Imitation with Transformer-Based Sensor Fusion for Autonomous Driving

Kashyap Chitta

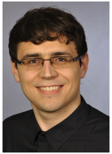
University of Tübingen
MPI for Intelligent Systems

Autonomous Vision Group

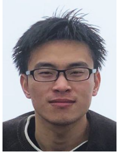
Team

Kashyap Chitta

Aditya Prakash



Bernhard Jaeger



Zehao Yu

Katrin Renz

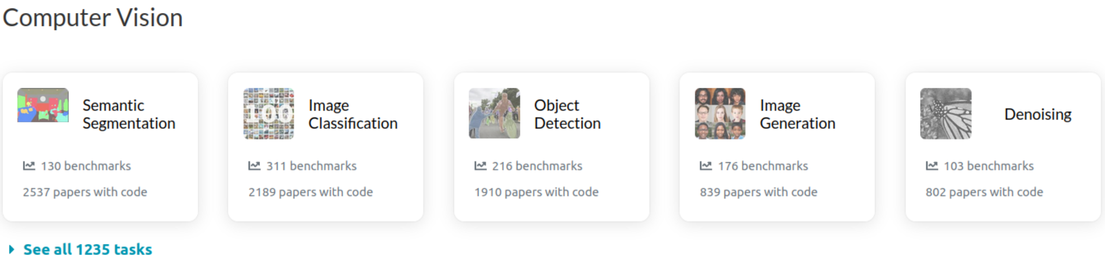
Andreas Geiger

Covered Papers

- **Multi-Modal Fusion Transformer for End-to-End Autonomous Driving**
  A. Prakash*, K. Chitta* and A. Geiger. CVPR, 2021.
- **TransFuser: Imitation with Transformer-Based Sensor Fusion**
  K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz and A. Geiger. PAMI, 2022.

Evaluating Self-Driving

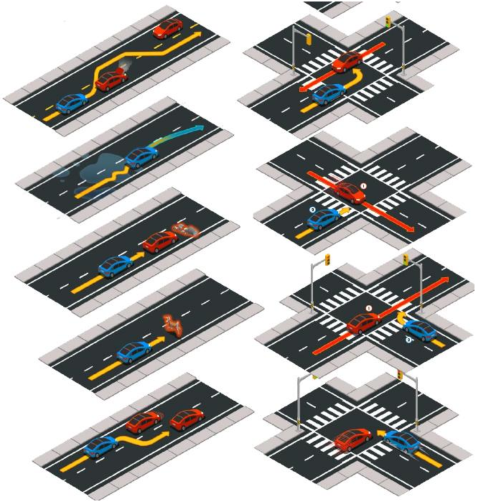
Common Task Framework



- 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



CARLA Leaderboard Evaluation

$$\frac{1}{n}\sum_{i=1}^{n} c_i p_i$$

Infraction penalty for route i

\# of routes

Completion of route i

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

Number of infractions of type j in route i

Penalty for infraction of type j

Imitation Learning for CARLA

Imitation Learning

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

Imitation Learning

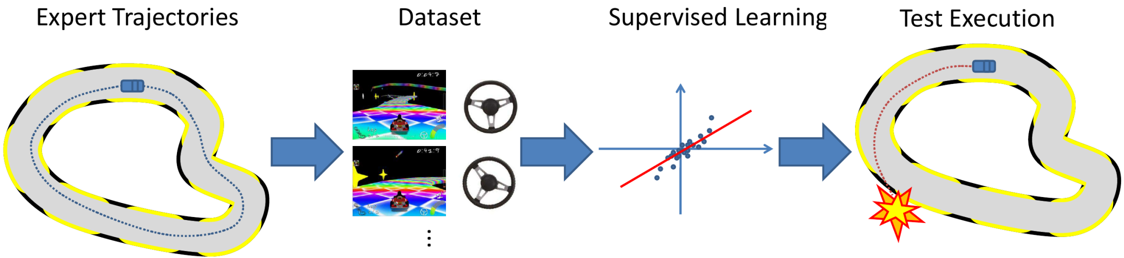
Motivation: Hand-designing a sensor-based driving policy is difficult

