BEV Map Based Perception for Autonomous Driving

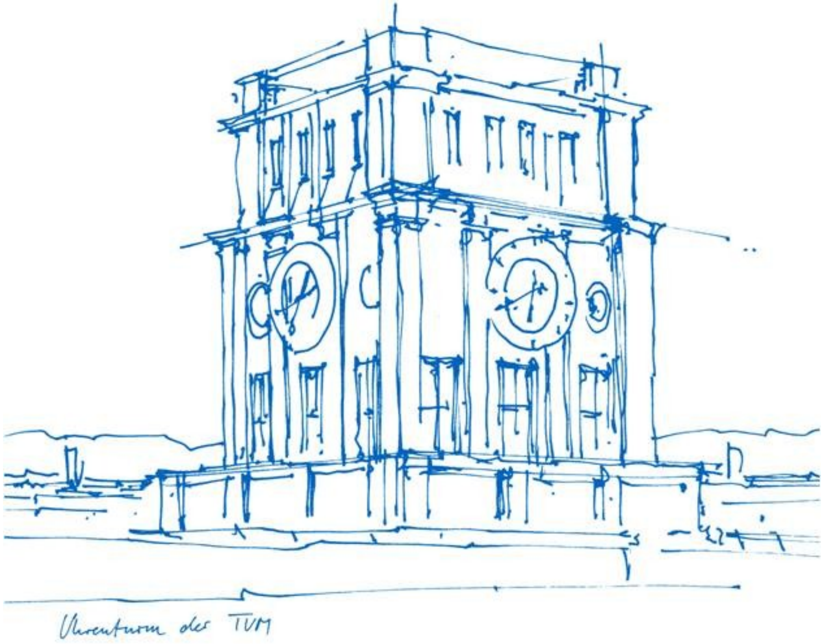
Yu Wu

Technical University of Munich

TUM School of CIT

Smart Robotics Lab

Munich, 16. January 2024



CONTENTS

1 Introduction

2 State of the art

3 Perception methods based on BEV

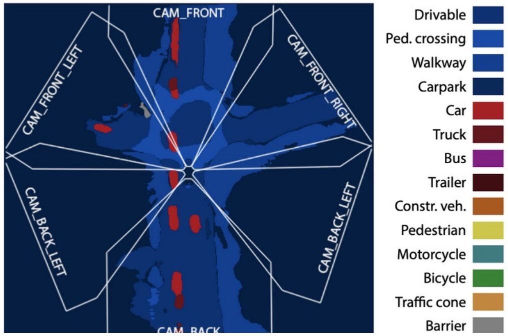
4 Methods comparison

5 Future work

PART 01

Introduction

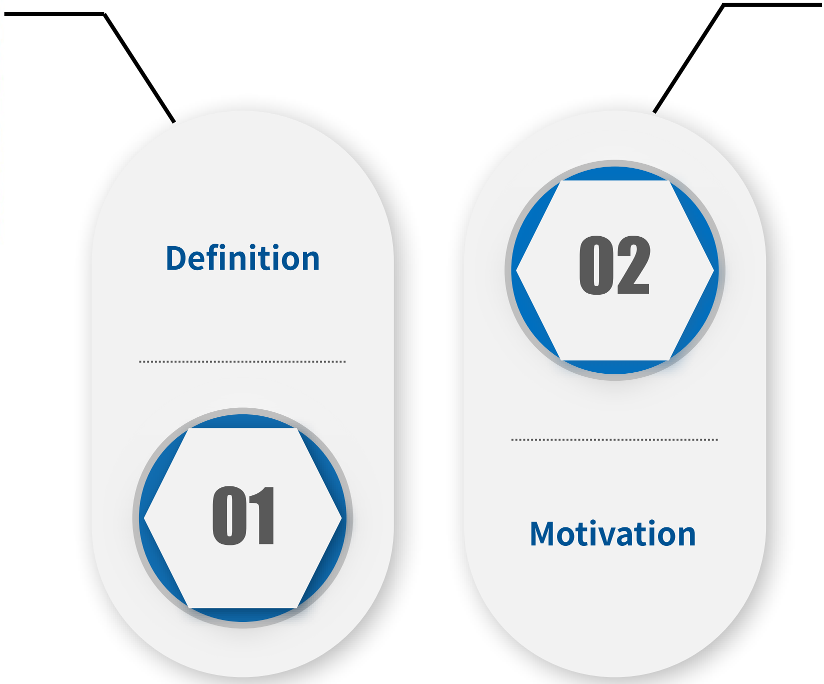
Introduction



BEV

↓

Bird's Eye View



- Strong perspective effect
- Targets obstruction

PART 02

State of the art

State of the art

2D-Perception

VS

