# Driving with Attention

Andreas Geiger

Autonomous Vision Group
University of Tübingen and MPI for Intelligent Systems

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN


ellis
European Laboratory for Learning and Intelligent Systems

# Covered Papers

**Multi-Modal Fusion Transformer for End-to-End Autonomous Driving**
Aditya Prakash, Kashyap Chitta and Andreas Geiger
CVPR 2021

**NEAT: Neural Attention Fields for End-to-End Autonomous Driving**
Kashyap Chitta, Aditya Prakash and Andreas Geiger
ICCV 2021

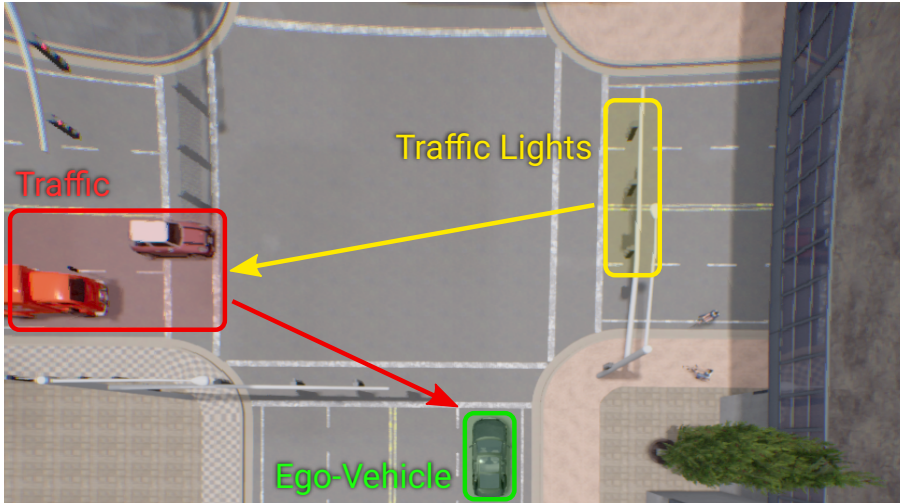
# Collaborators

Aditya Prakash

Kashyap Chitta

# TransFuser: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving

# Motivation



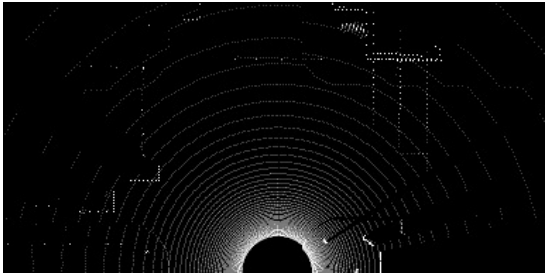

A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

# Sensors

## RGB Camera

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

## LiDAR Point Cloud



- \+ 3D information
- − Sparse input
- − No traffic light state

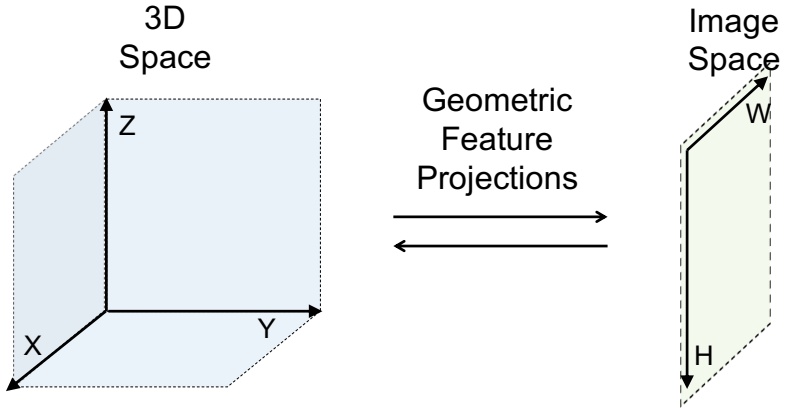
A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