- **Step 1:** Hand-design expert which uses privileged information

Imitation Learning

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

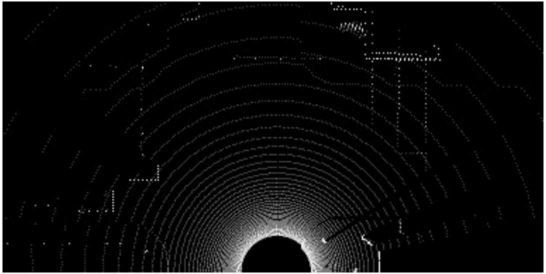
A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

Sensors

RGB Camera

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

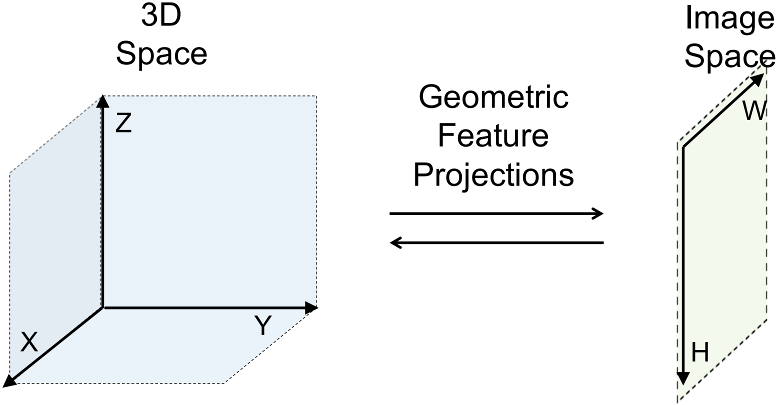
LiDAR Point Cloud



- 3D information
- Sparse input
- No traffic light state

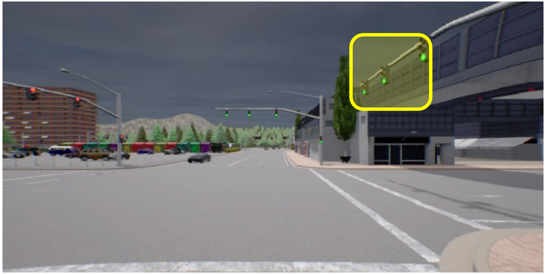
A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

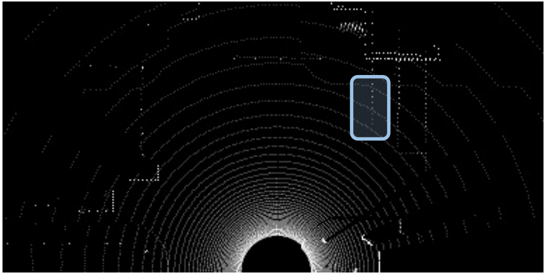
Geometric Fusion



A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

Geometric Fusion Lacks Global Context

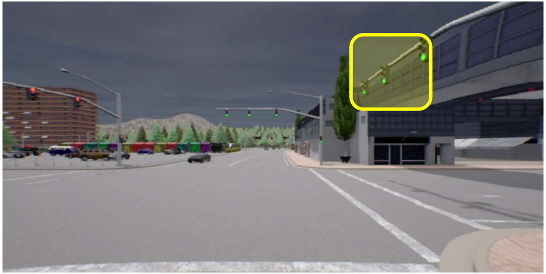


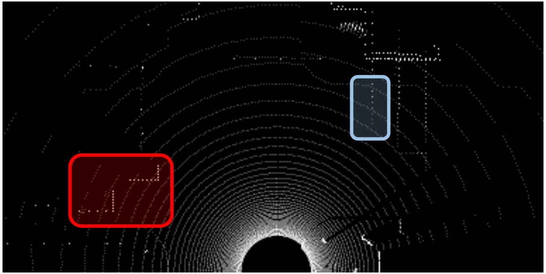


- From the yellow region, geometric fusion aggregates features to the blue region

