Exploring Autonomous Driving

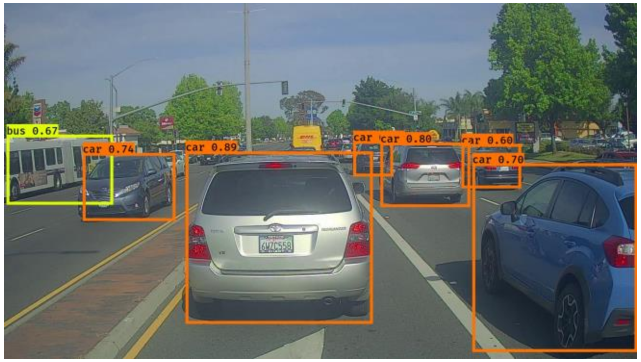
2023 Summer Seminar

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented by
Beoungwoo Kang

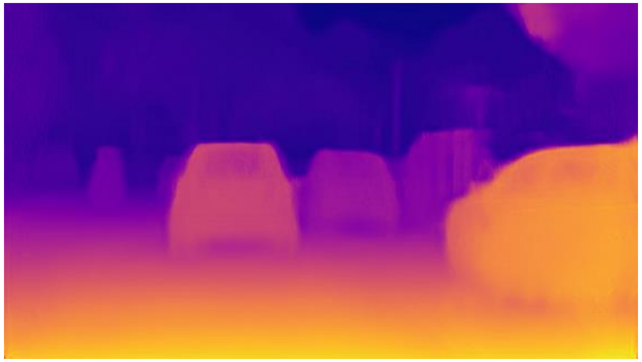
Abstract

- Do you believe autonomous driving is feasible?
  - I still believe that full autonomous driving is impossible

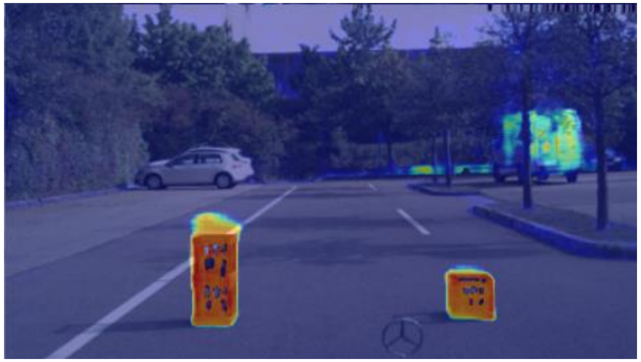


Object Detection

Segmentation



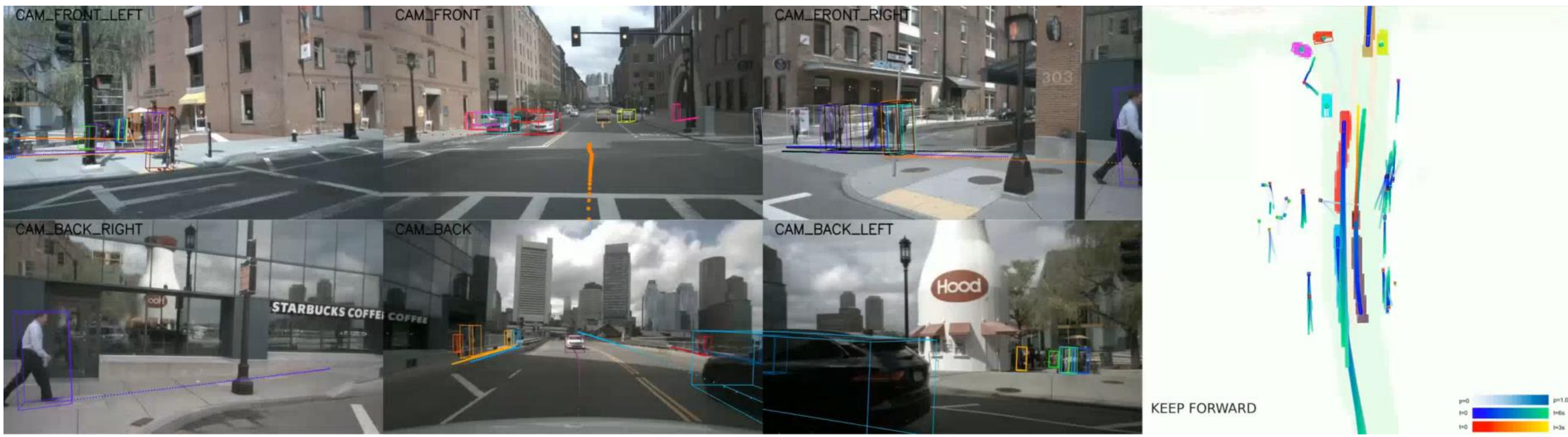
Depth Estimation



Anomaly Detection

Abstract

- [1] Planning-oriented Autonomous Driving [CVPR 2023 Best Paper]
  - Perception + Prediction + Planning

