INTISARI

Fusi Sensor Multi-modal Menggunakan *Transformer* Untuk *End-to-end* Path *Planning* Pada Kendaraan Otonom

Oleh

Novelio Putra Indarto

23/530959/PPA/06753

Sistem kemudi otonom umumnya terdiri atas beberapa modul terpisah seperti persepsi, perencanaan, dan kendali, yang sering menimbulkan kompleksitas integrasi serta hilangnya informasi penting antar-submodul. Pendekatan modular ini juga rentan terhadap propagasi kesalahan dan ketidaksinkronan konteks antar-modul. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan model *end-to-end* berbasis transformer yang mampu melakukan fusi data *multi-modal* (kamera RGB-D, IMU, dan GPS) secara langsung dalam satu alur inferensi guna menghasilkan prediksi *waypoint* yang akurat dan stabil.

Model yang diusulkan dirancang dengan arsitektur *end-to-end* yang terdiri atas blok persepsi berbasis EfficientNet-UNet untuk ekstraksi fitur visual, blok transformer untuk fusi kontekstual antar-modalitas, serta blok prediksi *waypoint* untuk melakukan regresi *waypoint*. Untuk meningkatkan efisiensi komputasi dan ketahanan terhadap noise sensor, diterapkan mekanisme Knowledge Distillation (KD) dari model teacher SegFormer-B5 guna mentransfer pengetahuan segmentasi, serta Entropy Masking (EM) untuk menyaring pseudo-label dengan tingkat ketidakpastian tinggi selama proses pelatihan.

Hasil eksperimen menunjukkan model mencapai rata-rata error *waypoint* sebesar 0,0846±0,0021 meter dan nilai mIoU sebesar 81,1833±0,0981, yang menegaskan keberhasilan model dalam memenuhi tujuan penelitian, khususnya dalam mencapai error *waypoint* di bawah 0,1 meter. Studi ablasi mengonfirmasi bahwa penghapusan modul transformer atau data sensor tertentu menyebabkan penurunan performa yang signifikan, menegaskan pentingnya fusi sensor dalam pembelajaran kontekstual. Pengujian pada perangkat NVIDIA Jetson AGX Orin menunjukkan efisiensi tinggi dengan penggunaan memori rendah dan waktu inferensi cepat, sehingga model layak diterapkan pada sistem kemudi otonom di perangkat dengan sumber daya komputasi terbatas.

Kata Kunci: kendaraan otonom, model *end-to-end*, *transformer*, multi-modal fusion, *knowledge distillation*, *entropy masking*.

ABSTRACT

Multi-Modal Sensor Fusion Using *Transformer* For *End-to-end* Autonomous Vehicles

By

Novelio Putra Indarto

23/530959/PPA/06753

Autonomous driving systems are commonly composed of multiple separate modules such as perception, planning, and control, which often introduce integration complexity and lead to the loss of critical contextual information across submodules. This modular approach is also susceptible to error propagation and contextual misalignment between components. Therefore, this research aims to develop a transformer-based end-to-end model capable of directly fusing multi-modal data (RGB-D cameras, IMU, and GPS) within a single inference pipeline to produce accurate and stable waypoint predictions.

The proposed model is designed with an end-to-end architecture consisting of an EfficientNet-UNet–based perception block for visual feature extraction, a transformer block for contextual multi-modal fusion, and a waypoint prediction block for waypoint regression. To improve computational efficiency and robustness against sensor noise, a Knowledge Distillation (KD) mechanism is applied to transfer segmentation knowledge from a SegFormer-B5 teacher model, while Entropy Masking (EM) is employed to filter pseudo-labels with high uncertainty during training.

Experimental results demonstrate that the proposed model achieves an average waypoint error of 0.0846±0.0021 meters and a mean Intersection over Union (mIoU) of 81.1833±0.0981, confirming that the model successfully meets the research objective of achieving a waypoint error below 0.1 meters. Ablation studies further reveal that removing the transformer module or specific sensor inputs leads to significant performance degradation, underscoring the importance of sensor fusion for contextual learning. Deployment experiments on the NVIDIA Jetson AGX Orin demonstrate high efficiency with low memory usage and fast inference time, indicating that the proposed model is well suited for autonomous steering systems on resource-constrained edge devices.

Keywords: autonomous vehicle, *end-to-end* model, *transformer*, multi-modal fusion, *knowledge distillation*, *entropy masking*.