A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

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

A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

TransFuser

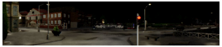
Key Idea

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

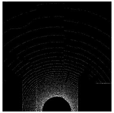
A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

TransFuser

RGB
Image

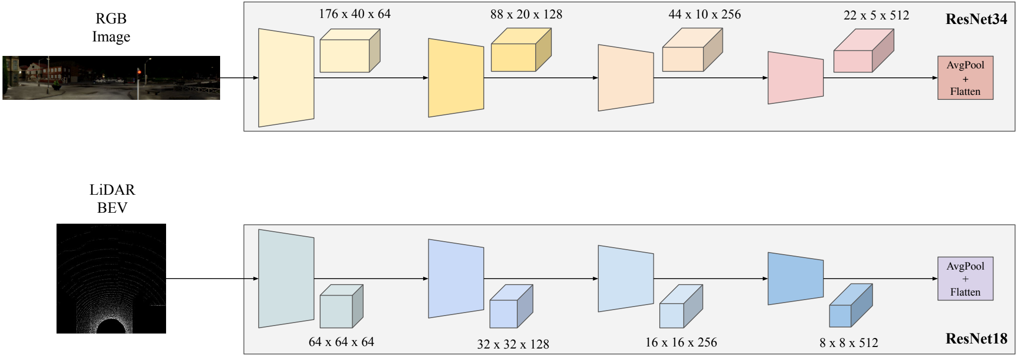


LiDAR
BEV



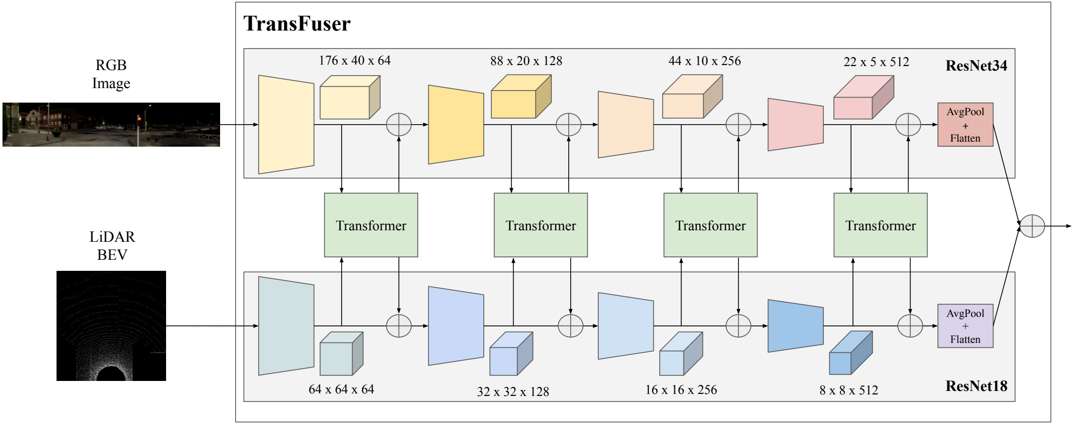
A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