3D-Perception

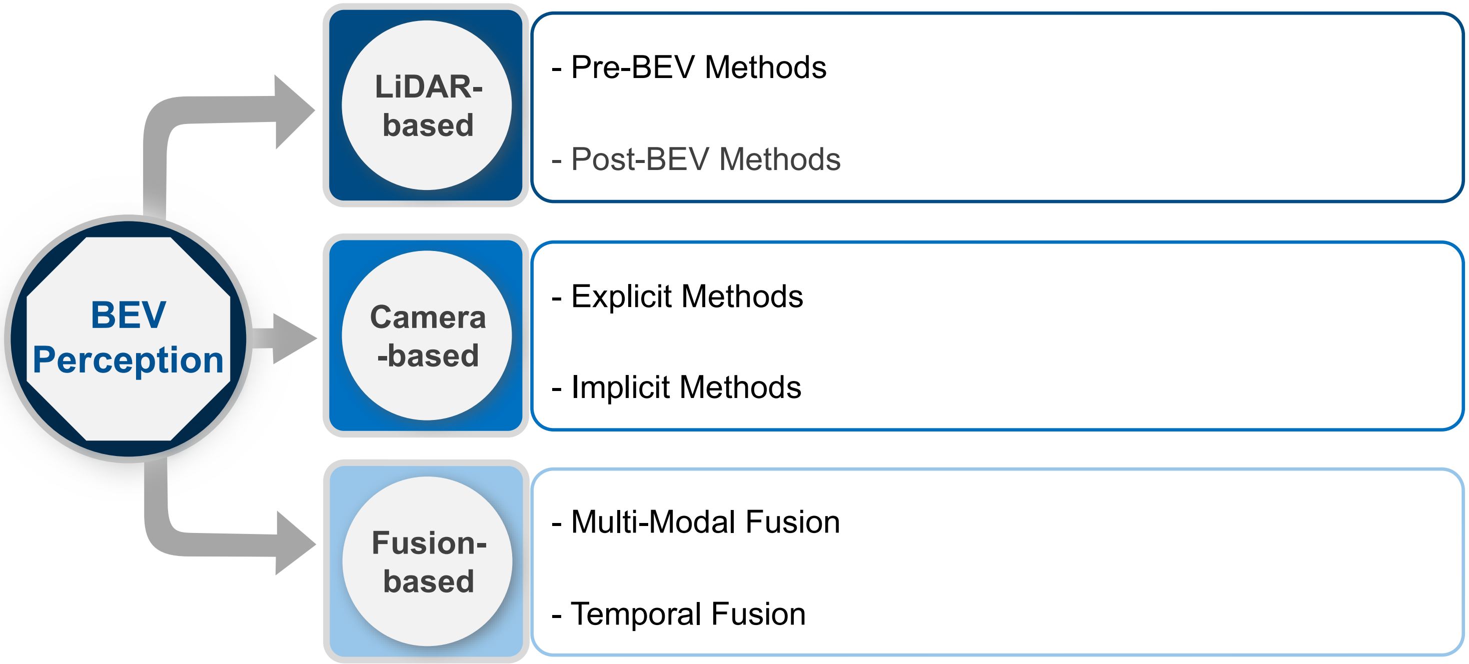
- No depth information
- Static
- Limited application scenarios

- Depth perception
- More comprehensive understanding of the environment
- Wider range of application scenarios

PART 03

Perception methods based on BEV

Perception methods based on BEV



BEV Perception

LiDAR-based

- Pre-BEV Methods
- Post-BEV Methods

Camera-based

- Explicit Methods
- Implicit Methods

Fusion-based

- Multi-Modal Fusion
- Temporal Fusion

Perception methods based on BEV

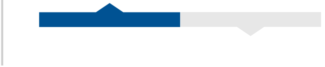
Pre-BEV Method:

