BAB I
PENDAHULUAN

1.1 Latar Belakang

Kendaraan otonom telah menjadi salah satu inovasi teknologi paling menjanjikan dalam sektor transportasi modern. Kendaraan ini memiliki potensi untuk mengurangi ketergantungan pada pengemudi manusia, sehingga dapat meningkatkan keselamatan jalan, mengurangi kecelakaan yang disebabkan oleh kesalahan manusia, meningkatkan efisiensi bahan bakar, serta mengurangi kemacetan melalui penjadwalan rute yang lebih baik (Chowdhuri et al., 2019) (Prakash et al., 2021). Dalam beberapa tahun terakhir, perkembangan teknologi kecerdasan buatan (AI) telah memungkinkan kendaraan otonom untuk semakin beradaptasi dengan kondisi jalan dan lingkungan yang dinamis (Mitchell et al., 2019).

Sistem kendaraan otonom modern bergantung pada integrasi berbagai jenis sensor untuk mengumpulkan data penting yang digunakan dalam proses persepsi dan pengambilan keputusan. Sensor-sensor seperti kamera RGB-D, LiDAR, dan GPS, misalnya, memberikan informasi visual, kedalaman, lokasi, dan kecepatan secara real-time, yang sangat penting untuk pemetaan lingkungan sekitar serta kendali kendaraan (Yan et al., 2020). Meskipun kemajuan teknologi sensor telah memfasilitasi peningkatan persepsi kendaraan otonom, tantangan besar dalam pengembangannya adalah integrasi data dari berbagai sensor secara efisien dalam waktu nyata. Pendekatan berbasis modular, yang memproses data dari setiap sensor secara terpisah, sering kali menimbulkan risiko kesalahan akumulatif dan kehilangan informasi penting, yang dapat mengurangi keandalan sistem secara keseluruhan(Gawlikowski et al., 2021). Untuk mengatasi tantangan ini, pendekatan end-to-end multi-task learning menjadi salah satu solusi yang menjanjikan.

Pendekatan multi-task learning memungkinkan berbagai tugas, seperti persepsi lingkungan, prediksi jalur (waypoint prediction), dan kendali kendaraan, untuk dipelajari secara bersamaan dalam satu model terintegrasi. Dalam pendekatan ini, model dapat berbagi representasi fitur antara tugas-tugas yang berbeda, yang dapat meningkatkan efisiensi dan akurasi sistem. Salah satu tugas utama dalam multi-task learning untuk kendaraan otonom adalah prediksi waypoints, yang bertujuan untuk memprediksi titik-

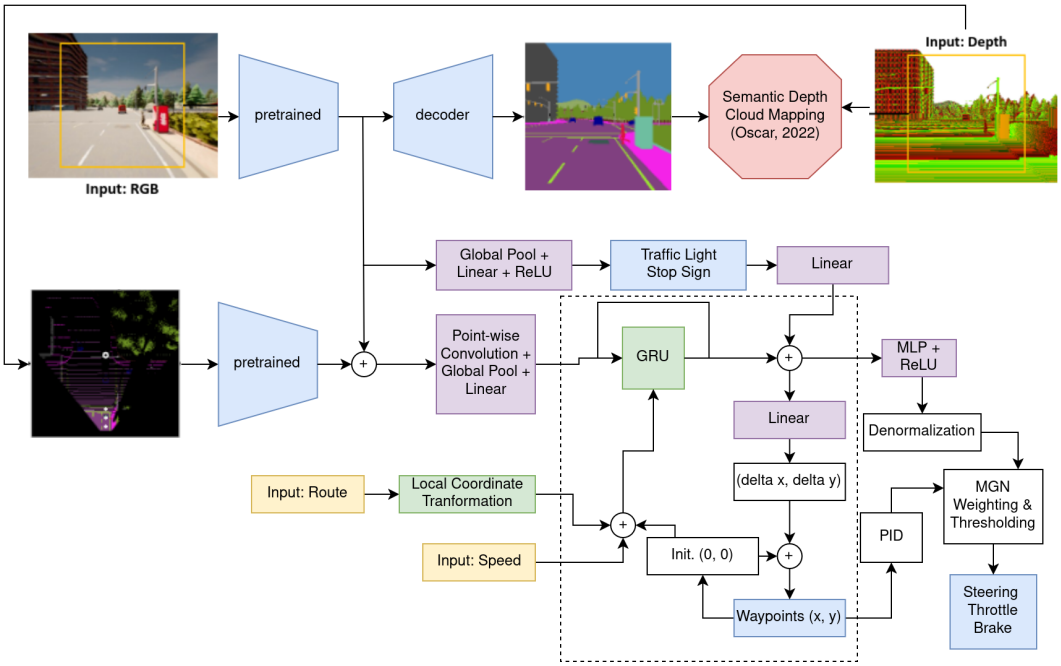
titik koordinat pada lintasan yang akan diikuti oleh kendaraan. Waypoints ini penting karena memberikan panduan bagi sistem kontrol agar kendaraan dapat bernavigasi dengan aman dan akurat di berbagai kondisi jalan (Gutiérrez et al., 2020).