TransFuser



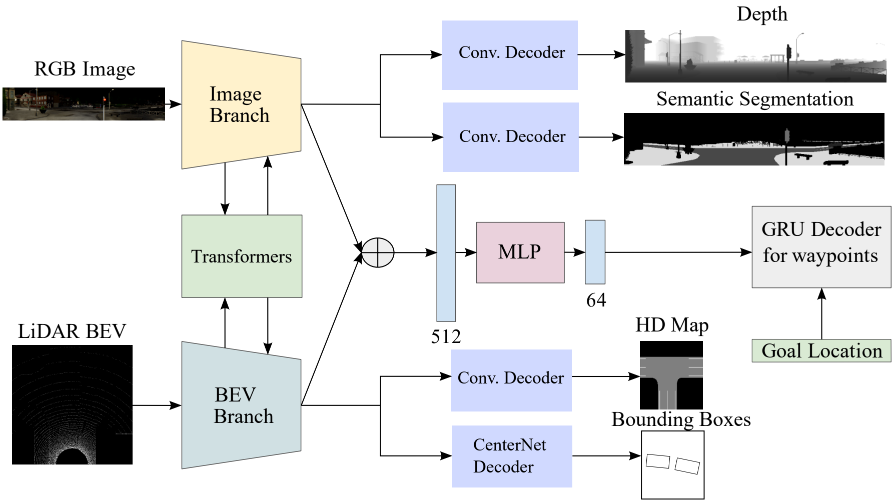
A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

TransFuser



A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

Full Architecture



K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz and A. Geiger: TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

Loss Functions

- L_1 loss on waypoints: $\mathcal{L} = \sum_{t=1}^{4} ||\mathbf{w}_t - \mathbf{w}_t^{gt}||_1$
- 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

K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz and A. Geiger: TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

Experiments

Dataset

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

K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz and A. Geiger: TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

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$\times$160 resolution, 132° FOV
- LiDAR: 32m range, 64 channels, 10 Hz rotation frequency

K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz and A. Geiger: TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

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$2km) with dense traffic
- Ensemble of 3 training runs to reduce variance

K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz and A. Geiger: TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

Results: Internal Benchmark

Method	**Driving Score** $\uparrow$	**Route Completion** $\uparrow$	**Infraction Score** $\uparrow$
Late Fusion	22 ± 4	83 ± 3	0.27 ± 0.03
Geometric Fusion	27 ± 1	91 ± 1	0.30 ± 0.02
TransFuser (Ours)	$\mathbf{47 \pm 6}$	$\mathbf{93 \pm 1}$	$\mathbf{0.50 \pm 0.00}$
Privileged Expert	*77 ± 2*	*89 ± 1*	*0.86 ± 0.03*

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

K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz and A. Geiger: TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

CARLA Leaderboard

Method	**Driving Score** $\uparrow$	**Route Completion** $\uparrow$	**Infraction Score** $\uparrow$
LAV	62	94	0.64
TransFuser (Ours)	61	**87**	**0.71**
GRIAD	37	62	0.60
WOR	31	58	0.56

- 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)

K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz and A. Geiger: TransFuser: Imitiation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

Summary

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

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

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

- Ohn-Bar et al.: Learning Situational Driving. CVPR, 2020.
  “Driving in diverse environments is eased by mixture policies.”

Other Work

- Ohn-Bar et al.: Learning Situational Driving. CVPR, 2020.
  "Driving in diverse environments is eased by mixture policies."

- Prakash et al.: Exploring Data Aggregation in Policy Learning. CVPR, 2020.
  "Vanilla DAGGER doesn't work well $\Rightarrow$ we must sample critical states."

Other Work

- Ohn-Bar et al.: Learning Situational Driving. CVPR, 2020.
  “Driving in diverse environments is eased by mixture policies.”

- Prakash et al.: Exploring Data Aggregation in Policy Learning. CVPR, 2020.
  “Vanilla DAGGER doesn’t work well $\Rightarrow$ we must sample critical states.”

- Behl et al.: Label Efficient Visual Abstractions. IROS, 2020.
  “Visual abstractions help, but annotating less can be more.”

