



TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving

11th Seminar in 2023, Paper Review

SSDC: Samsung Software Developer Community
Korea Vision & Robotics
Hongtae Kim
2023.05.27

Introduction

- 2021 CVPR, 'Multi-Modal Fusion Transformer for End-to-End Autonomous Driving'이라는 타이틀로 개제
- 2022 PAMI, 'TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving'이라는 타이틀로 개제

2021년



IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)
Impact Factor 22.39

2022년



IEEE Transactions on Pattern Analysis and Machine Intelligence
Impact Factor 24.31

Introduction

- Andreas Geiger : 컴퓨터 비전 및 자율주행 인지 연구자, KITTI Dataset을 만든 사람
- KITTI Dataset : 자율주행 분야의 ImageNet과 같은 Dataset



Andreas Geiger
 Professor of Computer Science, University of Tübingen and Tübingen AI Center
 Verified email at uni-tuebingen.de - [Homepage](#)
 computer vision machine learning robotics scene understanding

[FOLLOW](#)

[GET MY OWN PROFILE](#)

CITED BY YEAR

TITLE	CITED BY	YEAR
Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite	11468	2012
A. Geiger, P. Lenz, R. Urtasun Computer Vision and Pattern Recognition (CVPR), (Providence, USA)		
Vision meets robotics: The KITTI dataset	7241	2013
A. Geiger, P. Lenz, C. Stiller, R. Urtasun The International Journal of Robotics Research 32 (11), 1231-1237		
Object Scene Flow for Autonomous Vehicles	1960	2015
M. Menze, A. Geiger Proceedings of the IEEE Conference on Computer Vision and Pattern ...		
Occupancy networks: Learning 3d reconstruction in function space	1694	2019
L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, A. Geiger Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern ...		
Octnet: Learning deep 3d representations at high resolutions	1435	2017
G. Riegler, A. Osman Ulusoy, A. Geiger Proceedings of the IEEE Conference on Computer Vision and Pattern ...		
Stereoscan: Dense 3d reconstruction in real-time	1296	2011
A. Geiger, J. Ziegler, C. Stiller Intelligent Vehicles Symposium (IV), 2011 IEEE, 963-968		
Which Training Methods for GANs do actually Converge?	1197	2018
L. Mescheder, A. Geiger, S. Nowozin International Conference on Machine Learning, 3478-3487		

Cited by

Cited by	VIEW ALL
All	38242
Since 2018	121

Citations

Citations	45487
h-index	75
i10-index	128

Public access

Public access	VIEW ALL
0 articles	45 articles
not available	available

Based on funding mandates

The KITTI Vision Benchmark Suite
 A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago



[home](#) [setup](#) [stereo](#) [flow](#) [sceneflow](#) [depth](#) [odometry](#) [object](#) [tracking](#) [road](#) [semantics](#) [raw data](#) [submit results](#)

A. Geiger | P. Lenz | C. Stiller | R. Urtasun
Log in

New Dataset



KITTI-360
<http://www.cvlibs.net/datasets/kitti-360>



360° Velodyne Laserscanner
 Stereo Camera Rig
 GPS
 IMU
 Magnetometer
 Barometer
 Accelerometer
 Gyroscope

Welcome to the KITTI Vision Benchmark Suite!

We take advantage of our [autonomous driving platform Annieway](#) to develop novel challenging real-world computer vision benchmarks. Our tasks of interest are: stereo, optical flow, visual odometry, 3D object detection and 3D tracking. For this purpose, we equipped a standard station wagon with two high-resolution color and grayscale video cameras. Accurate ground truth is provided for depth, camera extrinsics and GPS location systems. Observations are generated by driving around the mid-size city of Middlebury in rural areas and on highways. Up to 15 cars and 30 pedestrians are visible per image. Besides providing all data in raw format, we extract benchmarks for each task. For each of our benchmarks, we also provide an evaluation metric and this evaluation website. Preliminary experiments show that methods ranking high on established benchmarks such as Middlebury perform below average when being moved outside the laboratory to the real world. Our goal is to reduce this bias and complement existing benchmarks by providing real-world benchmarks with novel difficulties to the community.