# 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.

# Geometric Fusion




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

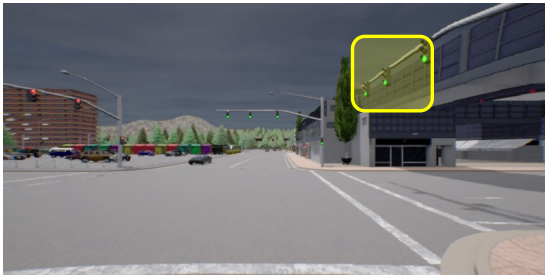
# Geometric Fusion under-performs in complex urban scenarios

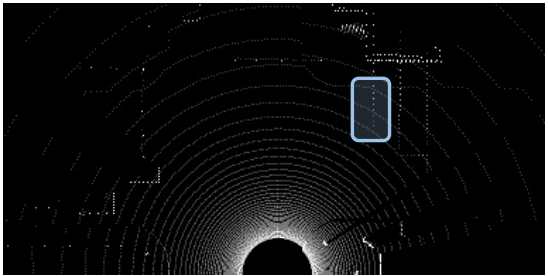

A. Prakash, K. Chitta and A. Geiger: Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. CVPR, 2021.

# Geometric Fusion under-performs in complex urban scenarios


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

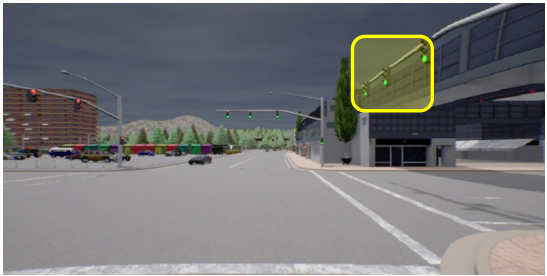
# Geometric Fusion lacks global context

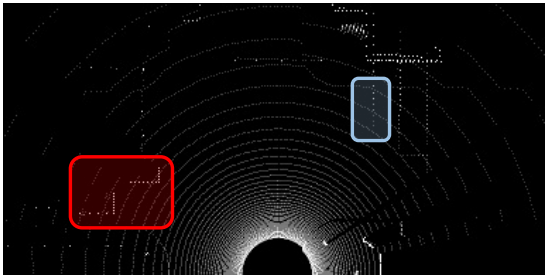




- For the yellow region, geometric fusion aggregates features from the blue region

# Geometric Fusion lacks global context



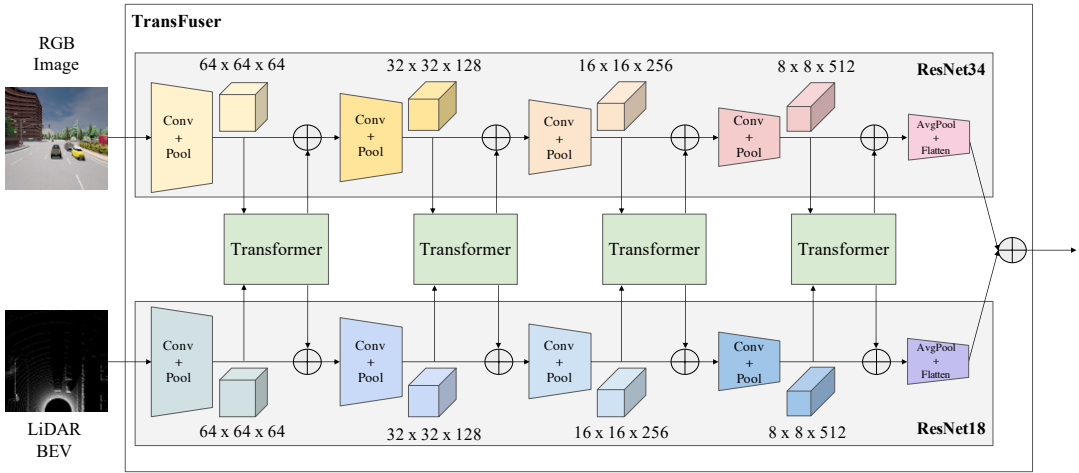


- However, for safe navigation, it is essential to aggregate features from the red region since it contains vehicles which are affected by the traffic light

