A SegFormer-Based Semantic Segmentation Approach for Autonomous Driving Decision Making Using Stereo Vision

Mahmudah Salwa Gianti, Muhammad Ramdani Subiyanto, Nurul Hasanah, Diky Zakaria, Bambang Wahono, Mulia Pratama, Muhammad Rizalul Wahid, Taufik Ibnu Salim

Presented By :
Bob Nolen Barnabas Munthe

International Conference on Smart Mechatronics
"Empowering Smart Mechatronics Innovations: Robotics, AI, and Intelligent Systems for a Connected Intelligent World"
Bandung, 9–11 June 2026



Organized by:
BRIN Badan Riset dan Inovasi Nasional
Telkom University

Outline

01 Research Background & Objectives

02 Methodology - System Pipline

03 Methodology - System Component

04 Methodology - Speed Decision Algorithm

05 Result - Segmentation and Distance Estimation Performance

06 Result - Spatial Reasoning in Practice

07 Conclusion - Future Work

Research Background & Objectives

The Problem

Bounding-box

Traditional bounding-box detection lacks full scene context



LIDAR Sensor

LiDAR sensors provide high accuracy but at prohibitive cost



CNN-based segmentation

CNN-based segmentation models are computationally heavy



transformer-based segmentation

Gap in transformer-based segmentation for real-time driving



Proposed Solution

Semantic Segmentation

The study utilizes pixel-level semantic segmentation instead of bounding boxes



ZED 2i stereo camera

Offers a cost-effective LiDAR alternative by combining RGB and depth sensing in one compact device.



SegFormer-based architecture

The integration of a hierarchical Transformer (MiT) encoder with a lightweight MLP decoder enhances computational efficiency



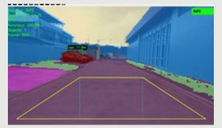
Rule-Based Driving

The research introduces a unified multi-perception and decision-making framework designed specifically for autonomous driving



Methodology

System Pipeline

Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7
				$\text{lanc_center} = \frac{\text{left_base} + \text{right_base}}{2}$ $\text{Curvature} = \frac{\left(1 + (2 \cdot fit_0 \cdot y_{eval} + fit_1)^2\right)^{1.5}}{\lvert 2 \cdot fit_0 \rvert}$		
Input Extracting image frames from the Cityscapes dataset or the ZED 2i camera to serve as raw visual input	**Segmentation** SegFormer classifies pixels into semantic categories such as road, vehicle, and pedestrian	**BEV Transformation** Projecting segmentation output into Bird's-Eye View to simplify road geometry.	**Object Detection** Detecting objects and measuring actual distance using stereo depth information	**Control** Computing steering parameters by deriving road curvature and lane center from BEV data	**Decision Algorithm** Rule-based speed and steering decisions from object distance and lane geometry.	**Result** Final interface integrating visual perception and autonomous navigation decisions

SegFormer-B5 for Cityscapes benchmarking, SegFormer-B0 for real-time stereo inference

Methodology

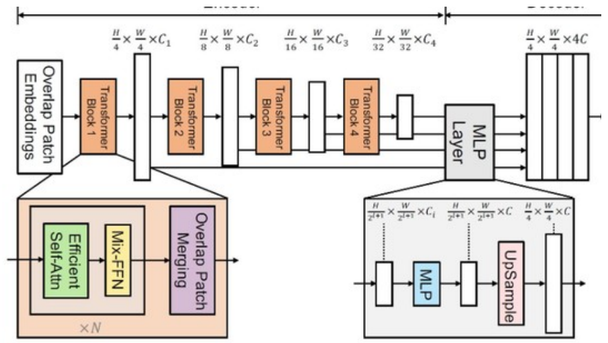
System Components

Dataset



- **Cityscapes Benchmark:** A total of **2,299** high-resolution images of Stuttgart were concatenated via ffmpeg into **30-FPS** video sequences to simulate continuous driving scenarios.
- **Experimental Camera Data:** Direct distance-testing was conducted against physical targets (humans and vehicles) within a spatial range of **1.0–10.0 meters**.

SegFormer



- **SegFormer-B5** : Employed as a theoretical benchmark to validate segmentation accuracy on the Cityscapes dataset.
- **SegFormer-B0** : Integrated with the ZED 2i stereo camera to perform distance estimation and real-time decision-making in physical experiments.

ZED 2i Stereo Camera



A stereo vision camera simultaneously produces RGB images and depth maps by leveraging the principle of disparity, thereby enabling accurate object distance estimation.

Key Output Classes: Road, Sidewalk, Person, Car, Truck, Bus, Motorcycle, Bicycle, Traffic Light.