[Share](#)

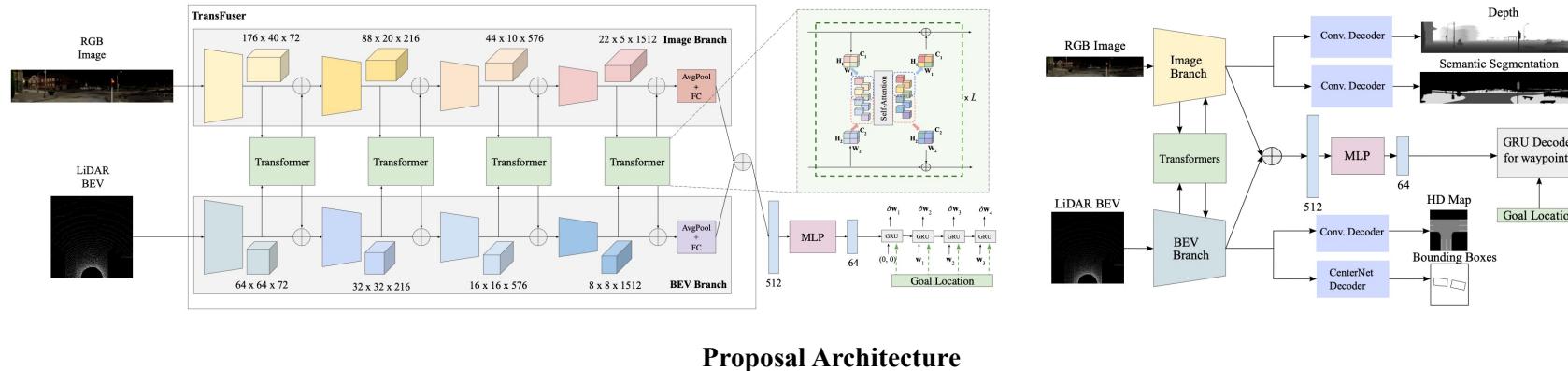
To get started, grab a cup of your favorite beverage and watch our video trailer (5 minutes):

Google Scholar profiles

KITTI Dataset

What is this paper for?

- Sensor Fusion 방식의 모방학습을 적용한 기존 End-to-end 자율주행 시스템은 동적 에이전트의 밀도가 높은 복잡한 도심 속 주행 시나리오에서 성능이 좋지 않았음
- 이를 극복하기 위해 Transformer의 Attention Mechanism을 이용하여 Multimodal(Camera, LiDAR) Sensor Fusion을 하는 ‘TransFuser’ 아키텍처를 제안
- 그 결과 복잡한 도심 속 주행에서 Global Context를 더 잘 이해하게 되어 높은 성능을 보여줌