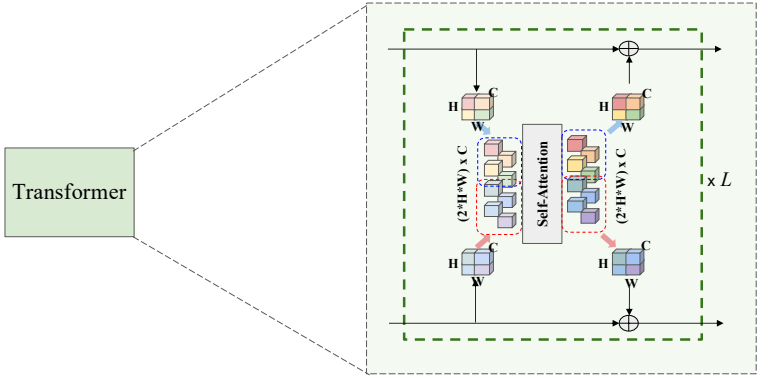
# Key Idea

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

# TransFuser Architecture



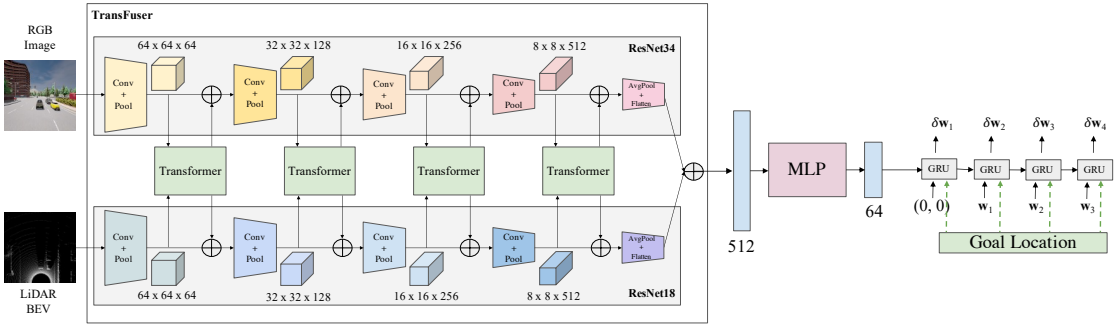
# Attention-based Feature Fusion




- 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.

# Architecture




- GRU-based **autoregressive waypoint prediction** network (conditioned on goal)
- Loss Function: $\mathcal{L} = \sum_{t=1}^{4} ||\mathbf{w}_t - \mathbf{w}_t^{gt}||_1$, waypoints are input to **PID controller**

# Experiments

**Dataset**

- 8 Towns and multiple weather conditions in CARLA
- Routes of varying length in complex urban scenarios
- Expert policy based on A* planner and collision avoidance heuristics

**Sensors**

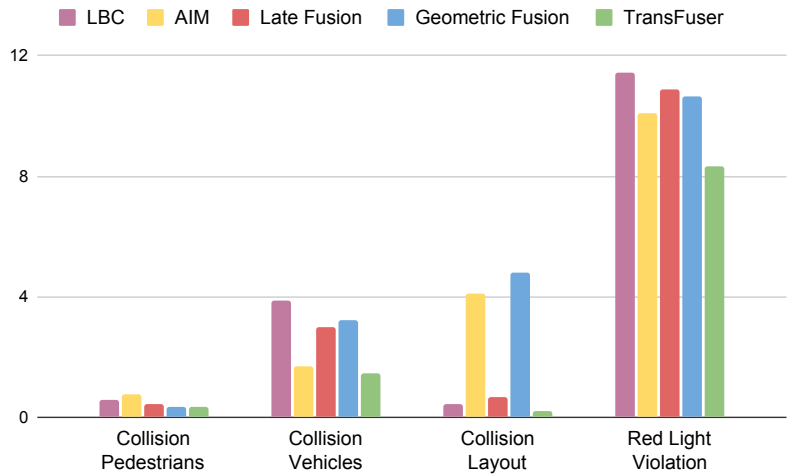
- RGB camera: $400 \times 300$ resolution, $100°$ FOV
- LiDAR: 80m range, 64 channels, 10 Hz rotation frequency

