Road Segmentation

Berat Kargin

April 2025

1 Introduction

With the rapid advancement of image processing and computer vision technologies, revolutionary developments have occurred in many industries, including transportation and traffic safety. One such area is road segmentation, a fundamental task that enables vehicles—especially autonomous systems—to understand their surroundings and make accurate decisions. Road segmentation is an image segmentation problem that aims to distinguish the road surface from the background in an image and classify it at the pixel level.

In autonomous driving systems, environmental perception must be carried out with high accuracy and low latency for vehicles to make real-time decisions. Therefore, accurately recognizing the road surface and performing segmentation without being affected by environmental factors (such as lighting changes, shadows, vehicles, pedestrians, etc.) is of critical importance. In this project, various methods will be used to address this problem, ranging from classical image processing techniques to modern deep learning architectures, and the performances of these methods will be compared.

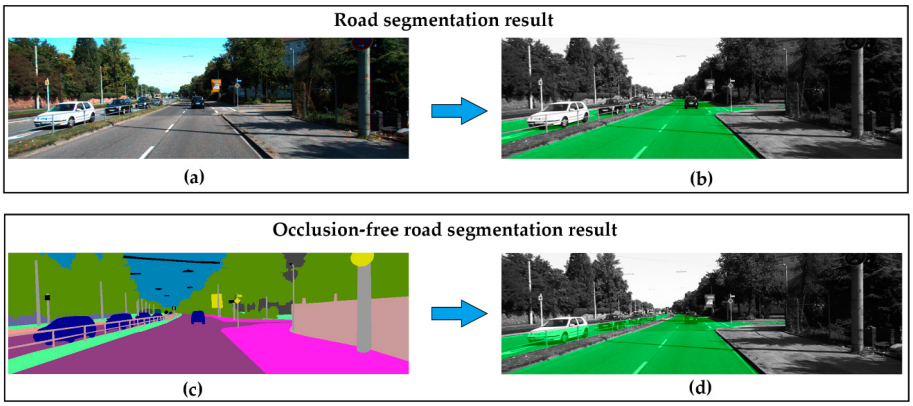

Figure 1: Road Segmentation Example
https://www.mdpi.com/1424-8220/19/21/4711

2 Problem Definition

The road segmentation problem is responsible for automatically detecting and separating road areas in images captured from the environment. However, this task is not as simple as it seems, as images obtained in natural conditions are often noisy and contain many variables. For example, factors such as weather conditions (rain, fog, snow), lighting changes (sunlight, shadows), road surface types (asphalt, dirt, gravel), and objects on the road (vehicles, pedestrians, traffic signs) directly affect the accuracy of segmentation.

In this context, the problem can be defined as follows:

• Each pixel in the image is to be classified as either 'road' or 'non-road.'

• A lightweight and fast model suitable for use in real-time systems is desired.

• High accuracy, generalization capability, and robustness are key criteria for successful segmentation.

Within the scope of this project, solutions will be developed using methods that incorporate both traditional image processing techniques and deep learning-based architectures such as CNNs and Vision Transformers. These solutions will be evaluated in terms of their performance.

3 Literature Survey

Road segmentation has been a topic of research in the literature for many years. Studies in the literature can be broadly divided into two main categories: traditional image processing methods and deep learning-based methods.

3.1 Traditional Methods

In early studies, road segmentation was performed using classical image processing techniques such as color thresholding, edge detection, and shape-based analysis.

• Edge Detection and Hough Transform (Canny + Hough Transform): Road lines are detected to estimate the road area. However, these methods fail when the line clarity is low.

• Color Thresholding (HSV, RGB Thresholding): The road surface typically falls within certain color ranges. This information is used for segmentation. However, different road types and lighting variations reduce accuracy.

• Histogram Analysis and Morphological Operations: The road boundaries can be detected through intensity analysis at the bottom of the image. However, these methods are generally insufficient under diverse conditions.

3.2 Deep Learning-Based Methods

In recent years, studies using deep neural networks have achieved great success. Especially Fully Convolutional Networks (FCNs) have been revolutionary in this field.

• FCN (Fully Convolutional Network): Introduced in 2015, this architecture was the first CNN model aimed at segmenting pixels rather than performing classification.

• U-Net: With its encoder-decoder structure, it combines features from both low and high resolutions to achieve effective segmentation. It performs particularly well on small datasets.

• SegNet: Similar to U-Net in using an encoder-decoder architecture, but it stores max-pooling indices to allow more accurate segmentation during decoding.

• DeepLabv3+: Uses atrous (dilated) convolutions and skip connections to preserve both detailed and contextual information. It is one of the most accurate models for road segmentation.

• ENet: A lightweight model designed for real-time systems. It offers low latency with fewer parameters, though with somewhat reduced accuracy.

• YOLOv8-Seg: Segmentation outputs are obtained via a ROIAlign-like method via the bounding box.

• SegFormer:It is a lightweight architecture that uses the MiT structure to extract multi-scale features in the encoder and combines them with MLP in the decoder to produce a segmentation map.