Features extraction by voxels

BEV map

Region Proposal Network

Post-BEV Method:



Perspective transformation

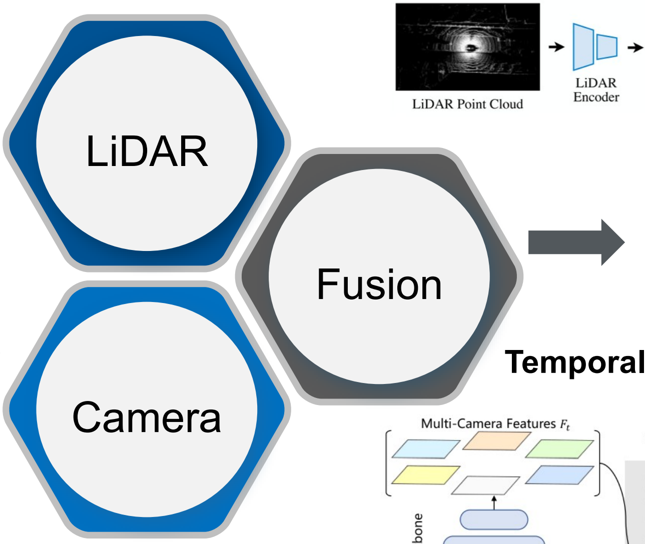
Features extraction under BEV map

Explicit:

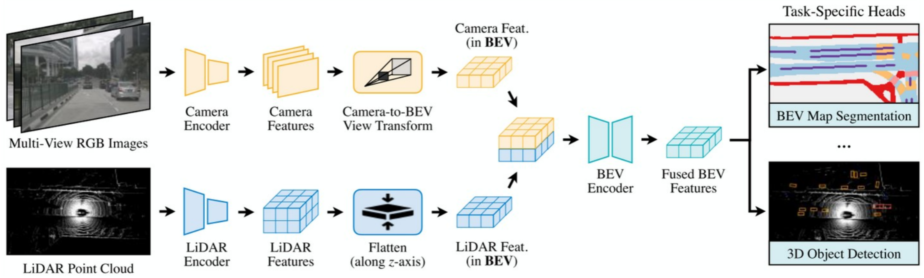
- Homography matrix

Implicit:

- Neural networks
- Loss = Euler distance between real and predicted 2D center points

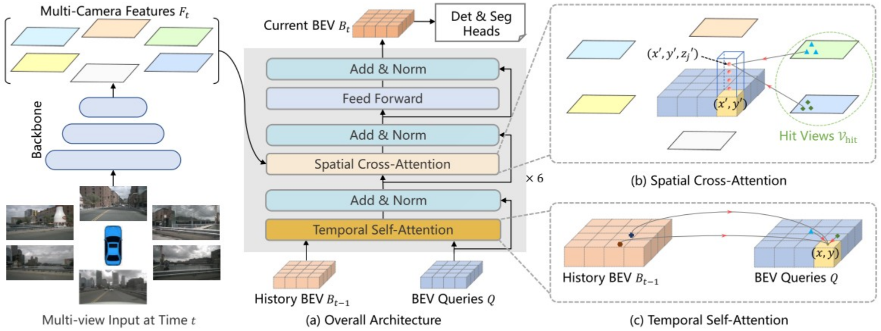


Multi-module fusion



[Structure of BEVFusion]

Temporal fusion



[Structure of BEVFormer]