# Results

| Method | Town05 Short | | Town05 Long | |
|---|---|---|---|---|
| | DS ↑ | RC ↑ | DS ↑ | RC ↑ |
| CILRS | 7.47 ± 2.51 | 13.40 ± 1.09 | 3.68 ± 2.16 | 7.19 ± 2.95 |
| Learning by Cheating | 30.97 ± 4.17 | 55.01 ± 5.14 | 7.05 ± 2.13 | 32.09 ± 7.40 |
| Waypoint Prediction | 49.00 ± 6.83 | 81.07 ± 15.59 | 26.50 ± 4.82 | 60.66 ± 7.66 |
| Late Fusion | 51.56 ± 5.24 | 83.66 ± 11.04 | 31.30 ± 5.53 | 68.05 ± 5.39 |
| Geometric Fusion | 54.32 ± 4.85 | **86.91** ± 10.85 | 25.30 ± 4.08 | **69.17** ± 11.07 |
| TransFuser | **54.52** ± 4.29 | 78.41 ± 3.75 | **33.15** ± 4.04 | 56.36 ± 7.14 |

Mean and standard deviation over 9 runs (3 training seeds, 3 rollouts per seed)

# Infraction Analysis




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

# CARLA Leaderboard

| Method | Driving Score ↑ | Route Completion ↑ |
|---|---|---|
| CILRS | 5.37 | 14.40 |
| Learning by Cheating | 8.94 | 17.54 |
| Waypoint Prediction | 12.88 | 41.52 |
| Late Fusion | 13.27 | 42.10 |
| Geometric Fusion | 14.47 | 40.99 |
| TransFuser | **16.93** | **51.82** |

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

# Qualtitative Results

# Qualtitative Results

# Attention Map Visualizations

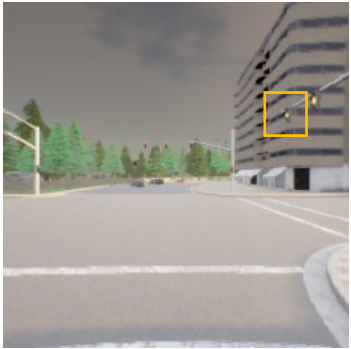
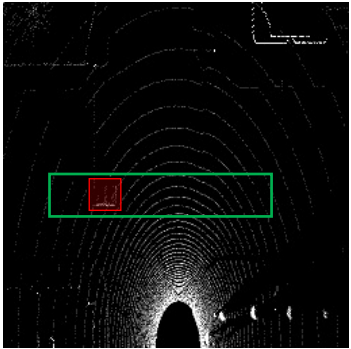


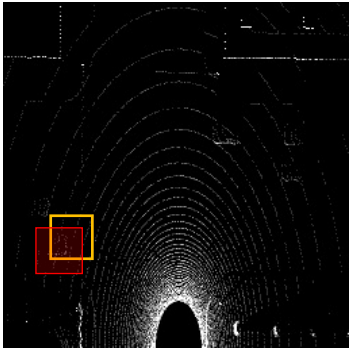
Image: Traffic Light



LiDAR: Vehicle

- Yellow: Query token in source modality. Green: Top-5 tokens in target modality.