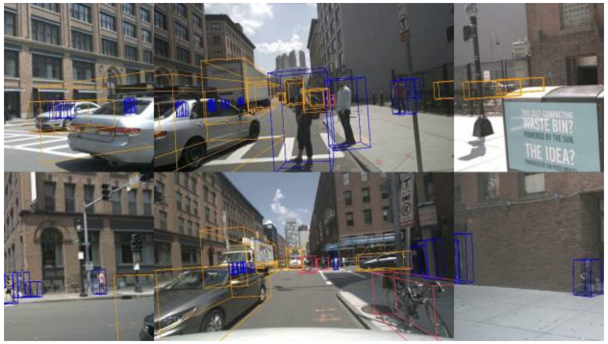


Outline

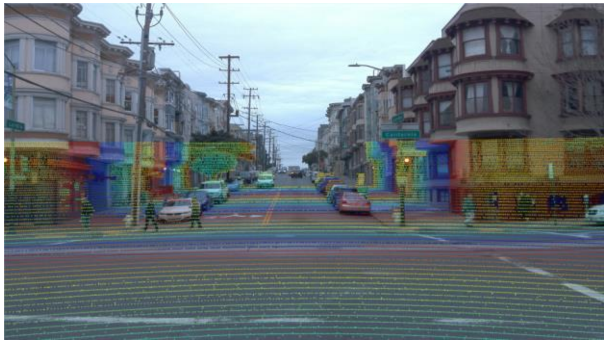
- Background
  - Datasets
  - Metrics
- [1] BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers [ECCV 2022]
  - BEVFormer architecture
  - Method
  - Experiments
- [2] Planning-oriented Autonomous Driving [CVPR 2023 Best Paper]
  - UniAD architecture
  - Method
  - Experiment

Background

- Datasets
  - nuScene dataset
    - Large-scale and diverse dataset for autonomous driving
    - Real-world scenes, images, lidar sweeps and 3D bounding boxes
  - Waymo open dataset
    - Also collected under various conditions and environments
    - Various weather conditions, from urban city center to landscapes



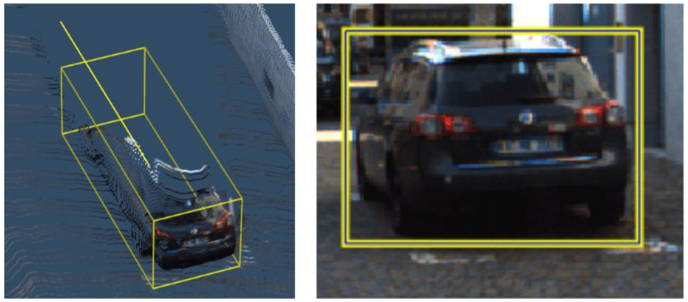
< nuScene dataset >



< Waymo open dataset >

Background

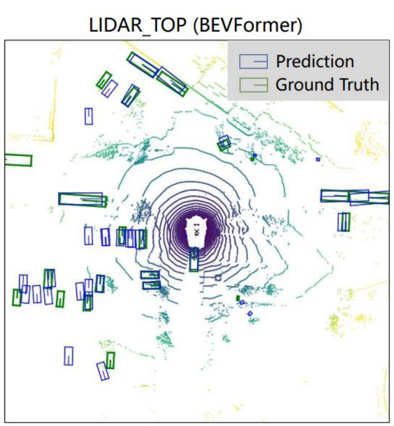
- Metrics
  - 3D detection
    - mAP (mean Average Precision)
    - mATE (mean Translation Error)
    - mASE (mean Scale Error)
    - mAOE (mean Orientation Error)
    - mAVE (mean Velocity Error)
    - mAAE (mean Attribute Error)
  - Autonomous driving
    - L2 error
    - Collision rate

