Imitation with Transformer-Based Sensor Fusion for Autonomous Driving

Kashyap Chitta

University of Tübingen
MPI for Intelligent Systems

Autonomous Vision Group
Team

Kashyap Chitta  Aditya Prakash  Bernhard Jaeger

Zehao Yu  Katrin Renz  Andreas Geiger
Covered Papers

- **Multi-Modal Fusion Transformer for End-to-End Autonomous Driving**

- **TransFuser: Imitation with Transformer-Based Sensor Fusion**
Evaluating Self-Driving
A common task framework accelerates research progress.

Computer vision: **static benchmarks**

How can the community compare **dynamic self-driving** agents?
CARLA Leaderboard

- 10 routes x 2 weathers x 5 repetitions
- 173 Km of driving experiences

Slide Credit: German Ros  https://carla.org/  https://leaderboard.carla.org/
CARLA Leaderboard Evaluation

\[ p_i = \prod_{j \in \mathcal{J}} (p^j)^{v_i} \]

- \# of routes
- Completion of route \( i \)
- Infraction penalty for route \( i \)
- Number of infractions of type \( j \) in route \( i \)
- Penalty for infraction of type \( j \)
Imitation Learning for CARLA
Motivation: Hand-designing a sensor-based driving policy is difficult
Imitation Learning

**Motivation:** Hand-designing a sensor-based driving policy is difficult

- **Step 1:** Hand-design expert which uses privileged information
Motivation: Hand-designing a sensor-based driving policy is difficult

- **Step 1:** Hand-design expert which uses privileged information
- **Step 2:** Train sensor-based policy to mimic demonstrator
Sensor Fusion
Sensors

**RGB Camera**

- Dense RGB input
- Lacks reliable 3D information
- Variation in weather
Sensors

**RGB Camera**
- Dense RGB input
- Lacks reliable 3D information
- Variation in weather

**LiDAR Point Cloud**
- 3D information
- Sparse input
- No traffic light state

Geometric Fusion

3D Space

Image Space

Geometric Feature Projections

Geometric Fusion Lacks Global Context

- From the yellow region, geometric fusion aggregates features to the blue region.
Geometric Fusion Lacks Global Context

- From the yellow region, geometric fusion aggregates features to the blue region.
- However, for safe navigation, it is useful to aggregate features for the red region since it contains vehicles which are affected by the traffic light.
TransFuser
Key Idea

Use attention-based feature fusion to capture the global context of the scene across modalities.
TransFuser

TransFuser

Full Architecture

RGB Image

LiDAR BEV

Image Branch

Transformers

BEV Branch

Conv. Decoder

Conv. Decoder

MLP

512

64

Depth

Semantic Segmentation

GRU Decoder for waypoints

HD Map

Goal Location

Bounding Boxes

Loss Functions

- $L_1$ loss on waypoints: $\mathcal{L} = \sum_{t=1}^{4} ||w_t - w_{gt}||_1$
- Cross-entropy loss on semantics
- $L_1$ loss on depth
- Cross-entropy loss on HD map
- Focal loss on CenterNet heatmaps
- $L_1$ loss on CenterNet offsets

Experiments

**Dataset**

- 8 Towns and randomized weather conditions in CARLA
- Expert policy based on MPC
- $\sim 3.5k$ short routes with hand-crafted scenarios
Experiments

**Dataset**
- 8 Towns and randomized weather conditions in CARLA
- Expert policy based on MPC
- ~3.5k short routes with hand-crafted scenarios

**Sensors**
- RGB cameras: 704×160 resolution, 132° FOV
- LiDAR: 32m range, 64 channels, 10 Hz rotation frequency
Experiments

**Dataset**
- 8 Towns and randomized weather conditions in CARLA
- Expert policy based on MPC
- \(\sim 3.5k\) short routes with hand-crafted scenarios

**Sensors**
- RGB cameras: 704×160 resolution, 132° FOV
- LiDAR: 32m range, 64 channels, 10 Hz rotation frequency

**Evaluation**
- Long routes (\(\sim 2\)km) with dense traffic
- Ensemble of 3 training runs to reduce variance

## Results: Internal Benchmark

<table>
<thead>
<tr>
<th>Method</th>
<th>Driving Score ( \pm )</th>
<th>Route Completion ( \pm )</th>
<th>Infraction Score ( \pm )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Late Fusion</td>
<td>22 ± 4</td>
<td>83 ± 3</td>
<td>0.27 ± 0.03</td>
</tr>
<tr>
<td>Geometric Fusion</td>
<td>27 ± 1</td>
<td>91 ± 1</td>
<td>0.30 ± 0.02</td>
</tr>
<tr>
<td>TransFuser (Ours)</td>
<td>47 ± 6</td>
<td>93 ± 1</td>
<td>0.50 ± 0.00</td>
</tr>
<tr>
<td>Privileged Expert</td>
<td>77 ± 2</td>
<td>89 ± 1</td>
<td>0.86 ± 0.03</td>
</tr>
</tbody>
</table>