# Attention Map Visualizations



LiDAR: Vehicles

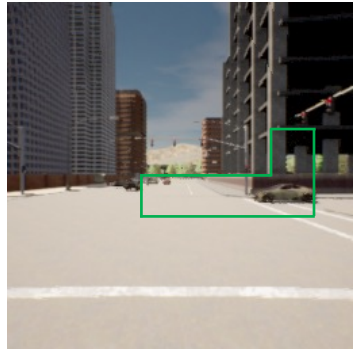
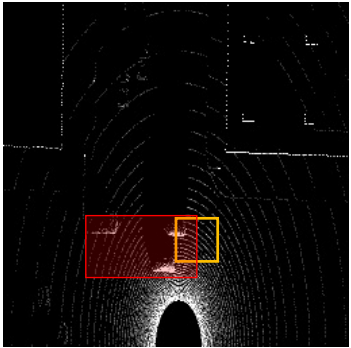


Image: Traffic Light and other Vehicles

- Yellow: Query token in source modality. Green: Top-5 tokens in target modality.

# Attention Map Visualizations



LiDAR: Vehicles

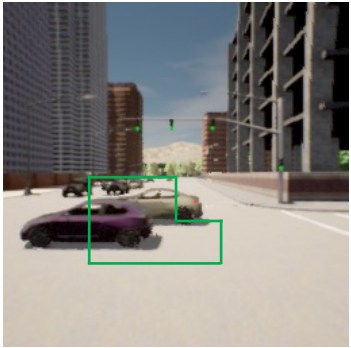


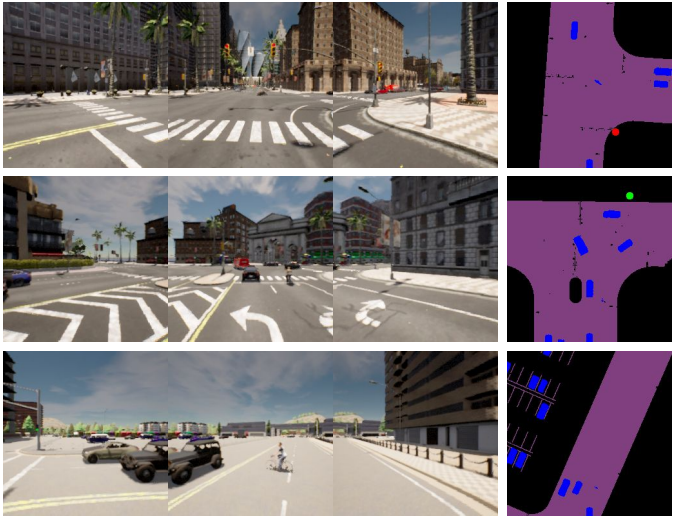
Image: Same Vehicles

- Yellow: Query token in source modality. Green: Top-5 tokens in target modality.

# NEAT: Neural Attention Fields for End-to-End Autonomous Driving

# Bird's-Eye-View (BEV)

- Efficient reasoning about the **semantic, spatial, and temporal structure** is crucial for self-driving
- Driving happens in **physical 3D space**
- BEV $\approx$ 3D space
- How to obtain compact, interpretable BEV repr. from **images** as input?

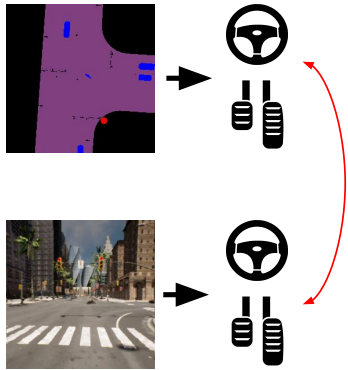


K. Chitta, A. Prakash and A. Geiger: NEAT: Neural Attention Fields for End-to-End Autonomous Driving. ICCV, 2021.

# BEV Semantics for Driving

**Prior work:**

- LBC [CoRL 2019]
- Roach [ICCV 2021]