Other Work

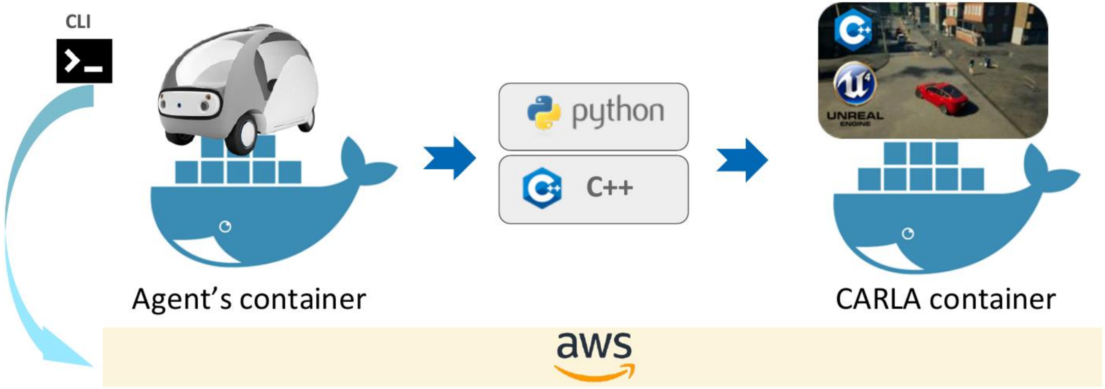
- Ohn-Bar et al.: Learning Situational Driving. CVPR, 2020.
  "Driving in diverse environments is eased by mixture policies."
- Prakash et al.: Exploring Data Aggregation in Policy Learning. CVPR, 2020.
  "Vanilla DAGGER doesn't work well $\Rightarrow$ we must sample critical states."
- Behl et al.: Label Efficient Visual Abstractions. IROS, 2020.
  "Visual abstractions help, but annotating less can be more."
- Chitta et al.: NEAT: Neural Attention Fields. ICCV, 2021.
  "BEV predictions from 2D images via neural fields can improve safety."

Extra Slides

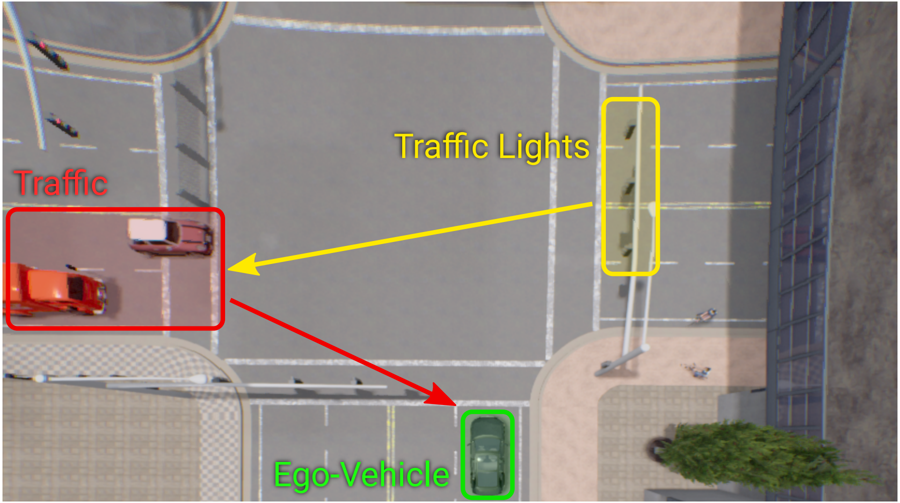
CARLA Leaderboard

- 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



Motivation



A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

Research Questions

- How to integrate representations from multiple modalities?

