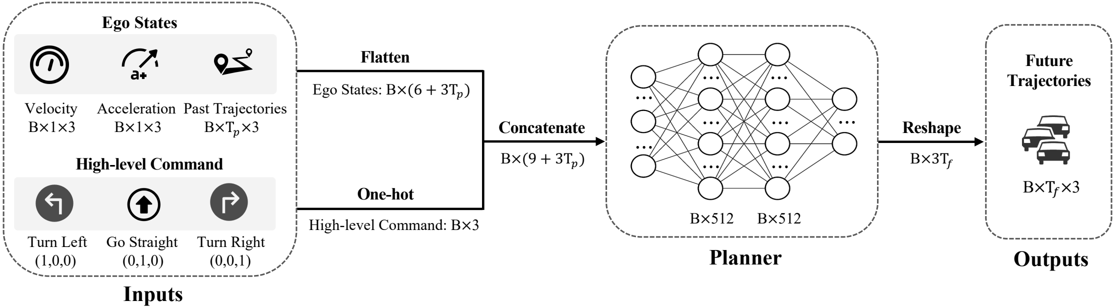
End-to-End Driving with Attention

Kashyap Chitta

ICRA Scalable AD Workshop, 02.06.2023

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Autonomous Driving with an MLP...



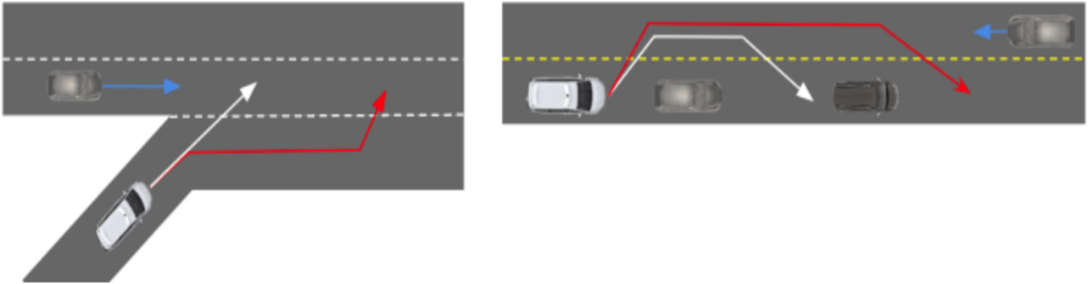
Zhai et al., Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes. ArXiv, 2023.

... is the current SoTA on nuScenes!

Method		Perception Information	Ego States	L2 (m) ↓	Collision (%) ↓
FF	(Hu et al., 2021)	✓	-	1.43	0.43
ST-P3	(Hu et al., 2022)	✓	-	2.11	0.71
EO	(Khurana et al. 2022)	✓	-	1.60	0.44
UniAD	(Hu et al., 2023)	✓	-	1.03	0.31
VAD	(Jiang et al., 2023)	✓	✓	0.37	0.14
AD-MLP	(Zhai et al., 2023)	-	✓	**0.23**	**0.12**

Zhai et al., Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes. ArXiv, 2023.

Open-Loop Evaluation is Flawed



Caesar et al., nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles. CVPR Workshops, 2021.

Closed-Loop Evaluation (CARLA)

https://leaderboard.carla.org/

Closed-Loop Evaluation (CARLA)