K. Chitta, A. Prakash and A. Geiger: NEAT: Neural Attention Fields for End-to-End Autonomous Driving. ICCV, 2021.

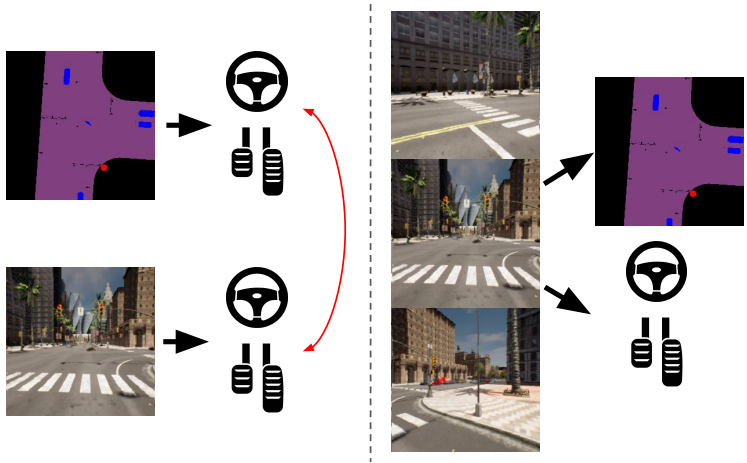
# BEV Semantics for Driving

**Prior work:**

- LBC [CoRL 2019]
- Roach [ICCV 2021]

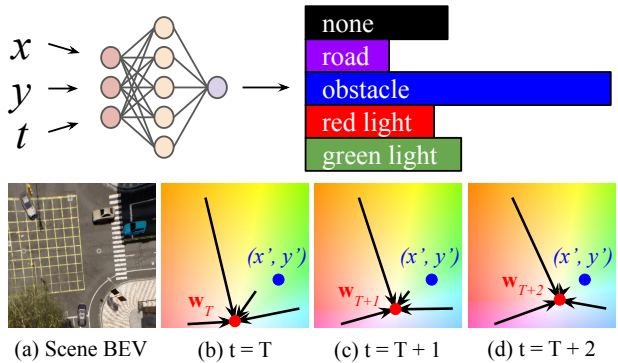
**NEAT:**

- Multi-task learning
- Single training stage
- Selective attention

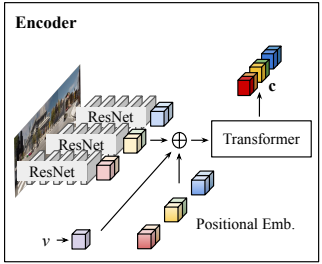


# Representation

- **Implicit** BEV semantics and waypoints
- Arbitrary spatial and temporal **resolution**
- Small **memory** footprint
- Can make use of sparse supervision signals



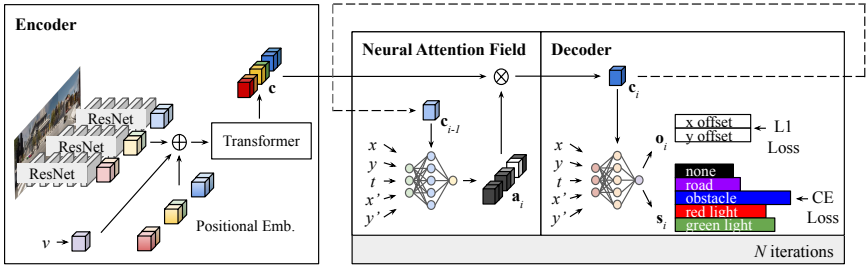
# Architecture: Encoder




- A **ResNet backbone** yields 64 patches/tokens per image
- Patch features are combined with **velocity** and **positional encoding**
- A **Transformer** integrates contextual cues through self-attention

