memory and compute time restricted larger experiments and dataset use.

4

OUTCOME

- Despite constraints, all models were implemented, trained, and evaluated successfully.



Conclusion

- **U-Net with Attention:** Top performer with highest mIoU and pixel accuracy, leveraging attention-enhanced encoder-decoder for superior spatial and contextual detail.
- **DeepLabv3+:** Strong on large background regions, weaker on fine object boundaries.
- **FCN:** Simplest model, reliable and fast but less accurate than U-Net and DeepLabv3+.
- **Future Work:** Extend to video segmentation and explore transformer models for better small-object detection.