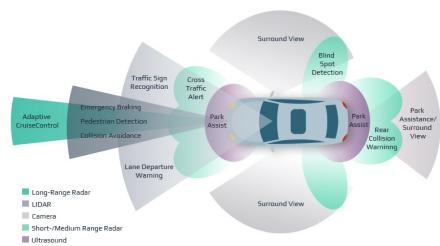
Keyword



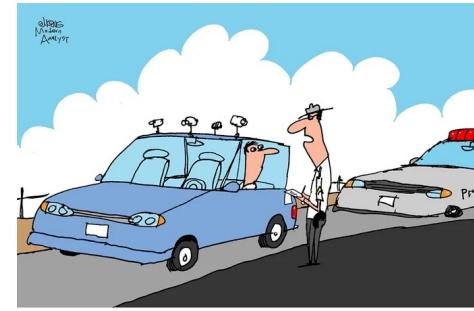
모방학습
Imitation Learning



트랜스포머
Transformer



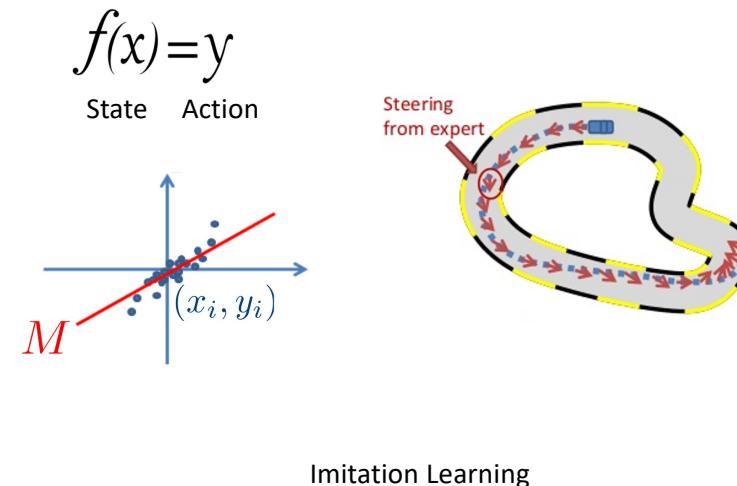
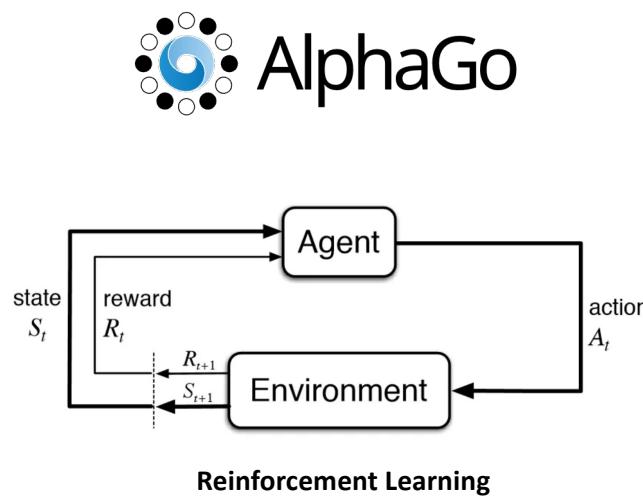
센서 퓨전
Sensor Fusion



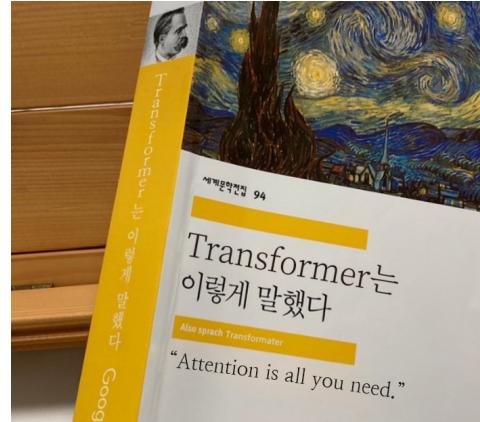
자율주행
Autonomous Driving

Imitation Learning

- 강화학습 : 현재 state에서 어떤 action을 취하는 것이 가장 좋은 결정인가에 대한 policy를 학습하는 것
- 모방학습 : expert의 policy를 모방하는 것
- “The goal of imitation Learning is to train a policy to mimic the expert’s demonstrations”



Transformer



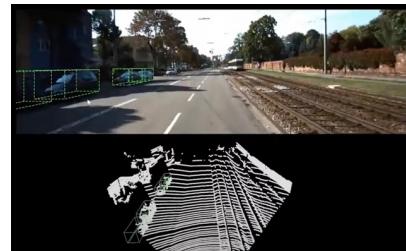
Self-Attention Mechanisms

Sensor Fusion

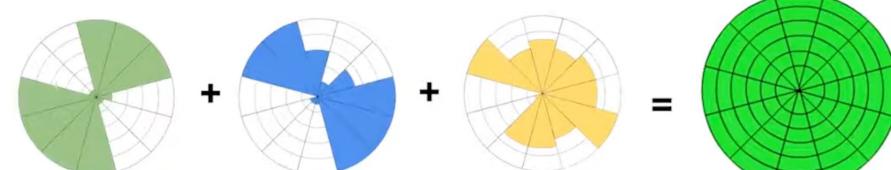
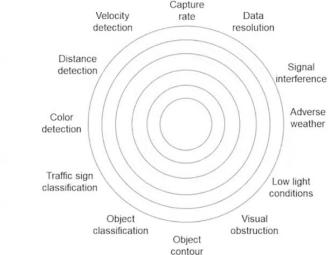
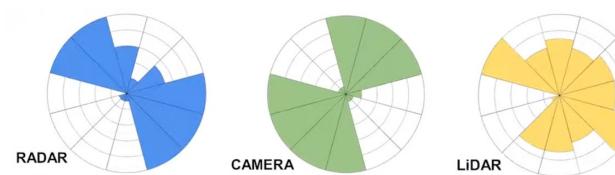
- 카메라, LiDAR, 레이더 등의 센서들은 서로 상호보완적인 관계를 가지고 있음
- 따라서 Sensor Fusion을 통해 주행 환경에 대한 정보를 더 잘 얻을 수 있음