Pada penelitian (Chitta et al., 2022)(Natan & Miura, 2023)(Codevilla et al., 2019)(Liang et al., 2020), CNN digunakan sebagai alur utama dalam arsitektur kendaraan otonom nya. Model berbasis Convolutional Neural Network (CNN), yang juga menjadi komponen penting dalam arsitektur kendaraan otonom, memiliki kelemahan utama dalam memahami hubungan global dalam data visual. CNN cenderung memahami hubungan global secara bertahap dan fokus pada fitur lokal di layer awal dari gambar melalui penerapan filter konvolusi, sehingga sulit untuk menangkap konteks global yang lebih kompleks(Raghu et al., 2021), seperti interaksi antara rambu lalu lintas dan lingkungan jalan secara keseluruhan. Kelemahan ini semakin menegaskan perlunya pendekatan baru yang mampu mengatasi keterbatasan arsitektur tradisional untuk mendukung kemampuan navigasi kendaraan otonom yang lebih andal dan adaptif.

Transformer muncul sebagai arsitektur alternatif yang lebih canggih untuk tugas-tugas sekuensial. Transformer, yang diperkenalkan pertama kali oleh (Vaswani et al., 2017), menggunakan mekanisme self-attention yang memungkinkan model untuk memproses keseluruhan urutan data secara paralel. Mekanisme ini memungkinkan Transformer untuk menangkap hubungan antar-elemen data dengan lebih baik, bahkan dalam urutan yang panjang (Prakash et al., 2021)(Ravanbakhsh et al., 2018). Oleh sebab itu, transformer sangat handal pada tugas pengolahan bahasa alami dan berhasil membawanya sebagai basis untuk Large Language Model di masa sekarang.

Pada Gambar 1.1, ditampilkan arsitektur x13 yang digunakan dalam sistem prediksi waypoints dan navigasi kendaraan otonom. Arsitektur ini terdiri dari beberapa komponen utama, termasuk EfficientNet sebagai backbone untuk ekstraksi fitur dari input RGB, Gated Recurrent Unit (GRU) untuk memproses data sekuensial dan menghasilkan prediksi waypoints, serta encoder untuk ekstraksi fitur dari semantic depth cloud. Dalam arsitektur ini, input berupa gambar RGB dan data kedalaman (depth) diproses melalui modul segmentasi untuk menghasilkan peta kedalaman semantic depth cloud (SDC), yang kemudian digunakan dalam proses pengambilan keputusan terkait navigasi.

Setelah tahap segmentasi dan SDC, 2 fitur di gabung, dilakukan pooling global dan



Gambar 1.1: Arsitektur x13 (Natan & Miura, 2023).

pemetaan data melalui layer linear untuk mendeteksi elemen-elemen penting seperti rambu lalu lintas. Informasi itu dan data sensor lainnya kemudian diintegrasikan melalui GRU untuk memprediksi waypoints berikutnya berdasarkan data sebelumnya. Waypoints ini membantu dalam menentukan lintasan kendaraan. Hasil dari waypoint prediction selanjutnya diproses melalui modul kendali (control module) yang mengatur respons kendaraan, termasuk pengaturan kemudi, throttle, dan rem, melalui metode Proportional-Integral-Derivative (PID).