- Geometric Fusion, TransFuser and Expert have similar route completion
- Clear trend in infraction score (Expert > TransFuser > Baselines)
CARLA Leaderboard

<table>
<thead>
<tr>
<th>Method</th>
<th>Driving Score ↑</th>
<th>Route Completion ↑</th>
<th>Infraction Score ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAV</td>
<td>62</td>
<td>94</td>
<td>0.64</td>
</tr>
<tr>
<td>TransFuser (Ours)</td>
<td>61</td>
<td>87</td>
<td>0.71</td>
</tr>
<tr>
<td>GRIAD</td>
<td>37</td>
<td>62</td>
<td>0.60</td>
</tr>
<tr>
<td>WOR</td>
<td>31</td>
<td>58</td>
<td>0.56</td>
</tr>
</tbody>
</table>

- Simple end-to-end IL (competitors have complex multi-stage training pipelines)
- Rank 2 at submission (April), with **best infraction score** among top methods
- Still **gets blocked** more often than LAV
- DS > 60, rapid overall progress on leaderboard since 2020 (DS < 20)
Conclusions

- Global contextual reasoning is crucial in complex urban scenarios
Summary

Conclusions

- Global contextual reasoning is crucial in complex urban scenarios
- Attention is effective in aggregating information from multiple modalities
Conclusions

- Global contextual reasoning is crucial in complex urban scenarios
- Attention is effective in aggregating information from multiple modalities
- Driving Score of simple Imitation Learning baseline is competitive
Summary

Conclusions

- Global contextual reasoning is crucial in complex urban scenarios
- Attention is effective in aggregating information from multiple modalities
- Driving Score of simple Imitation Learning baseline is competitive

Code

- www.github.com/autonomousvision/transfuser
Other Work

  
  “Driving in diverse environments is eased by mixture policies.”
Other Work

  “Driving in diverse environments is eased by mixture policies.”

  “Vanilla DAGGER doesn’t work well ⇒ we must sample critical states.”
Other Work

  "Driving in diverse environments is eased by mixture policies."

  "Vanilla DAGGER doesn’t work well ⇒ we must sample critical states."

  "Visual abstractions help, but annotating less can be more."
Other Work

   “Driving in diverse environments is eased by mixture policies.”

   “Vanilla DAGGER doesn’t work well ⇒ we must sample critical states.”

   “Visual abstractions help, but annotating less can be more.”

   “BEV predictions from 2D images via neural fields can improve safety.”
Extra Slides
CARLA Leaderboard

- Open **test bed** to evaluate AD agents for the driving task
- Common maps, situations, and metrics
- Built upon the CARLA simulator
- Aim to **accelerate progress** in the research community
CARLA Leaderboard Submission

Agent’s container

CARLA container

CLI

aws
Motivation

Research Questions

▶ How to integrate representations from multiple modalities?
Research Questions

- How to integrate representations from multiple modalities?
- To what extent should the different modalities be processed independently?
Research Questions

▶ How to integrate representations from multiple modalities?
▶ To what extent should the different modalities be processed independently?
▶ What kind of fusion mechanism to use for maximum performance?
Attention-based Feature Fusion

▶ Consider feature maps as **sets of tokens** (cells of grid = tokens)
▶ Pass all tokens to **self-attention** module and reshape back into grid form

Overall Pipeline

- **Step 1 - Privileged Agent (Data Collection)**
  - Demonstrator
  - Routes
  - Sensors
Overall Pipeline

► **Step 1 - Privileged Agent (Data Collection)**
  ► Demonstrator
  ► Routes
  ► Sensors

► **Step 2 - Sensorimotor Agent (Training)**
  ► Architecture
  ► Loss function
  ► Controller
Demonstrator: Components

**Lateral Control**

- Input: HD Map
- A* Planner
- PID controller
Demonstrator: Components

**Longitudinal Control**

- Input: traffic light states
- Input: nearby actor states
  - Position
  - Orientation
  - Velocity
- Kinematic bicycle model
- PID controller
Demonstrator

- Simplified version of Model Predictive Control (MPC)
- 2 candidate trajectories using HD map + PID controllers
  - Greedy: target speed = 4 m/s
  - Conservative: target speed = 0 m/s
- Roll out greedy trajectory with bicycle model
- Choose conservative trajectory if infraction is detected
Routes

- ~3000 Junctions (~100m long)
- ~500 Curves (~400m long)
- 8 CARLA towns (1, 2, 3, 4, 5, 6, 7, 10)
- 7 CARLA scenarios (1, 3, 4, 7, 8, 9, 10)
Routes