자율주행에 사용하는 주요 센서들



Sensor Fusion 예시



상호보완적인 센서들

기존 연구 한계

Limitation

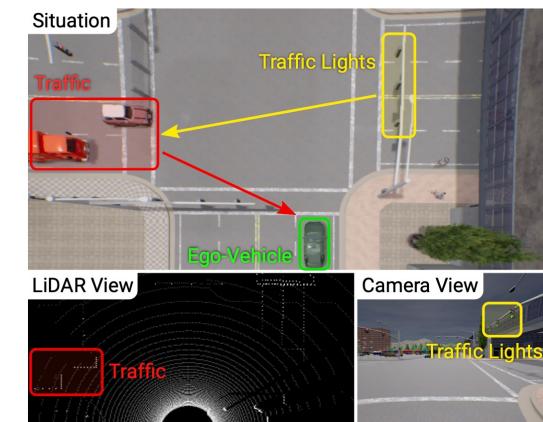
Locality Assumption

- 성능 향상을 극대화하기 위해 어떤 종류의 융합 메커니즘을 사용해야 하는가?
- Sensor Fusion의 선행 연구에서는 주로 2D 및 3D Object detection, motion forecasting, depth estimation 등 state representation을 인지하는 측면에 초점을 둠
- 이러한 인지적 측면에서는 global context를 이해하는 것이 목표가 아님, 도로 위 객체들을 잘 인지하는 것이 목표임
- “Information is typically aggregated from a local neighborhood around each feature in the projected 2D or 3D space”
- 이러한 locality assumption 아키텍처 디자인은 복잡한 도심 속 환경에서 좋지 않은 성능을 보임



Figure

기존 Sensor Fusion의 문제점

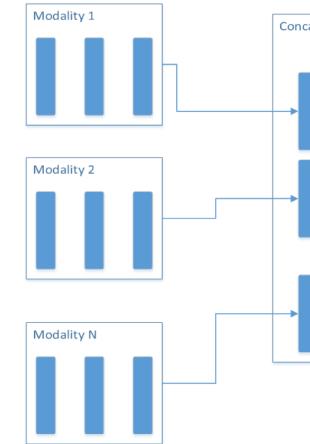
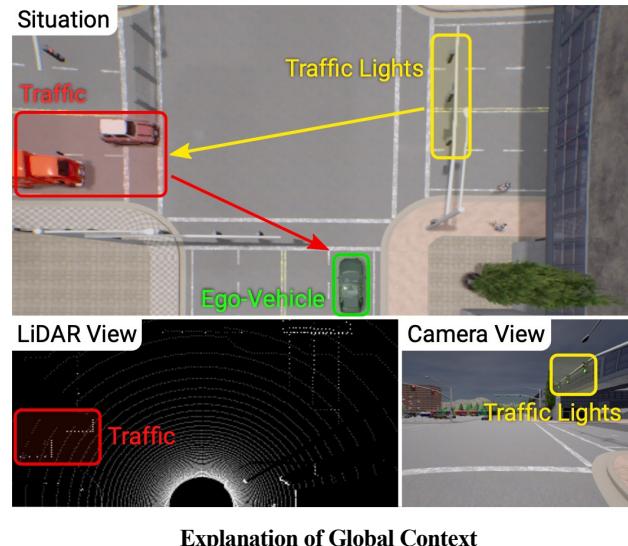


제안된 아키텍쳐

Limitation – Global Context

Global Context?