PART 04

Methods comparison

Methods comparison

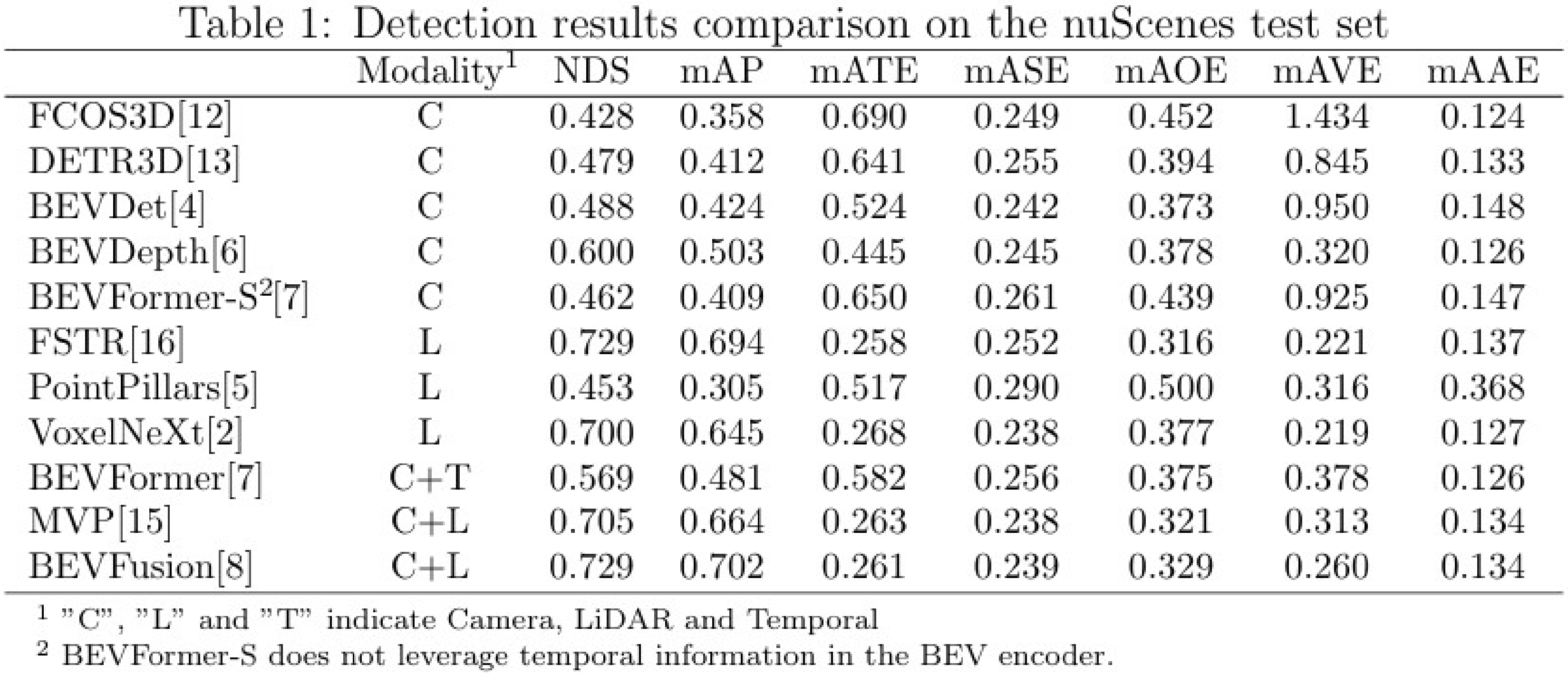


Table 1: Detection results comparison on the nuScenes test set

	Modality[1]	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
FCOS3D[12]	C	0.428	0.358	0.690	0.249	0.452	1.434	0.124
DETR3D[13]	C	0.479	0.412	0.641	0.255	0.394	0.845	0.133
BEVDet[4]	C	0.488	0.424	0.524	0.242	0.373	0.950	0.148
BEVDepth[6]	C	0.600	0.503	0.445	0.245	0.378	0.320	0.126
BEVFormer-S[2][7]	C	0.462	0.409	0.650	0.261	0.439	0.925	0.147
FSTR[16]	L	0.729	0.694	0.258	0.252	0.316	0.221	0.137
PointPillars[5]	L	0.453	0.305	0.517	0.290	0.500	0.316	0.368
VoxelNeXt[2]	L	0.700	0.645	0.268	0.238	0.377	0.219	0.127
BEVFormer[7]	C+T	0.569	0.481	0.582	0.256	0.375	0.378	0.126
MVP[15]	C+L	0.705	0.664	0.263	0.238	0.321	0.313	0.134
BEVFusion[8]	C+L	0.729	0.702	0.261	0.239	0.329	0.260	0.134