A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

Research Questions

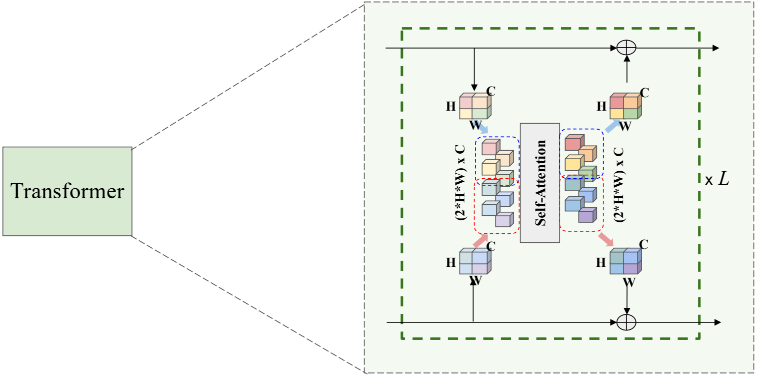
- How to integrate representations from multiple modalities?
- To what extent should the different modalities be processed independently?

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?

A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

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

A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

Overall Pipeline

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

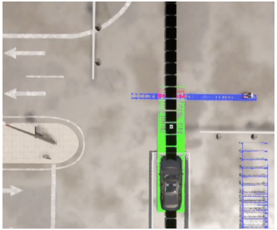
Overall Pipeline

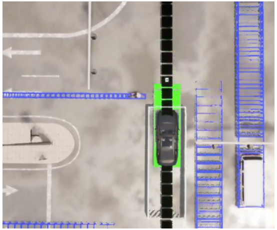
- **Step 1 - Privileged Agent (Data Collection)**
  - Demonstrator
  - Routes
  - Sensors
- **Step 2 - Sensorimotor Agent (Training)**
  - Architecture
  - Loss function
  - Controller

Demonstrator: Components

Lateral Control

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

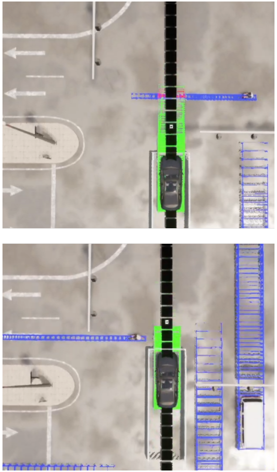




Demonstrator: Components

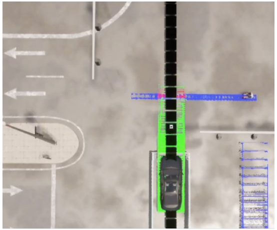
Longitudinal Control

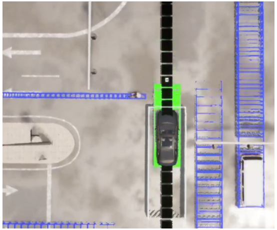
- Input: traffic light states
- Input: nearby actor states
  - Position
  - Orientation
  - Velocity
- Kinematic bicycle model
- PID controller



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

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

Routes

- $\sim$ 3000 Junctions ($\sim$100m long)
- $\sim$ 500 Curves ($\sim$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

Sensors

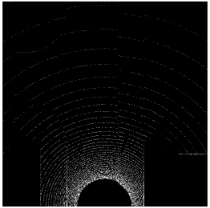
RGB cameras

- 3 cameras: front, 60° left, 60° right
- Field of view: 60° each
- Resolution: 320×160 pixels each
- Composited into 704×160 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