3.3 Datasets Used

Some of the common datasets used for segmentation today are:

• Cityscapes: A rich dataset with segmentation labels, captured from vehicle cameras in urban areas.

• CamVid: A smaller-scale labeled dataset containing road scenes.

• KITTI Road Dataset: Specifically prepared for road segmentation, including stereo camera images.

• BDD100K: Another dataset with a wide variety of environmental conditions in road scenes.

4 Methods to be Used and Compared

4.1 YOLOv8-Seg

YOLO (You Only Look Once) is a high-speed object detection and recognition algorithm that identifies objects in an image in a single forward pass. YOLOv8 is the latest version developed by Ultralytics and includes support for instance segmentation. With YOLOv8-Seg, masks (segmentation contours) for each object can be extracted, allowing pixel-level segmentation of areas like roads.

4.1.1 YOLO Architecture

Segmentation outputs are derived using ROIAlign-like methods based on bounding boxes.

• Backbone: CSPDarknet-like structures (lightweight and fast).

• Neck: Feature Pyramid Network (FPN) + PAN.
• Head: Three branches for classification, bounding box regression, and mask generation.

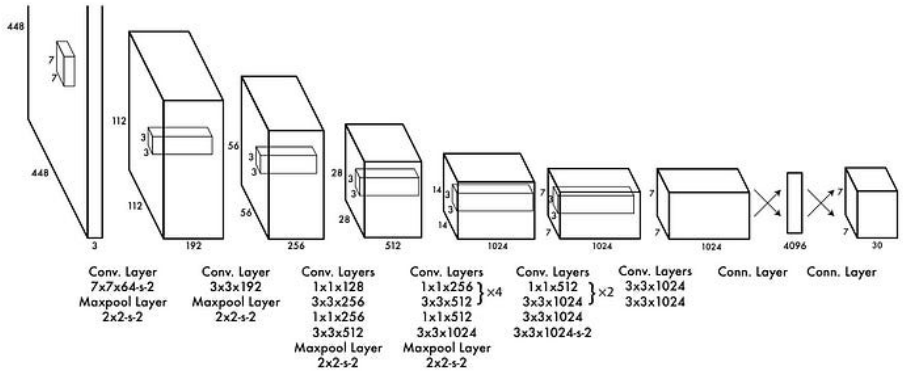

Figure 2: Road Segmentation Example
https://medium.com/@divyapoojitha999/yolo-architecture-6a584081363b

4.1.2 Advantages

• Can operate in real-time (including on Raspberry Pi, Jetson, and mobile devices).
• Combines all tasks in a single network (detection + segmentation).
• Easy and fast training (with Ultralytics API).

4.1.3 Disadvantages

• Segmentation quality is not as detailed as semantic segmentation architectures.
• Performance may drop in complex scenes (with many objects or highly irregular surfaces).

4.2 SegFormer

SegFormer is a Vision Transformer (ViT) architecture optimized for semantic segmentation, introduced by Nvidia in 2021. Unlike CNNs that rely on local filters, it is based on a global self-attention mechanism. It offers both lightweight and high-accuracy versions (models range from B0 to B5).

4.2.1 SegFormer Architecture

• Encoder (MiT – Mix Transformer): Analyzes the image using four levels of windows to generate multi-scale feature maps.

• Decoder: Unlike CNN-based decoders, it directly combines transformer outputs to generate the segmentation map.
• Uses alternatives to BatchNorm and ReLU that require fewer parameters.

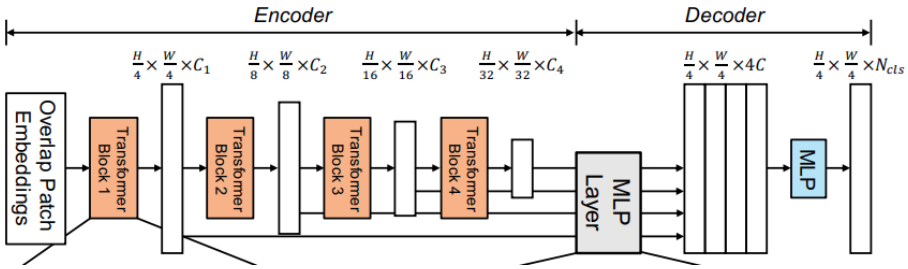

Figure 3: Road Segmentation Example
https://medium.com/data-science/implementing-segformer-in-pytorch-8f4705e2ed0e

4.2.2 Advantages

• Excellent contextual understanding: can correctly classify ambiguous areas such as the edges of roads
• Produces more detailed and sharp results at segmentation boundaries.
• Strong generalization across different data scales.

4.2.3 Disadvantages

• Higher computational cost compared to YOLO.
• Weaker performance for real-time applications (especially with B4-B5 models).

4.3 Comparison Criteria

The CNN-based YOLO architecture and the vision transformer-based SegFormer architectures will be compared under four main headings.

- Accuracy
- Model size and number of parameters
- Real time (FPS - Frame per second)
- Hardware compatibility (Raspberry Pi, Jetson, GPU etc.)