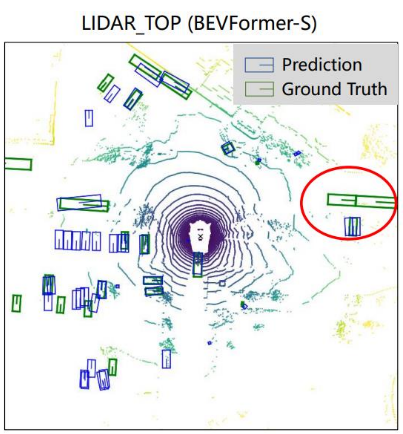


< 3D detection >

BEVFormer

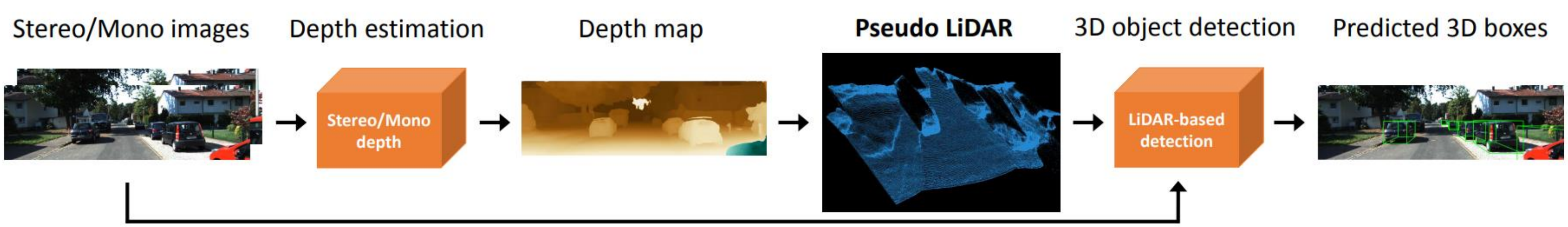
- BEV (Bird's Eye View)
  - Definition
    - Viewpoint from a high altitude, as if observed by a bird in flight
    - Representation of 3D space into 2D plane
  - Advantages
    - Cheaper than Lidar
    - Capable of detecting features that can only be seen in images





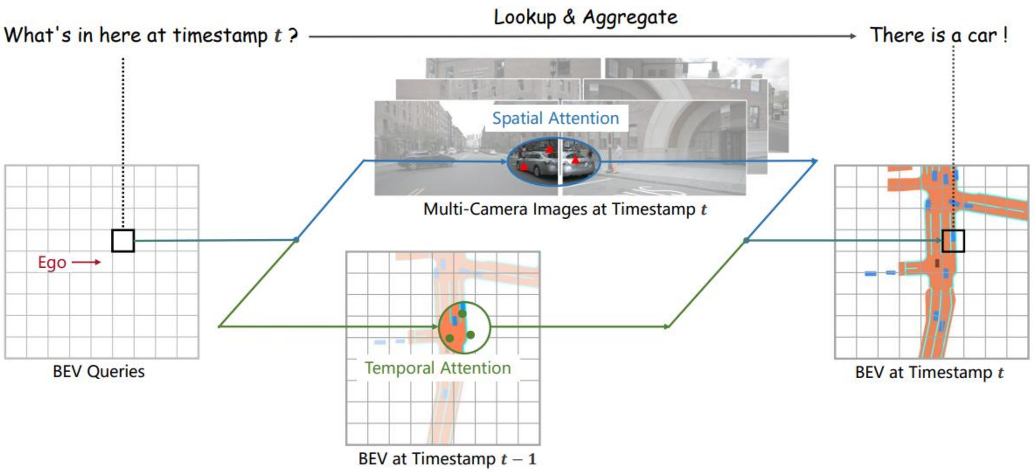
BEVFormer

- BEV (Bird's Eye View)
  - Conventional BEV framework
    - Creating BEV features based on depth information
      - Accuracy responds too sensitively to depth values or distributions
  - Proposal BEV framework
    - Designing depth-independent BEV feature to evade compounding errors
    - Connecting temporal and spatial information
      - Using sequential video data for perception