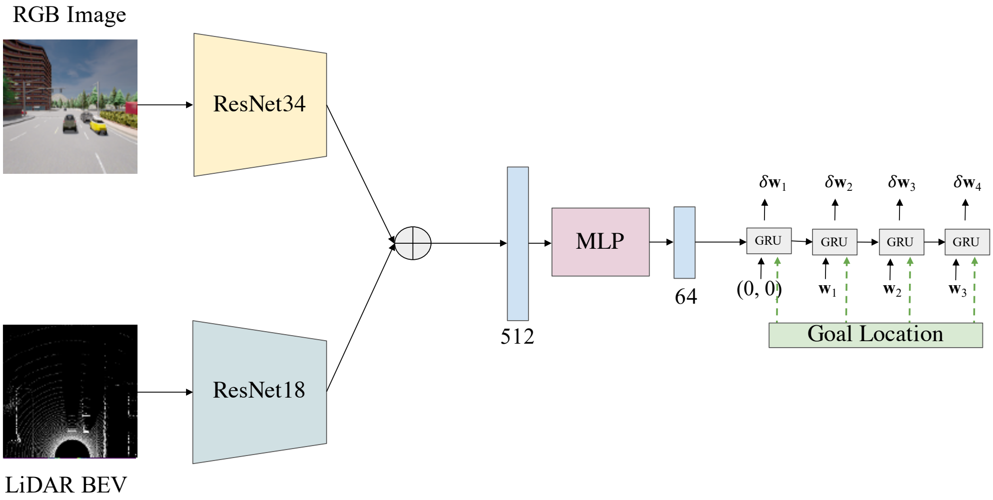


Sensors

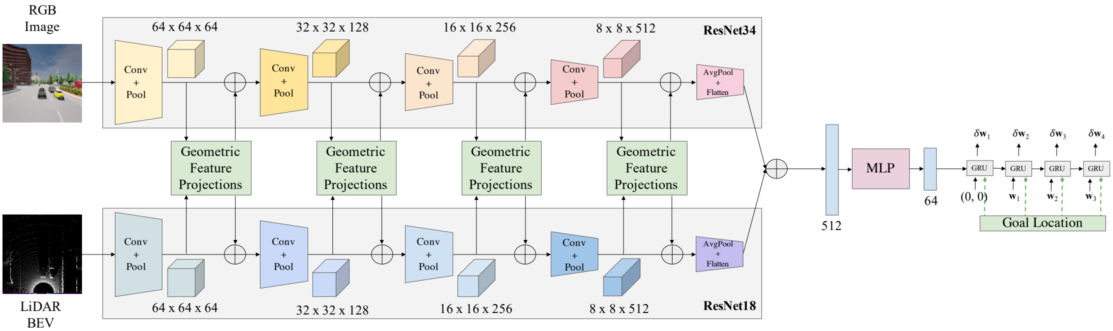
Additional sensors used for auxiliary supervision

- Semantic Segmentaion
- Depth
- HD Map: same coordinate frame as LiDAR

Baselines - Late Fusion



Baselines - Geometric Fusion



Controller

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

Runtime

Method	Single Model	Ensemble (3)
Late Fusion (LF)	23.5	46.7
Geometric Fusion (GF)	43.5	69.1
TransFuser (Ours)	27.6	59.6

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

Auxiliary Losses	**DS** ↑	**RC** ↑	**IS** ↑
None	44	78	0.58
No Depth	56	91	0.61
No Semantics	53	88	0.61
No HD Map	50	89	0.58
No Vehicle Detection	53	88	0.60
All Losses (Worst Seed)	49	90	0.55
All Losses (Best Seed)	**56**	**92**	**0.62**

Table: **Auxiliary Tasks.** Training without auxiliary losses leads to a significant reduction in RC and DS.

Architecture

Parameter	Value	DS ↑	RC ↑	IS ↑
Fusion Direction	LiDAR → Camera	46	87	0.55
	Camera → LiDAR	47	86	0.57
Fusion Scales	1	49	84	0.57
	2	53	91	0.59
	3	48	85	0.60
Attention Layers	2	53	90	0.60
	6	56	92	0.61
	8	56	**92**	0.61
Default Config	Worst Seed	49	90	0.55
	Best Seed	**56**	92	**0.62**

Table: **Architecture Ablations.** The default configuration fuses in both directions. It uses 4 fusion scales, 4 attention layers.

Model Inputs

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

Table: **Model Input Ablations.** The default configuration uses a 32m × 32m LiDAR range and 132° camera FOV.

Inertia Problem

Velocity Input?	Creeping?	DS ↑	RC ↑	IS ↑
-	-	46	78	0.63
	✓	**56**	**92**	0.62
✓	-	37	64	**0.65**
	✓	45	86	0.52

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).