Pada penelitian (Natan & Miura, 2023), model arsitektur prediksi waypoints yang digunakan menunjukkan akurasi driving score sebesar 35,982 dalam simulasi CARLA. Sebagai perbandingan, driving score yang dihasilkan oleh pengemudi ahli mencapai 41,579. Hal ini menunjukkan adanya kesenjangan yang signifikan antara performa model dengan standar yang dicapai oleh pengemudi manusia. Kesenjangan tersebut menjadi salah satu motivasi utama dalam penelitian ini untuk mengembangkan arsitektur yang lebih canggih dan adaptif, terutama melalui penerapan model berbasis transformer, guna meningkatkan akurasi driving score serta mendekatkannya ke tingkat performa pengemudi ahli.

Dalam penelitian ini, fokus utama adalah pada rekayasa arsitektur prediksi waypoints dan navigasi untuk kendaraan otonom dengan memanfaatkan kapabilitas model berbasis

transformer. Arsitektur yang diusulkan memiliki tiga komponen penting: backbone untuk ekstraksi fitur, modul segmentasi untuk menghasilkan peta kedalaman semantik, dan prediksi waypoints melalui unit rekursif. Penggantian backbone berbasis CNN dengan model transformer seperti Swin Transformer (Wu et al., 2022)dan Vision Transformer (ViT) (Dosovitskiy et al., 2020). Model-model ini telah terbukti unggul dalam berbagai tugas pengenalan objek dan segmentasi, berkat pendekatan mereka yang efisien dalam menangani gambar sebagai patch atau urutan sub-bagian(Z. Yang et al., 2022)(S. He et al., 2024). Dengan demikian, pemanfaatan transformer dapat memperkaya pemahaman terhadap elemen-elemen gambar, seperti rambu lalu lintas dan marka jalan, yang sangat penting dalam navigasi kendaraan otonom.

Selanjutnya, komponen kedua yang akan dimodifikasi adalah encoder pada SDC. Penggantian encoder SDC berbasis CNN dengan metode segmentasi berbasis ViT dan Swin Transformer dapat memberikan fleksibilitas dan ketajaman dalam membedakan elemen-elemen penting pada gambar, terutama dalam kondisi jalan yang dinamis (Rasoulian et al., 2022)(Rasoulian et al., 2022; Strudel et al., 2021). Dengan keunggulan yang dimiliki SDC yaitu penglihatan tampak atas yang tidak memerlukan lidar. Dengan demikian transformer dapat menangkap fitur tampak atas dengan lebih baik dibandingkan CNN seperti pejalan kaki atau kendaraan beberapa langkah kedepan yang tidak bisa dilakukan oleh semantik segmentasi biasa.

Modifikasi terakhir yang akan dilakukan adalah pada model prediksi waypoints. Penggantian Gated Recurrent Unit (GRU) dengan model berbasis transformer dapat meningkatkan akurasi prediksi wapoints. Model-model ini dirancang khusus untuk mempertimbangkan interaksi jangka panjang antara waypoints sebelumnya dan lingkungan, yang memungkinkan mereka untuk memprediksi waypoints dengan lebih baik dalam situasi yang kompleks. Dengan kemampuan untuk mempertimbangkan seluruh urutan waypoints secara bersamaan, model berbasis transformer ini diharapkan dapat mengatasi keterbatasan yang ada pada GRU yang hanya mempertimbangkan hubungan lokal. Selain modifikasi arsitektur, penelitian ini juga akan melibatkan perbandingan performa antara model. Perbandingan ini akan mencakup akurasi prediksi waypoints dan kemampuan adaptasi terhadap variasi lingkungan jalan (Gu et al., 2019)(Xie et al., 2021).