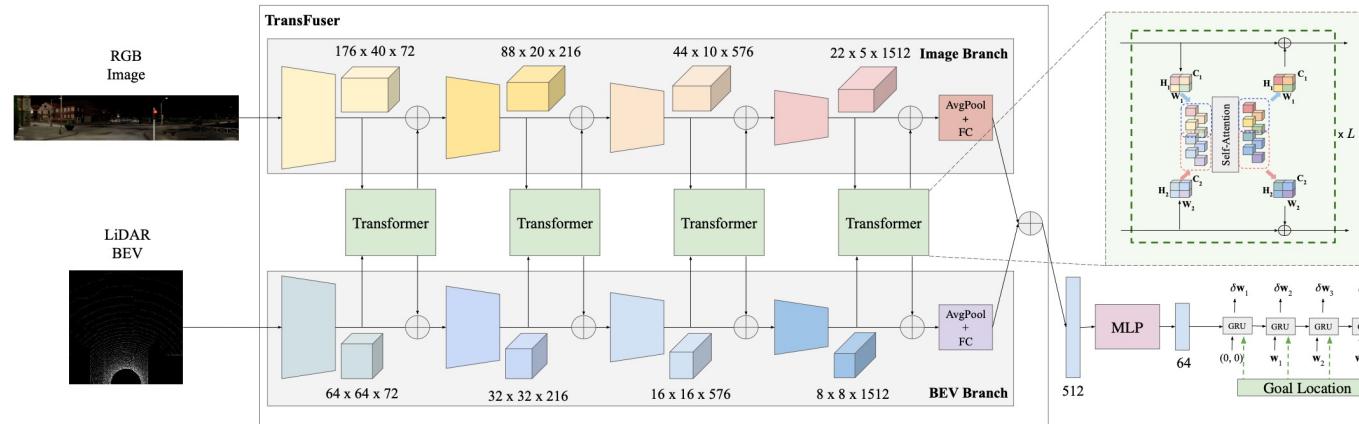
- 여러 차선이 있는 교차로에서 정보를 처리할 때 근처에 있는 동적 에이전트와 멀리 있는 신호등 간의 상호 작용을 고려해야 함
 - 1) LiDAR에는 보이지 않는 신호등 정보와 차량 정보를 어떻게 통합할 것인가? (Context)
 - 2) Camera로는 탐지할 수 없지만, LiDAR로는 탐지할 수 있는 정보를 어떻게 융합할 것인가? (Invisible)
- Deep Convolutional Network는 하나의 Modal 안에서는 Global context를 캡쳐하는 것이 가능
- 하지만 두 가지 Modal이 있을 때, Modal 간의 Feature를 상호작용하면서 Global Context를 캡쳐하는 것은 어려움



* Sobh, Ibrahim, et al. "End-to-end multi-modal sensors fusion system for urban automated driving." (2018).

Limitation

- 이렇게 서로 다른 Modal간 Global context를 통합하지 못하는 한계점을 극복하기 위해서 Transformer의 Attention Mechanism을 사용
- 카메라와 라이다는 서로 상호 보완적인 성질을 가지고 있기 때문에, 이 두 Modal을 통합하는 데에 집중하였음
- 각 Input들은 Transformer를 사용하여 상호 연결된 두 개의 독립적인 convolutional encoder branches에 의해 처리됨
- 이러한 아키텍처를 **TransFuser**라고 한다
- TransFuser의 결과를 Autoregressive waypoint prediction 프레임워크에 통합하여 end-to-end 자율주행을 구현



제안된 아키텍쳐

Contribution

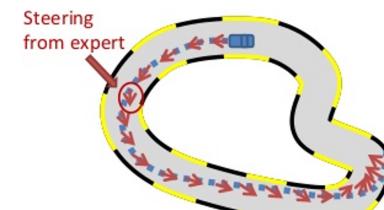
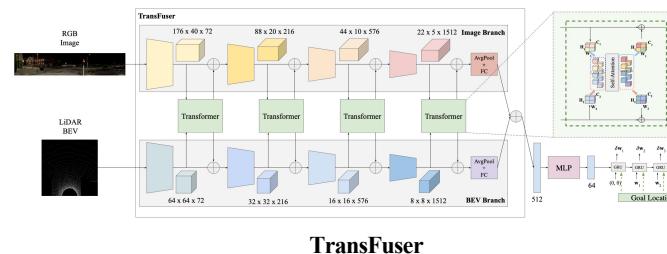
- 1) 기존 Sensor Fusion 접근법에 기반에 모방 학습 정책이 트래픽이 밀집된 까다로운 시나리오에서 좋지 않은 성능을 보인다는 것을 입증하는 새로운 평가 설정을 CARLA에서 설계함
- 2) 다양한 Input Modal의 Feature Extractor Layer에 Global Context와 Pairwise Interactions을 통합하기 위해 새로운 TransFuser (Multi-modal Fusion Transformer)를 제안합니다.
- 3) 제안된 평가 환경과 공식 CARLA Leaderboard에서 TransFuser를 사용하여 주행 성능을 보여주는 상세한 경험적 분석을 수행함.
 - 이 분석은 End-to-end 주행 모델의 현재 한계를 탐구함

