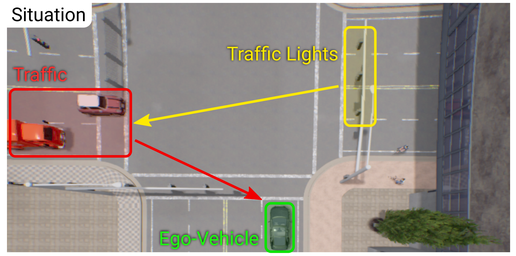
Multi-Modal Fusion Transformer for End-to-End Autonomous Driving

Aditya Prakash*[1] Kashyap Chitta*[1,2] Andreas Geiger[1,2]

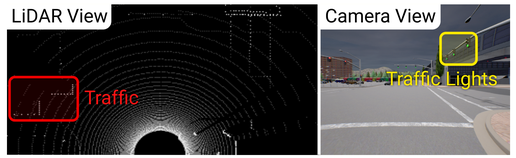
[1]Max Planck Institute for Intelligent Systems, Tübingen [2]University of Tübingen

Motivation

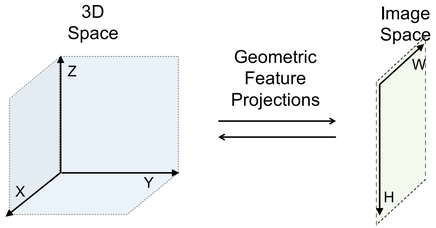
Global context is essential in complex scenarios, e.g. relation between traffic lights and vehicles



Fusion-based methods can **capture geometric and semantic information of the 3D scene** using multiple sensors (e.g. camera and LiDAR)

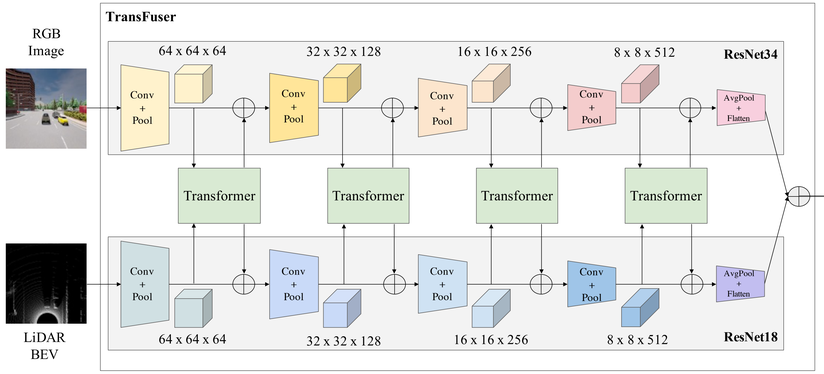


Prior work focus on using **geometry-based feature projections** to aggregate features from a **local region in projected 3D or image space**.



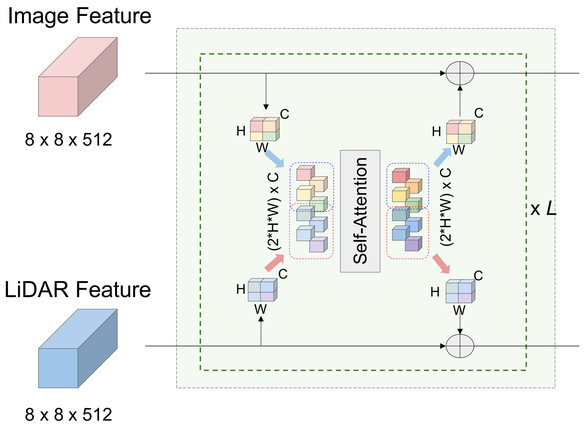
Our Method

Our key idea is to use **attention-based feature fusion** to incorporate global context of the 3D scene.

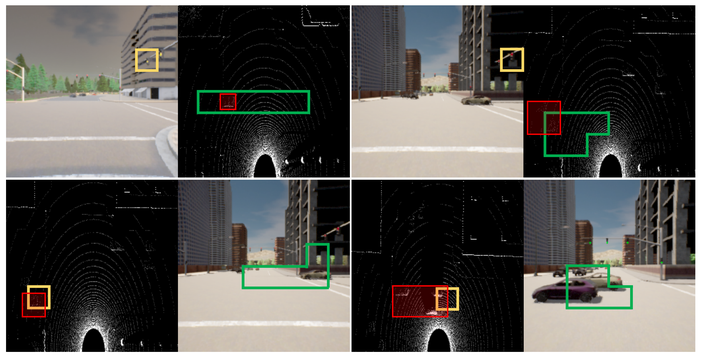


$$\mathcal{L} = \sum_{t=1}^{T} ||\mathbf{w}_t - \mathbf{w}_t^{gt}||_1$$

Attention-based Feature Fusion



Attention Map Visualizations

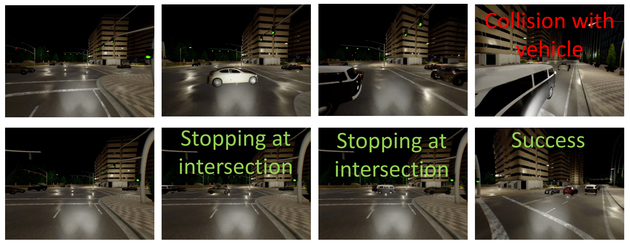


yellow: source token, green: top-5 attended tokens, red: vehicles in LiDAR point cloud

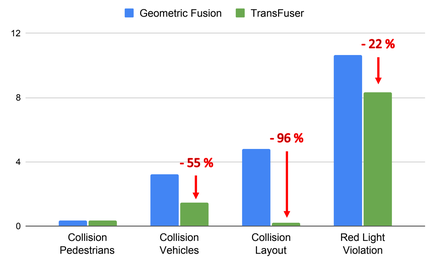
Driving Results

Generalization to New Town (GF: Geometric Fusion, TF: TransFuser)

Generalization to New Weather



Infraction Analysis



TransFuser focuses on vehicles and traffic lights at intersections and can safely navigate difficult scenarios.

aditya.prakash@tue.mpg.de