Infraction penalty for route i

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

# of routes

Completion of route i

Number of infractions of type j in route i

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

Penalty for infraction of type j

https://leaderboard.carla.org/

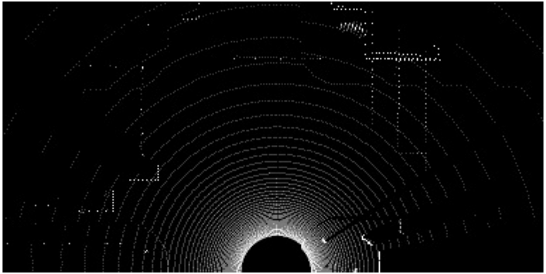
Part 1: Sensor Fusion

Sensor Inputs

RGB Image

- Dense input
- Unreliable 3D information
- Low FOV

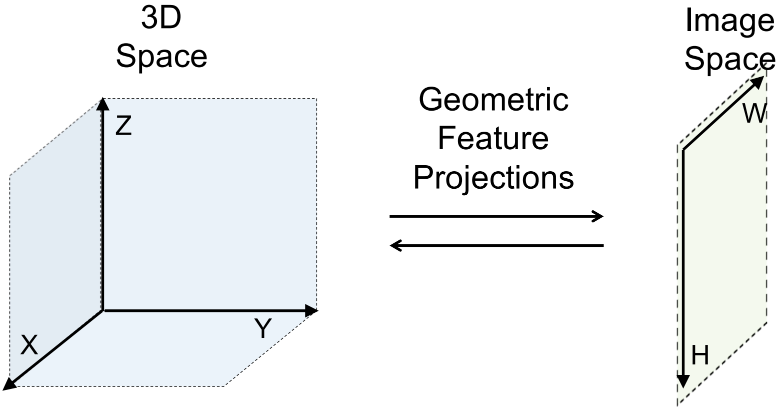
LiDAR Point Cloud



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

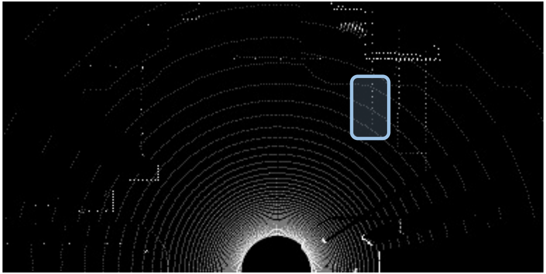
Chitta et al., TransFuser: Imitiation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

Geometric Fusion



Chitta et al., TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

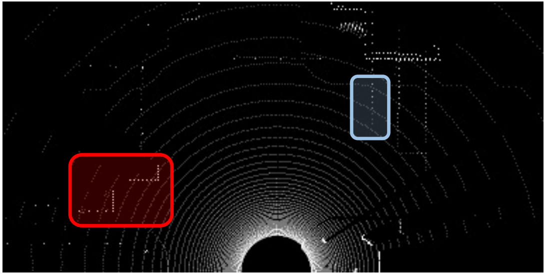
Geometric Fusion Lacks Global Context



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

Chitta et al., TransFuser: Imitiation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

Geometric Fusion Lacks Global Context



- From the yellow region, geometric fusion aggregates features to the blue region
- It is useful to aggregate to the red region (vehicles affected by the traffic light)

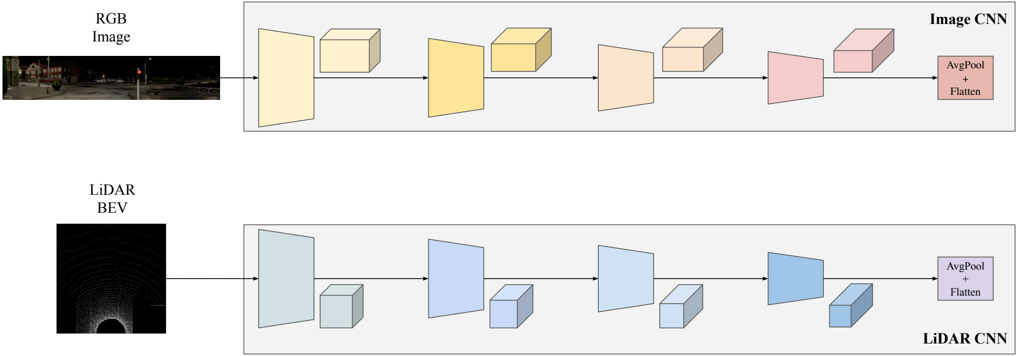
Chitta et al., TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