4.4 Longest6 Benchmark

The CARLA simulator provides an official evaluation leaderboard consisting of 100 secret routes. However, teams using this platform are restricted to only 200 hours of evaluation time per month. A single evaluation takes over 100 hours, making the official leaderboard unsuitable for ablation studies or obtaining detailed statistics involving multiple evaluations of each model. Therefore, we propose the Longest6 Benchmark, which shares several similarities to the official leaderboard, but can be used for evaluation on local resources without computational budget restrictions.

The CARLA leaderboard repository provides a set of 76 routes as a starting point for training and evaluating agents. These routes were originally released along with the 2019 CARLA challenge. They span 6 towns and each of them is defined by a sequence of waypoints. However, there is a large imbalance in the number of routes per town, e.g.

Longest6 Benchmark



아키텍쳐

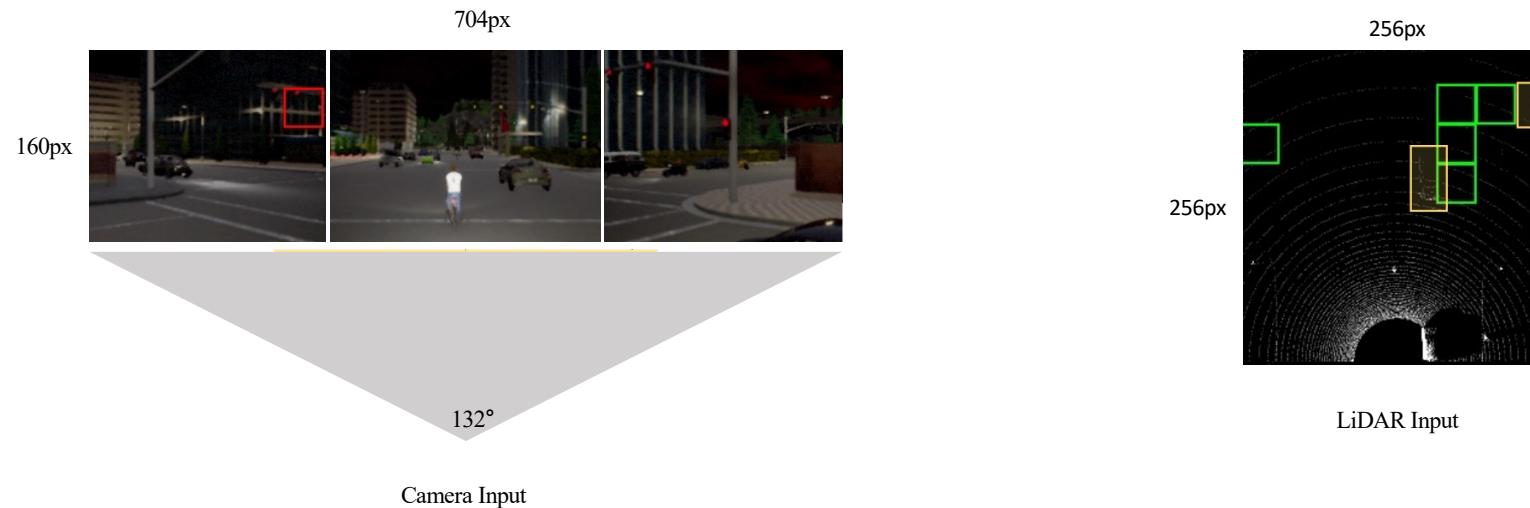
TransFuser – Input

Camera Input

- 세 대의 카메라가 사용되며, 한 대는 정면을 향하고 다른 카메라는 각각 왼쪽과 오른쪽으로 60° 각도를 유지함