Methodology

Speed Decision Algorithm

Decision Module Input

- **Object Perception:** Person, Car, Obstacle Detection
- **Spatial Data:** Measured Object Distance
- **Lane Geometry:** Lane Curvature & Offset
- **Road Stability:** Left/Right Lane Radius Calculation
- **Environment:** Free space around the vehicle
- **Radius of Curvature :**

$$R = \frac{\left(1 + \left(\frac{d_x}{d_y}\right)^2\right)^{\frac{3}{2}}}{\left|\frac{d^{2x}}{dy^2}\right|}$$

Speed Decision Rule

Distance < 15 m	MAX Speed : 40 km/h
Distance < 10 m	Reduce to ~ 6 km/h
Distance < 5 m	Low Speed to STOP

**Rules integrated based on proximity and spatial clearance.*

OUPUT : Continue, Turn Left, Turn Right, Stop

Result

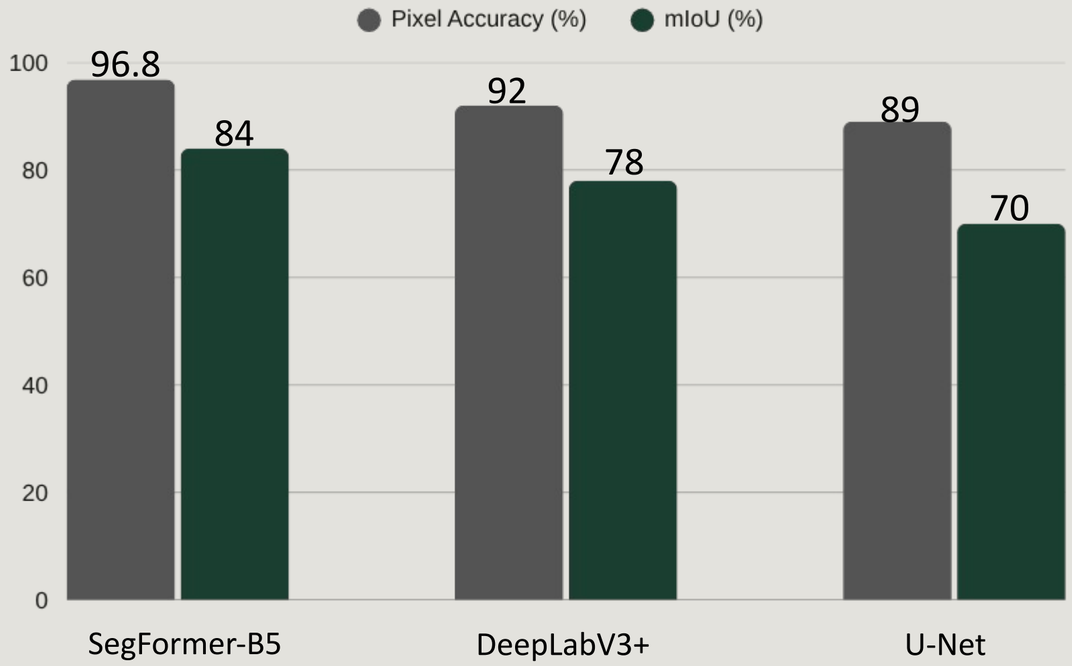
Segmentation and Distance Estimation Performance

96.8%
Pixel Accuracy

~84%
Mean IoU

~15 FPS
Processing Speed

94.7–99.8%
Distance Accuracy



Distance Measurement (Table 2)

Object	Real (m)	Measured (m)	Accuracy
Person	2.0	2.014	99.3%
Car	9.0	8.990	99.8%
Person	10.0	10.060	99.4%

Distance (B0 + ZED 2i)

- Avg. Distance Error: **2–3%** (Indoor Exp.)
- Range: 1–10m (High reliability for urban)
- Utilizes stereo disparity for depth precision

Segmentation (B5)

- Surpasses DeepLabV3+ and U-Net baselines
- Superior performance on Cityscapes dataset
- Throughput: ~15 FPS (vs baselines ~10 FPS)

Result

Spatial Reasoning in Practice

Object Perception	Distance (m)	Object Position (Spatial Context)	Driving Decision	Speed (km/h)
Person	1.37	Center	Stop	0
Person	2.14	Right Side Path Obstruction	Turn Left	5.00
Car	7.05	Center-Right Lane Blocked	Turn Left	6.70
Car	9.74	Center Safe Distance	Continue	6.37
Person	1.86	Center-Left Lane Blocked	Turn Right	5.00

Lateral Analysis

The system uses stereo depth to identify the **lateral position** of objects within the lane, not just linear distance

Obstacle Avoidance

When an obstacle is on the **Right**, the module verifies **Free Space** on the Left to trigger a turn command.

Safety Buffer

Decisions are updated in real-time. If spatial clearance disappears, the system defaults to **STOP** protocols.

Conclusion - Future Work

CONCLUSION

Dense Scene Understanding

Semantic segmentation provides pixel-level context (road, obstacles) that outperforms traditional bounding boxes for lane planning.

Robust Depth Estimation

Stereo vision integration with SegFormer-B0 delivers high-accuracy depth measurement (94.7%–99.8%) without expensive LiDAR

Reliable Low-Speed Control

The rule-based decision module effectively translates spatial cues into safe navigation commands (Stop, Turn, Continue).

FUTURE WORK

Real-World Outdoor Scenarios

Expanding testing beyond indoor corridors to complex urban environments with dynamic lighting and multi-lane traffic.

Learning-Based Strategies

Transitioning from static rule-based logic to adaptive AI-driven decision models for handling unpredictable real-world splits.

Embedded Platform Optimization

Optimizing the pipeline for deployment on high-performance autonomous embedded systems to ensure minimal latency.



Thank You!

BRIN
BADAN RISET
DAN INOVASI NASIONAL

Telkom
University