UNIVERSITAS
GADJAH MADA
OPTIMASI MODEL SEGMENTASI SEMANTIK MENGGUNAKAN KUANTISASI UNTUK INFERENSI REAL-TIME PADA SINGLE BOARD COMPUTER
Ilyaz Raukhillah Adzaqi, Dr. Andi Dharmawan, S.Si., M.Cs.; Oskar Natan, S.ST., M.Tr.T., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

OPTIMASI MODEL SEGMENTASI SEMANTIK *REAL-TIME* MENGGUNAKAN KUANTISASI PADA *SINGLE BOARD COMPUTER*

Oleh

Ilyaz Raukhillah Adzaqi

22/503561/PA/21646

Segmentasi semantik krusial untuk *autonomous driving*, namun beban komputasinya menjadi tantangan besar pada *Single Board Computer* (SBC) berdaya rendah. Penelitian ini mengoptimasi model segmentasi semantik (SegFormer-B0, SegFormer-B1, dan FCN8s-ResNet18) pada Raspberry Pi 5 menggunakan akselerator NPU Hailo-8. Optimasi dilakukan melalui kompresi presisi 8-bit Integer (INT8) menggunakan *Quantization-Aware Training* (QAT) dan *Post-Training Quantization* (PTQ) pada dataset Cityscapes. Hasil pengujian menunjukkan eksekusi CPU murni tidak memadai untuk kebutuhan *real-time* (hanya mencapai 10 FPS). Pendelegasian ke NPU Hailo-8 berhasil mendongkrak kecepatan SegFormer-B0 hingga 27-31 FPS dan SegFormer-B1 hingga 20-23 FPS, sementara FCN8s-ResNet18 meningkat hingga 81-90 FPS. Namun, arsitektur *Mix Transformer* (SegFormer) terbukti lebih tangguh terhadap *quantization noise* dibandingkan arsitektur CNN (FCN8s-ResNet18) yang akurasinya terdegradasi 6,19-24,03 mIoU akibat *downsampling* ekstrem. Varian SegFormer-B1 dengan QAT mampu mempertahankan akurasi spasial secara optimal hingga 48,97 mIoU. Kesimpulannya, integrasi SBC, NPU, dan SegFormer terkuantisasi INT8 berbasis QAT menawarkan solusi persepsi visual *real-time* yang ideal untuk komputasi *edge*.

Kata Kunci: Segmentasi Semantik, SegFormer, *Single Board Computer*, Hailo-8, Kuantisasi, *Edge Computing*.

UNIVERSITAS
GADJAH MADA
OPTIMASI MODEL SEGMENTASI SEMANTIK MENGGUNAKAN KUANTISASI UNTUK INFERENSI REAL-TIME PADA SINGLE BOARD COMPUTER
Ilyaz Raukhillah Adzaqi, Dr. Andi Dharmawan, S.Si., M.Cs.; Oskar Natan, S.ST., M.Tr.T., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

OPTIMIZATION OF REAL-TIME SEMANTIC SEGMENTATION MODELS USING QUANTIZATION ON A SINGLE BOARD COMPUTER

By

Ilyaz Raukhillah Adzaqi

22/503561/PA/21646

Semantic segmentation is crucial for autonomous driving, yet its high computational load presents a significant challenge for real-time implementation on low-power Single Board Computers (SBCs). This research optimizes SegFormer-B0, SegFormer-B1, and FCN8s-ResNet18 models on a Raspberry Pi 5 integrating the Hailo-8 Neural Processing Unit (NPU). Optimization uses 8-bit Integer (INT8) precision compression via Quantization-Aware Training (QAT) and Post-Training Quantization (PTQ) evaluated on the Cityscapes dataset. Experimental results indicate pure CPU execution is insufficient for real-time scenarios (achieving only 10 FPS). Delegating computation to the Hailo-8 NPU significantly boosts inference speeds to 27-31 FPS for SegFormer-B0 and 20-23 FPS for SegFormer-B1, while the FCN8s-ResNet18 peaks at 81-90 FPS. Furthermore, Mix Transformer-based architectures (SegFormer) demonstrate superior resilience against quantization noise compared to Convolutional Neural Networks (FCN8s-ResNet18), which suffer extreme accuracy degradation (dropping to 6.19-24.03 mIoU) due to aggressive downsampling. The SegFormer-B1 variant utilizing QAT successfully maintains optimal spatial accuracy up to 48.97 mIoU. In conclusion, combining an SBC, an NPU accelerator, and INT8-quantized SegFormer architectures via QAT offers an ideal real-time visual perception solution for edge computing.

Keywords: *Semantic Segmentation, SegFormer, Single Board Computer, Hailo-8, Quantization, Edge Computing.*