- ~ 3000 Junctions (~100m long)
- ~ 500 Curves (~400m long)
- 8 CARLA towns (1, 2, 3, 4, 5, 6, 7, 10)
- 7 CARLA scenarios (1, 3, 4, 7, 8, 9, 10)
- Time of day: custom distribution around 6 preset values
- Weathers: 7 CARLA presets
- Dataset size: 226k frames
RGB cameras

- 3 cameras: front, 60° left, 60° right
- Field of view: 60° each
- Resolution: 320 \times 160 \text{ pixels each}
- Composited into 704 \times 160 \text{ input}
Sensors

64 beam LiDAR

- 10 Hz frequency: use alternate frames
- Field of view: 180°
- Rasterized into BEV (256×256, 32m range)
- 2 channels: ground plane, objects
Sensors

Additional sensors used for auxiliary supervision

- Semantic Segmentation
- Depth
- HD Map: same coordinate frame as LiDAR
Baselines - Late Fusion

RGB Image

ResNet34

ResNet18

LiDAR BEV

MLP

Goal Location

GRU

GRU

GRU

GRU

Goal Location

512

64
Controller

- Heading and target speed from waypoints
- PID controllers
- Inertia problem: creep forward if still for \(~1\) minute
  - Safety check: no creeping when LiDAR indicates close proximity
## Runtime

<table>
<thead>
<tr>
<th>Method</th>
<th>Single Model</th>
<th>Ensemble (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Late Fusion (LF)</td>
<td>23.5</td>
<td>46.7</td>
</tr>
<tr>
<td>Geometric Fusion (GF)</td>
<td>43.5</td>
<td>69.1</td>
</tr>
<tr>
<td>TransFuser (Ours)</td>
<td>27.6</td>
<td>59.6</td>
</tr>
</tbody>
</table>

Table: We show the runtime per frame in ms for each method averaged over all timesteps in a single evaluation route. We measure runtimes for both a single model and an ensemble of three models. A single TransFuser model runs in real-time on an RTX 3090 GPU.
# Auxiliary Tasks

<table>
<thead>
<tr>
<th>Auxiliary Losses</th>
<th>DS ↑</th>
<th>RC ↑</th>
<th>IS ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>44</td>
<td>78</td>
<td>0.58</td>
</tr>
<tr>
<td>No Depth</td>
<td>56</td>
<td>91</td>
<td>0.61</td>
</tr>
<tr>
<td>No Semantics</td>
<td>53</td>
<td>88</td>
<td>0.61</td>
</tr>
<tr>
<td>No HD Map</td>
<td>50</td>
<td>89</td>
<td>0.58</td>
</tr>
<tr>
<td>No Vehicle Detection</td>
<td>53</td>
<td>88</td>
<td>0.60</td>
</tr>
<tr>
<td>All Losses (Worst Seed)</td>
<td>49</td>
<td>90</td>
<td>0.55</td>
</tr>
<tr>
<td>All Losses (Best Seed)</td>
<td><strong>56</strong></td>
<td><strong>92</strong></td>
<td><strong>0.62</strong></td>
</tr>
</tbody>
</table>

Table: **Auxiliary Tasks**. Training without auxiliary losses leads to a significant reduction in RC and DS.
Table: Architecture Ablations. The default configuration fuses in both directions. It uses 4 fusion scales, 4 attention layers.
## Model Inputs

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>DS ↑</th>
<th>RC ↑</th>
<th>IS ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiDAR Range</td>
<td>64m × 32m</td>
<td>49</td>
<td>91</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>64m × 64m</td>
<td>47</td>
<td>90</td>
<td>0.52</td>
</tr>
<tr>
<td>LiDAR Encoder</td>
<td>PointPillars</td>
<td>50</td>
<td>91</td>
<td>0.55</td>
</tr>
<tr>
<td>Camera FOV</td>
<td>120°</td>
<td>49</td>
<td>90</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>90°</td>
<td>42</td>
<td>88</td>
<td>0.51</td>
</tr>
<tr>
<td>No Rasterized Goal</td>
<td>-</td>
<td>54</td>
<td>91</td>
<td>0.60</td>
</tr>
<tr>
<td>No Rotation Aug</td>
<td>-</td>
<td>56</td>
<td>92</td>
<td>0.61</td>
</tr>
<tr>
<td>Default Config</td>
<td>Worst Seed</td>
<td>49</td>
<td>90</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>Best Seed</td>
<td>56</td>
<td>92</td>
<td><strong>0.62</strong></td>
</tr>
</tbody>
</table>

Table: **Model Input Ablations.** The default configuration uses a 32m × 32m LiDAR range and 132° camera FOV.
Table: **Inertia Problem.** Creeping improves the RC in both the setting where we input the velocity to our encoder and our default configuration (no velocity input).