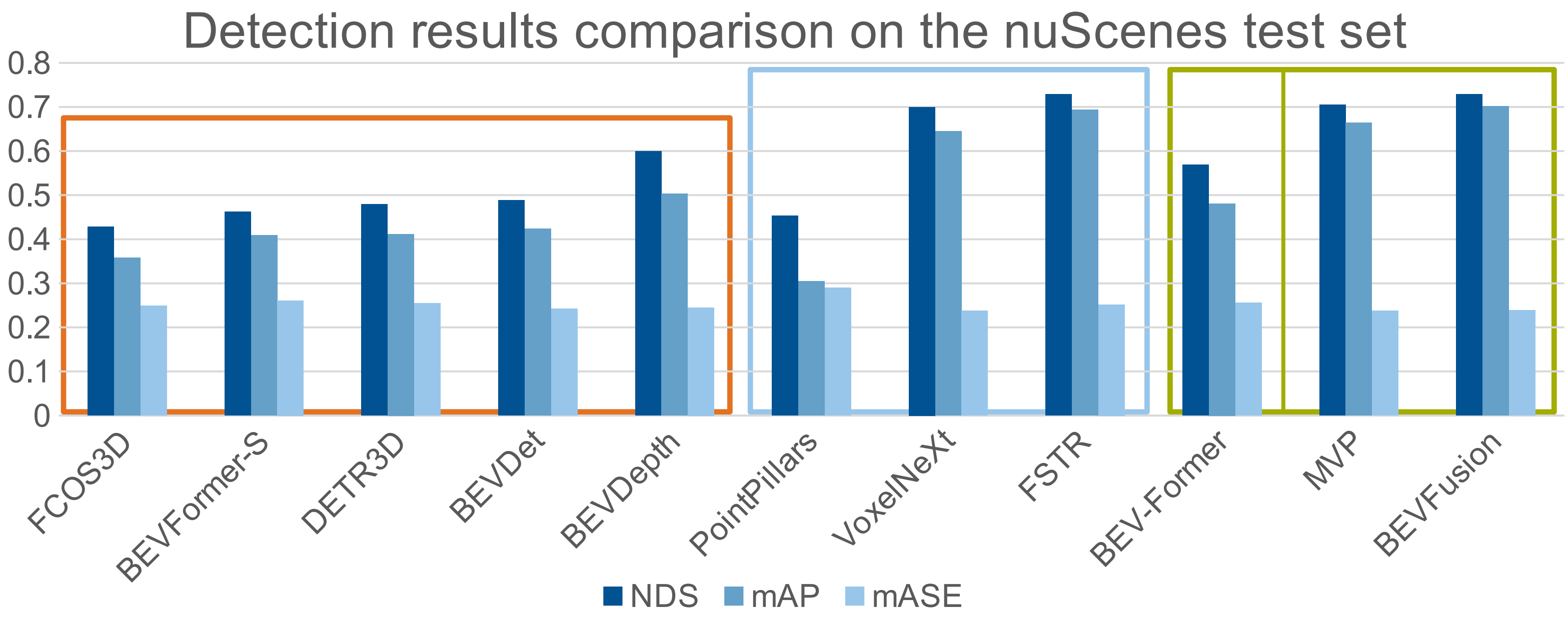
[1] "C", "L" and "T" indicate Camera, LiDAR and Temporal

[2] BEVFormer-S does not leverage temporal information in the BEV encoder.

- NDS: nuScenes Detection Score
- mAP: mean Average Precision
- mATE: mean Average Translation Error

- mASE: Average Scale Error
- mAOE: mean Average Orientation Error
- mAVE: mean Average Velocity Error
- mAAE: mean Average Attribute Error

Methods comparison



- mAP (mean Average Precision): Distance from the 2D center points under BEV map
- NDS (nuScenes detection score): Weighted average of all evaluation indicators in the table
- mASE (Average Scale Error): 1 – IoU (Intersection over Union) under perspective view

PART 05

Future work

Future work

For Camera-based

- Proposing new approaches to obtain the depth information

For LiDAR-based

- Focus on researching backbone networks

For Fusion-based

- Integration of different modules has greater potential
- Reducing redundant information in temporal fusion

Other directions

- exploring how to reduce information loss during perspective transformation
- constructing new BEV detection heads

Thank you for listening!

Yu Wu

Munich, 16. January 2024