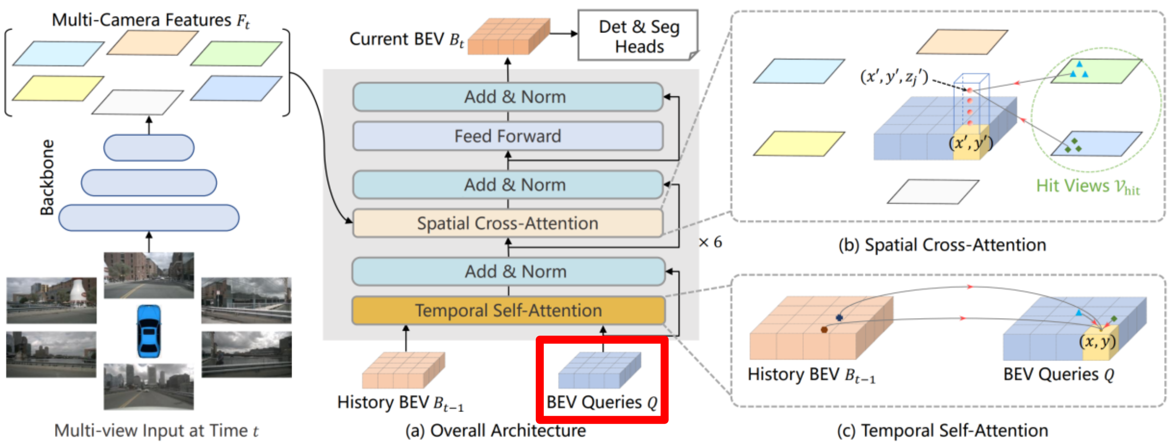
BEVFormer

- BEVFormer
  - BEV query
    - To better represent BEV features
  - Spatial cross-attention
    - To efficiently capture spatial dependencies across different views
  - Temporal self-attention
    - To incorporate temporal information from previous frames



BEVFormer

- BEV query
  - Set of learnable parameters
    - Positional embedding is added to the BEV queries
      - Capturing the spatial information of each grid cell in the BEV plane
    - Generating strong BEV features
      - Crucial for accurate 3D bounding box prediction



서강대학교 SOGANG UNIVERSITY

VDS LAB

BEVFormer

- Spatial cross-attention
  - Capturing spatial dependencies across different views
    - Set reference view point each $\mathcal{V}_{hit}$ for 2D multi-camera features
    - Cross-attention between the BEV features and the 6 multi views

\# of reference point

$$\mathrm{SCA}(Q_p, F_t) = \frac{1}{|\mathcal{V}_{\mathrm{hit}}|} \sum_{i \in \mathcal{V}_{\mathrm{hit}}} \sum_{j=1}^{N_{\mathrm{ref}}} \mathrm{DeformAttn}(Q_p, \mathcal{P}(p,i,j), F_t^i)$$

matched viewpoint

BEV query

Projection function (Q_p to 2D)

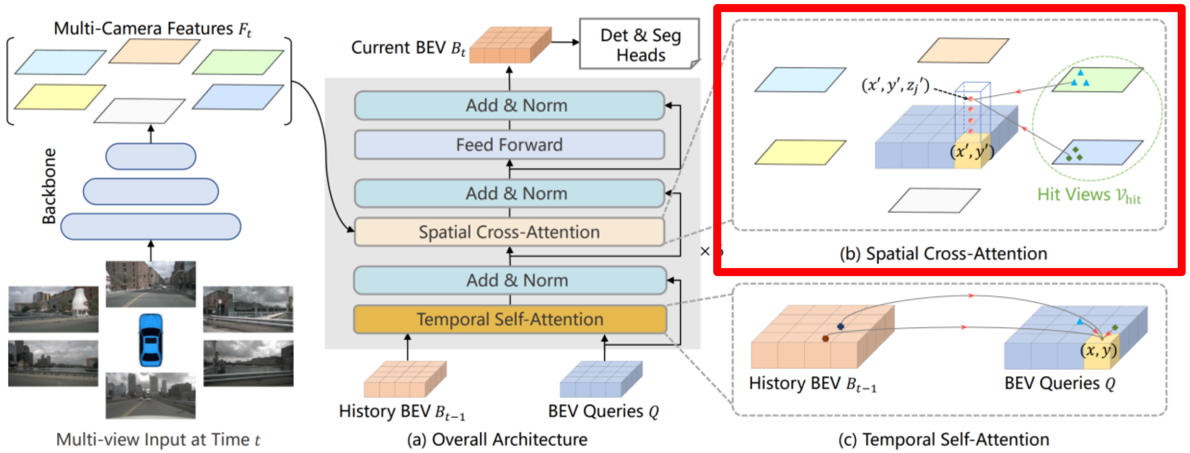
ith camera view feature

$$x' = (x - \frac{W}{2}) \times s; \quad y' = (y - \frac{H}{2}) \times s$$

$$\mathcal{P}(p,i,j) = (x_{ij}, y_{ij})$$

$$\text{where } z_{ij} \cdot \begin{bmatrix} x_{ij} & y_{ij} & 1 \end{bmatrix}^T = T_i \cdot \begin{bmatrix} x' & y' & z'_j & 1 \end{bmatrix}^T$$