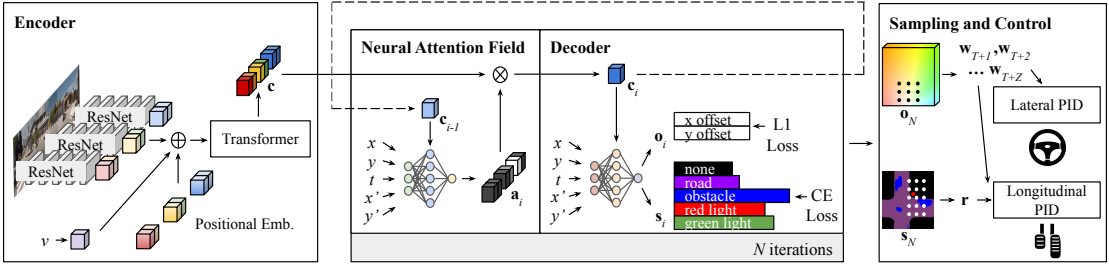
K. Chitta, A. Prakash and A. Geiger: NEAT: Neural Attention Fields for End-to-End Autonomous Driving. ICCV, 2021.

# Architecture: NEAT and Decoder




- The **encoder** is not informed by the query $(x, y, t)$ or target $(x', y')$ location
- Instead, **NEAT** identifies relevant patch features recurrently $(\dim(\mathbf{c}_i) \ll \dim(\mathbf{c}))$
- The **decoder** predicts the semantic class and waypoint offset from $\mathbf{c}_N$

# Architecture: Sampling and Control




- At test time, we decode **waypoints** and a **red light indicator** from our model
- To extract both, we uniformly sample a sparse grid and **query the decoder**
- Grid pooling provides robustness at inference time; control via **PID controllers**

# Experiments

**Sensors**

- 3 RGB cameras: (-60°, 0° and 60°)
- No LiDAR

**Evaluation**

- 42 mixed-length routes (200-3000m) from 6 different towns
- 7 weather conditions (Clear, Cloudy, Wet, MidRain, WetCloudy, HardRain, SoftRain)
- 6 daylight conditions (Night, Twilight, Dawn, Morning, Noon, Sunset)
- High density of dynamic agents and complex scenarios

# Results

| Method | Aux. Sup. | DS ↑ |
|---|---|---|
| CILRS | Velocity | $22.97 \pm 0.90$ |
| Learning by Cheating | BEV Sem | $29.07 \pm 0.67$ |
| Waypoint Prediction | None | $51.25 \pm 0.17$ |
| | 2D Sem | $57.95 \pm 2.76$ |
| | BEV Sem | $60.62 \pm 2.33$ |
| NEAT | BEV Sem | $\mathbf{65.10 \pm 1.75}$ |

- CILRS / LBC use discrete / image-based target locations
- WP and NEAT use **BEV-based target locations**

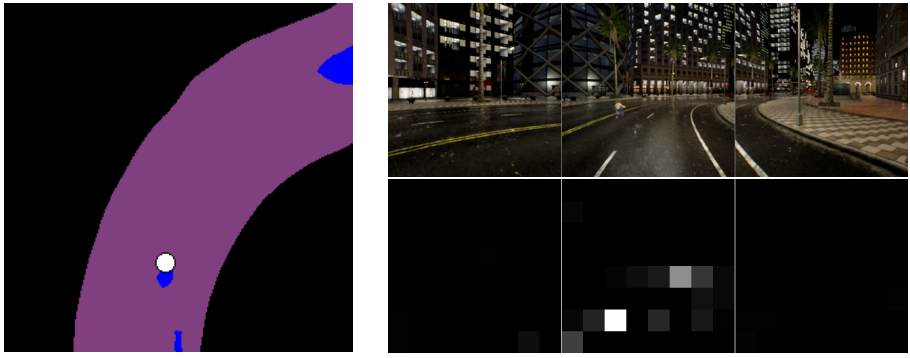
# CARLA Leaderboard

| Method | Driving Score ↑ |
|---|---|
| WOR | **31.37** |
| MaRLn | 24.98 |
| NEAT | 21.83 |
| Waypoint Prediction | 19.38 |
| TransFuser | 16.93 |
| Learning by Cheating | 8.94 |
| CILRS | 5.37 |