- 각 카메라의 FOV(Field of View, 수평 시야각)은 120°임
- 이미지는 960 x 480픽셀의 해상도로 캡처한 다음 가장자리 방사형 왜곡을 제거하기 위해 320 x 160픽셀로 crop 함
- 그런 다음 각 카메라에서 얻은 320 x 160의 이미지들을 병합하여, 704 x 160픽셀의 해상도와 132°의 확장된 FOV를 가진 이미지 생성

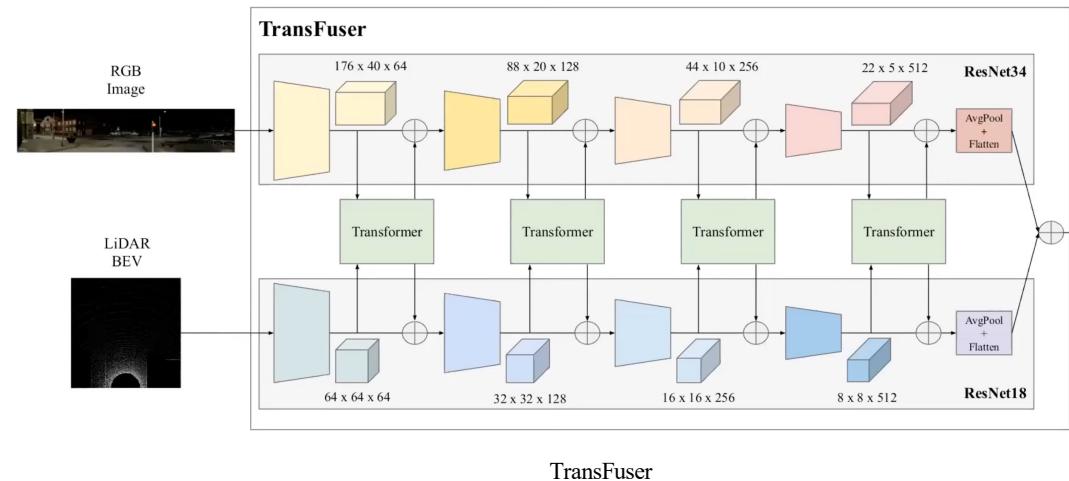
LiDAR Input

- 전방 32m, 양측 16m의 영역을 커버하여 32m x 32m의 BEV 그리드를 생성함
- 그리드를 각각 0.125m x 0.125m의 작은 블록으로 분할하면 256 x 256 픽셀의 해상도를 가지는 이미지가 됨



TransFuser – Multi-Modal Fusion Transformer

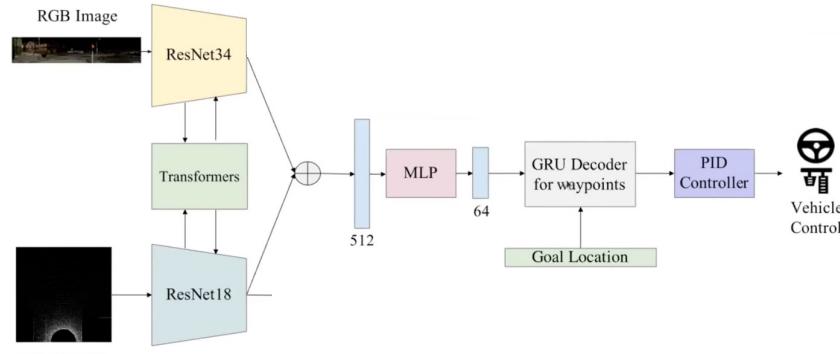
- 이미지와 LiDAR BEV 입력에 대한 convolutional feature extractors는 Input에 대해서 서로 다른 해상도에서 인코딩을 하게 됨
- 따라서 여러 Scale에서 융합을 하게 됨
- Transformer의 출력 Feature map은 각 Branch의 Feature map과의 element-wise summation을 통해 다시 피드백 됨
- 높은 해상도에서 Feature map을 처리하는 것은 계산 비용이 큼
- 따라서 Average Pooling을 통해 해상도를 Downampling 후 Transformer의 입력으로 전달
- 최종적으로 $22 \times 5 \times 512$, $8 \times 8 \times 512$ 크기의 Feature map을 얻음
- Avg Pooling + FC를 거쳐 512 차원의 특징 벡터를 element-wise summation을 통해 결합
- 이 512차원 Feature vector는 3D 장면의 Global context를 인코딩하여 환경의 압축적인 표현을 구성
- 그런 다음 다음에서 설명하는 waypoint prediction network이 정보를 제공