Key Idea #1

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

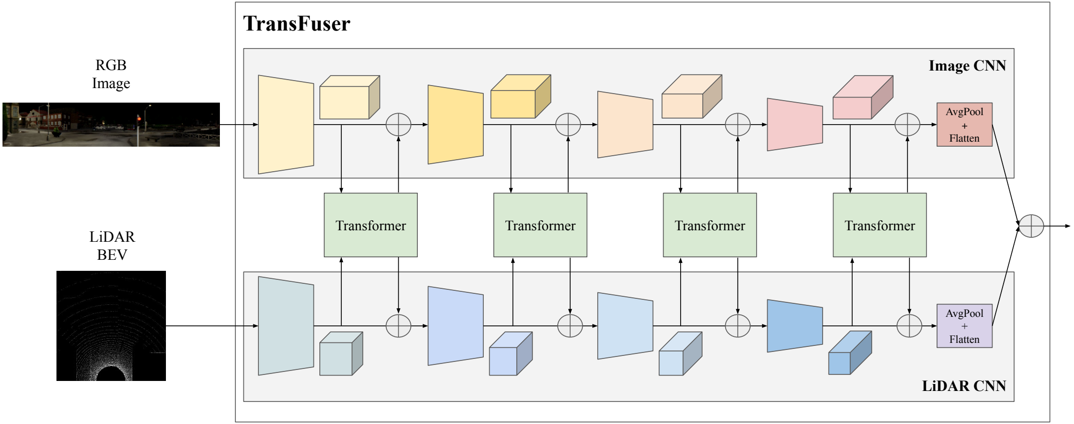
Chitta et al., TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

TransFuser



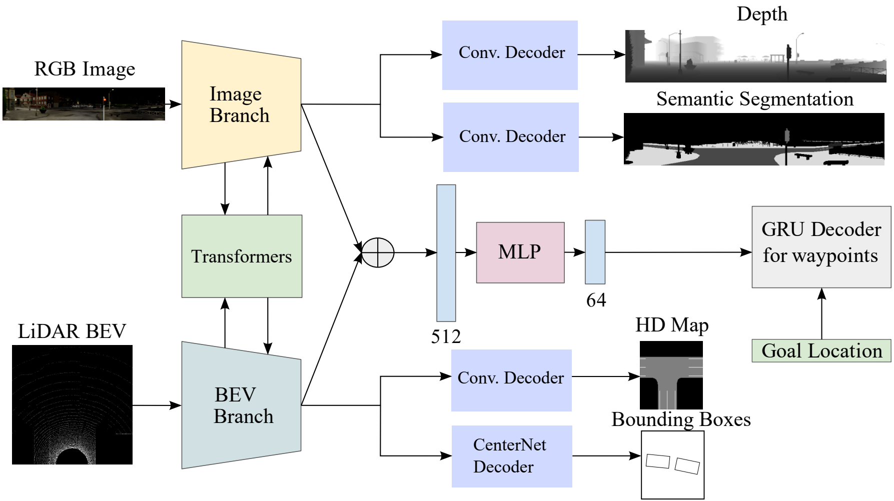
Chitta et al., TransFuser: Imitiation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

TransFuser



Chitta et al., TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

Multi-Task Imitation Learning

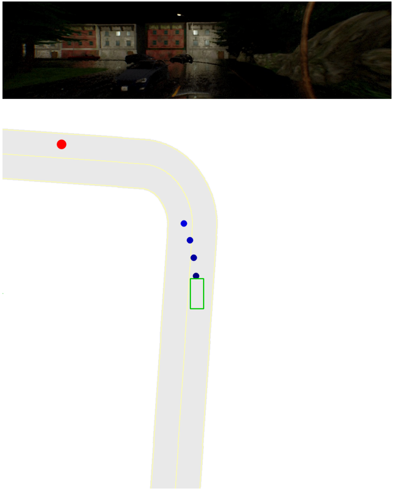


Chitta et al., TransFuser: Imitiation with Transformer-Based Sensor Fusion for Autonomous Driving. PAMI, 2022.