1.2 Rumusan Masalah

1. Bagaimana pemanfaatan arsitektur Transformer dalam komponen arsitektur berbasis CNN, seperti EfficientNet dan GRU, untuk meningkatkan performa prediksi waypoints dan navigasi pada kendaraan otonom?
2. Apakah penggunaan Transformer mampu meningkatkan akurasi dan responsivitas sistem kendaraan otonom terhadap perubahan kondisi jalan yang dinamis?
3. Apa saja kendala teknis yang muncul dalam penerapan arsitektur Transformer, khususnya terkait kebutuhan komputasi dan kecepatan inferensi?

1.3 Tujuan Penelitian

1. Mengimplementasikan arsitektur Transformer sebagai pada arsitektur berbasis CNN tradisional dalam tugas multi-task learning untuk prediksi waypoints dan navigasi kendaraan otonom.
2. Mengeksplorasi dampak penggunaan Transformer pada akurasi dan responsivitas sistem kendaraan otonom dalam menghadapi kondisi jalan yang dinamis.
3. Mengidentifikasi tantangan teknis dan kebutuhan sumber daya yang muncul dalam penerapan arsitektur Transformer pada model multi-task learning kendaraan otonom.

1.4 Batasan Masalah

Dalam penelitian ini, terdapat beberapa batasan yang perlu diperhatikan untuk memperjelas ruang lingkup dan fokus studi yang akan dilakukan, antara lain:

1.4.1 Lingkup Data Sensor yang Digunakan

Penelitian ini akan menggunakan data dari sensor kamera RGB-D dan informasi lokasi dari GPS untuk proses persepsi lingkungan dan pengambilan keputusan. Data dari sensor

tambahan seperti LiDAR atau radar tidak akan dipertimbangkan dalam penelitian ini, sehingga hasil penelitian terbatas pada arsitektur yang mengandalkan data visual dan kedalaman dari kamera RGB-D.

1.4.2 Jenis Arsitektur Transformer yang Digunakan

Penelitian ini hanya akan mengevaluasi dua jenis arsitektur Transformer, yaitu Swin Transformer dan Vision Transformer (ViT). Arsitektur Transformer lain seperti BERT atau generative Transformers tidak akan dieksplorasi dalam penelitian ini. Hal ini dilakukan untuk membatasi ruang lingkup dan fokus pada Transformer yang secara spesifik dirancang untuk pemrosesan visual.

1.4.3 Evaluasi Kinerja pada Lingkungan Simulasi

Pengujian dan evaluasi kinerja model akan dilakukan di lingkungan simulasi untuk kendaraan otonom, bukan dalam kondisi nyata di jalan raya. Hal ini dilakukan untuk menjaga keselamatan serta mengurangi biaya, namun juga berarti hasil penelitian mungkin memiliki keterbatasan dalam penerapan langsung pada kondisi dunia nyata yang lebih kompleks dan beragam.

1.4.4 Pengukuran Kinerja Terbatas pada Tugas Prediksi Waypoints dan Segmentasi Semantik

Penelitian ini akan fokus pada evaluasi performa dalam dua tugas utama: prediksi waypoints dan segmentasi semantik. Metrik seperti akurasi prediksi waypoints, kualitas segmentasi semantik, dan waktu respons model akan menjadi parameter evaluasi utama. Penelitian ini tidak mencakup tugas lain seperti deteksi objek atau klasifikasi jalan, sehingga hasil penelitian hanya terbatas pada evaluasi kedua tugas ini.

1.4.5 Dataset Terbatas untuk Latihan dan Pengujian

Dataset yang digunakan untuk pelatihan dan pengujian model akan berasal dari dataset publik untuk kendaraan otonom yang tersedia, yaitu Carla, yang mungkin tidak sepenuhnya mencakup variasi kondisi jalan dan cuaca di dunia nyata. Hal ini bisa mempengaruhi generalisasi model dalam kondisi lingkungan yang tidak ada di dataset pelatihan.