핵심 아이디어는 Transformer의 self-attention 메커니즘을 활용하여 이미지와 LiDAR 양식이 상호 보완적인 특성을 가지고 있다는 점을 감안하여 global context를 통합하는 것

Waypoint Prediction Network & Controller

- 512차원 feature vector를 연산 효율성을 위해 64차원으로 줄인 후 GRU를 사용하여 구현된 Auto-regressive Waypoint Network에 전달함
- 단일 레이어 GRU를 사용하여 차량의 현재 좌표 프레임에서 $T = 4$ 개의 미래 시간에 대해 waypoint를 예측함
- 종방향, 횡방향 제어를 위해 두 개의 PID 컨트롤러를 사용하여 예측된 Waypoint에서 스티어, 스로틀 및 브레이크 값을 얻음
- Creeping: 차량이 장시간(55초, 적색 신호에서 예상 대기 시간보다 높게 설정) 움직이지 않는 경우, 짧은 시간(1.5초) 동안 PID 컨트롤러의 목표 속도를 4m/s로 설정하여 차량이 앞으로 전진하도록 함
- Safety Heuristic: creeping 동작만으로는 안전하지 않을 수 있음. 예를 들어, 에이전트가 앞으로 나아가는 것이 충돌로 이어질 수 있는 교통 체증 상황에서는 creeping 동작이 안전하지 않을 수 있음. 이를 방지하기 위해 차량 앞의 작은 직사각형 영역에 라이더가 닿으면 creeping 동작을 덮어쓰는 동작을 함.



Waypoint Prediction Network & Controller

Metrics

- CARLA 리더보드 평가. 공식 평가 서버의 100개 비밀 경로에 대한 DS, RC, IS
- DS(Driving Score) : 위반 가중치를 이용하여 경로 완료의 가중 평균
- RC(Route Completion) : 차량이 경로 i에서 완료한 경로 거리의 백분율(R_i), N개의 경로를 평균한 값
- IS(Infraction Scroe) : 차량이 경로 중 발생한 모든 위반 사례 j, 보행자 충돌 시 0.5, 차량과 충동 시 0.6, 정지선 위반 0.65, 신호 위반 0.7

$$DS = \frac{1}{N} \sum_i^N R_i P_i$$

$$RC = \frac{1}{N} \sum_i^N R_i$$

$$IS = \prod_j^{Ped, Veh, Stat, Red} (p^j)^{\# \text{ infractions}^j}$$

Method	LiDAR?	Map?	DS ↑	RC ↑	IS ↑
CaRINA [5]	✓	-	4.56	23.80	0.41
CIRLS [6]	-	-	5.37	14.40	0.55
LBC [7]	-	-	8.94	17.54	0.73
CaRINA [5]	✓	✓	15.55	40.63	0.47
Pilot [8]	✓	✓	16.70	48.63	0.50
TF CVPR [9]	✓	-	16.93	51.82	0.42
AIM-MT [10]	-	-	19.38	67.02	0.39
NEAT [10]	-	-	21.83	41.71	0.65
MaRLn [11]	-	-	24.98	46.97	0.52
Late Fusion	✓	-	26.07	64.67	0.47
WOR [12]	-	-	31.37	57.65	0.56
TF+ [13]	✓	-	34.58	69.84	0.56
GRIAD [14]	-	-	36.79	61.86	0.60
Geometric Fusion	✓	-	41.70	87.85	0.47
Latent TF (Ours)	-	-	45.20	66.31	0.72
TF (Ours)	✓	-	61.18	86.69	0.71
LAV [15]	✓	-	61.85	94.46	0.64

Attention Map