Part 2: A Hidden Bias

TransFuser Extrapolates Predictions to Goal Locations

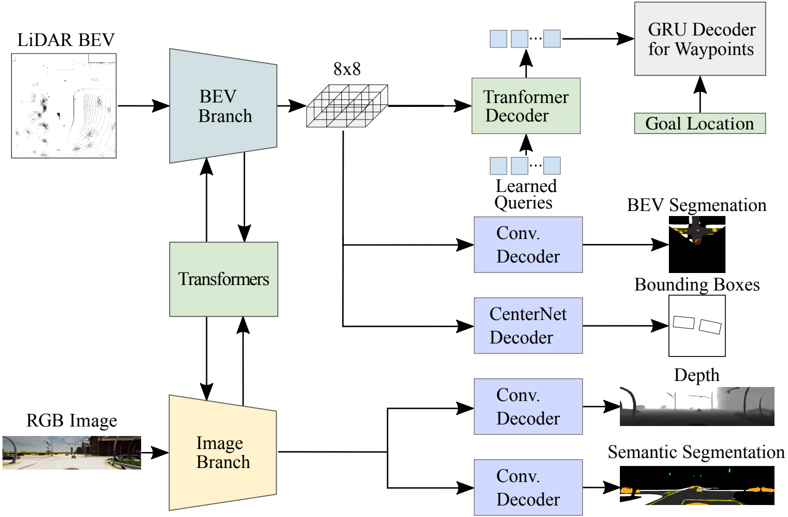
LAV and TCP Extrapolate Predictions to Goal Locations



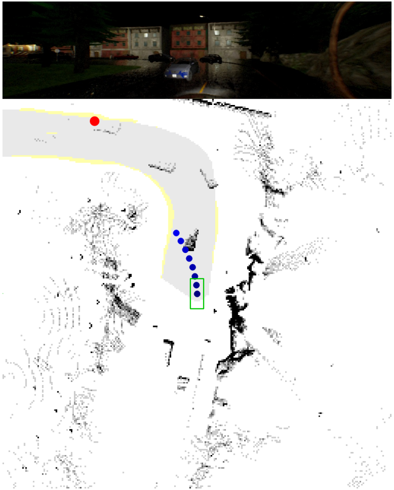
Key Idea #2

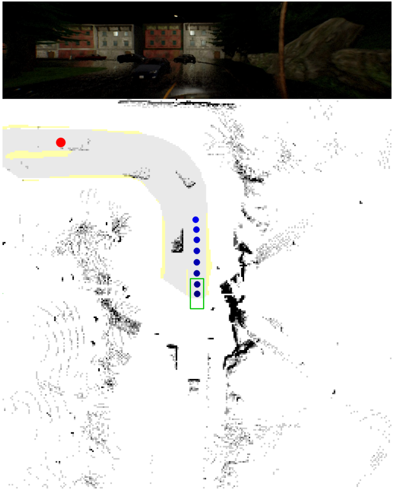
Use **attention-based** feature pooling to preserve the **spatial information** of the encoder features.

Mitigating the Bias

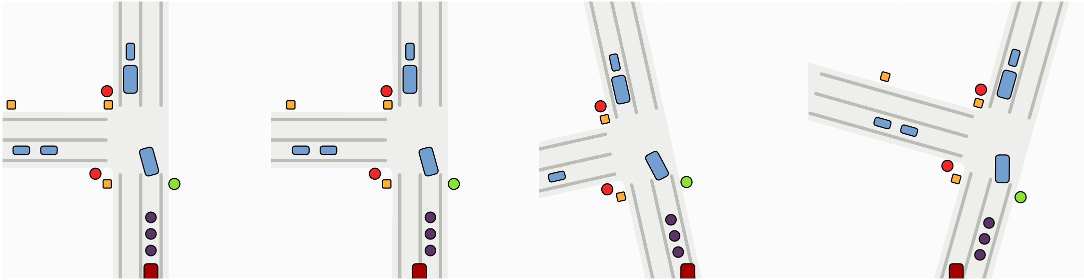


Mitigating the Bias





Shift and Rotation Augmentation



Chen et al., Learning by Cheating. CoRL, 2019.