- Sometimes too **cautious** and doesn't complete route in time
- But **safest driving** and trained with 1-2 orders **less data** than WOR/MaRLn

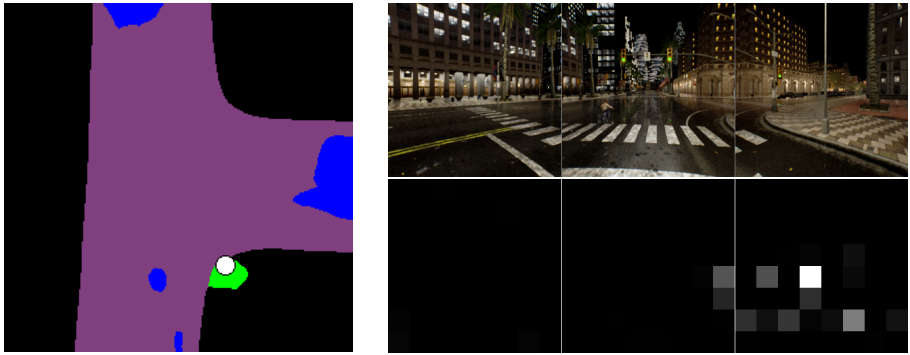
K. Chitta, A. Prakash and A. Geiger: NEAT: Neural Attention Fields for End-to-End Autonomous Driving. ICCV, 2021.

# Attention Visualizations



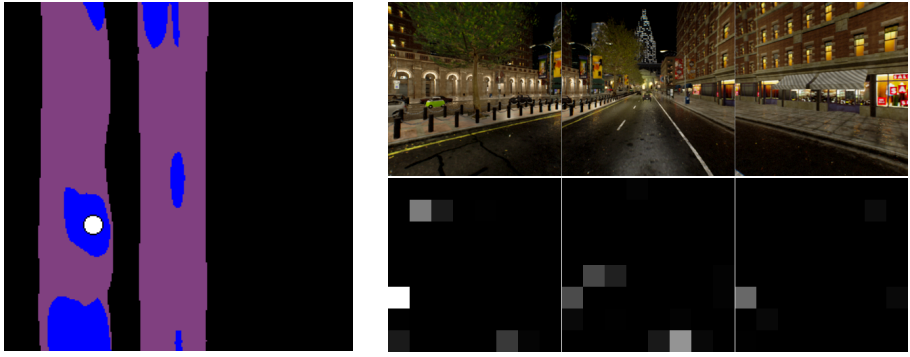
- White circle = query location $(x, y)$
- NEAT consistently **attends** to the region corresponding to the **object of interest**

# Attention Visualizations



- White circle = query location $(x, y)$
- NEAT consistently **attends** to the region corresponding to the **object of interest**

# Attention Visualizations



- White circle = query location $(x, y)$
- NEAT consistently **attends** to the region corresponding to the **object of interest**

# Attention Visualizations

- White circle = query location $(x, y)$
- NEAT consistently **attends** to the region corresponding to the **object of interest**

# Video


K. Chitta, A. Prakash and A. Geiger: NEAT: Neural Attention Fields for End-to-End Autonomous Driving. ICCV, 2021.

# Summary

**Conclusions:**

- Global contextual reasoning is crucial in complex urban scenarios
- Attention is effective in aggregating information from multiple modalities
- This allows for navigation of challenging intersection scenarios
- Joint BEV semantic prediction and trajectory planning leads to safer driving
- An implicit intermediate representation can efficiently solve this task

**Code and Models:**

- www.github.com/autonomousvision/transfuser
- www.github.com/autonomousvision/neat

K. Chitta, A. Prakash and A. Geiger: NEAT: Neural Attention Fields for End-to-End Autonomous Driving. ICCV, 2021.

# Thank you!

http://autonomousvision.github.io



Federal Ministry
of Education
and Research

Federal Ministry
for Economic Affairs
and Energy


Microsoft
Research