CARLA Longest6 Benchmark Results

Method	**Driving Score** $\uparrow$	**Route Completion** $\uparrow$
Geometric Fusion	27 ± 1	91 ± 1
TransFuser	49 ± 2	87 ± 0
+ Transformer Decoder	63 ± 4	93 ± 3
+ Augmentation	$\mathbf{71} \pm \mathbf{3}$	$\mathbf{95} \pm \mathbf{3}$

- **+81%** from attention in sensor fusion
- **+45%** from attention in aggregation and augmentation

CARLA Longest6 Benchmark Results

Method		**Driving Score** $\uparrow$	**Route Completion** $\uparrow$
TCP	(Wu et al., NeurIPS 2022)	54 ± 2	78 ± 2
Perc. PlanT	(Renz et al., CoRL 2022)	58 ± 5	88 ± 1
CaT	(Zhang et al., CVPR 2023)	58 ± 2	79 ± 2
ThinkTwice	(Jia et al., CVPR 2023)	61	73
Ours		$\mathbf{71} \pm \mathbf{3}$	$\mathbf{95} \pm \mathbf{3}$

Summary

Key Takeaways

- Open-loop evaluation is flawed

Summary

Key Takeaways

- Open-loop evaluation is flawed
- Attention-based sensor fusion captures global context

Summary

Key Takeaways

- Open-loop evaluation is flawed
- Attention-based sensor fusion captures global context
- Attention-based feature pooling mitigates shortcut learning

Summary

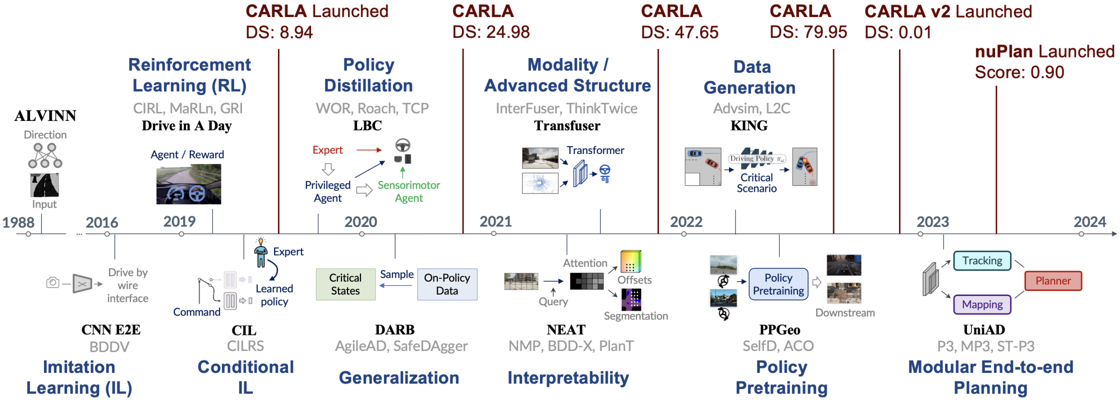
Key Takeaways

- Open-loop evaluation is flawed
- Attention-based sensor fusion captures global context
- Attention-based feature pooling mitigates shortcut learning

Code

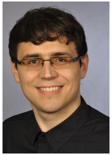
- www.github.com/autonomousvision/transfuser

Rapidly Growing Field



Thank You!

Kashyap Chitta



Bernhard Jaeger

Katrin Renz

Aditya Prakash

Zehao Yu

Andreas Geiger

kashyap7x.github.io