Coordinate conversion (2D to 3D)



서강대학교 SOGANG UNIVERSITY

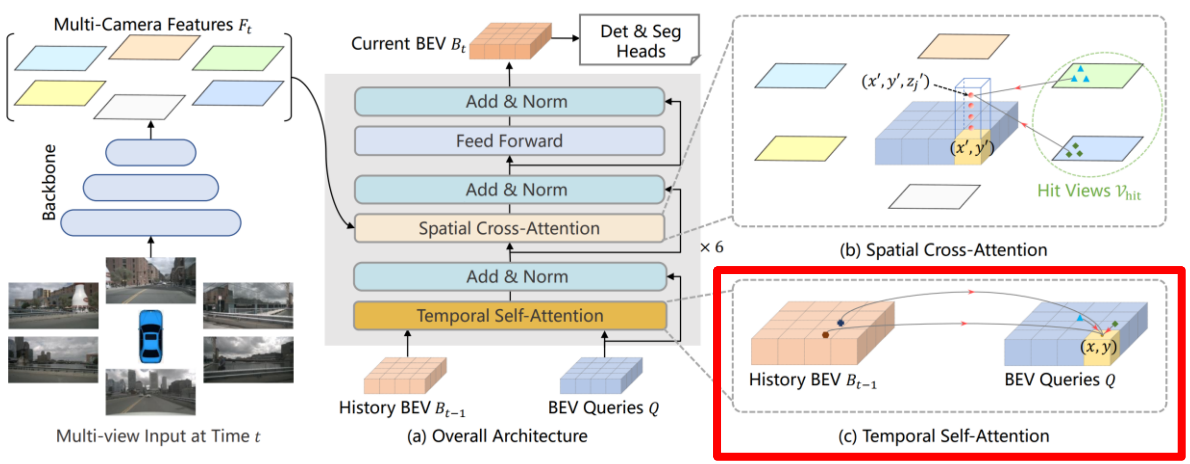
VDS LAB

BEVFormer

- Temporal self-attention
  - Incorporate temporal information from previous frames
    - Align the historical BEV features with the current BEV queries
      - Historical BEV is aligned with the ego-vehicle at the center
    - Temporal connections to verify the consistency of an object's identity

$$\mathrm{TSA}(\underbrace{Q_p}_{\text{Current BEV}}, \{Q, \underbrace{B'_{t-1}}_{\text{Historical BEV}}\}) = \sum_{V \in \{Q, B'_{t-1}\}} \mathrm{DeformAttn}(Q_p, p, V)$$

∴ Paper set t as 6



서강대학교 SOGANG UNIVERSITY

VDS LAB

BEVFormer

- Experimental results
  - Comparable performance with Lidar based models
  - BEVFormer outperforms BEVFormer-S
    - It indicates importance of considering temporal information

Table 1: **3D detection results on nuScenes `test` set.** $*$ notes that VoVNet-99 (V2-99) [21] was pre-trained on the depth estimation task with extra data [31]. "BEVFormer-S" does not leverage temporal information in the BEV encoder. "L" and "C" indicate LiDAR and Camera, respectively.

Method	Modality	Backbone	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
SSN [55]	L	-	0.569	0.463	-	-	-	-	-
CenterPoint-Voxel [52]	L	-	0.655	0.580	-	-	-	-	-
PointPainting [43]	L&C	-	0.581	0.464	0.388	0.271	0.496	0.247	0.111
FCOS3D [45]	C	R101	0.428	0.358	0.690	0.249	0.452	1.434	**0.124**
PGD [44]	C	R101	0.448	0.386	**0.626**	**0.245**	0.451	1.509	0.127
BEVFormer-S	C	R101	0.462	0.409	0.650	0.261	0.439	0.925	0.147
BEVFormer	C	R101	**0.535**	**0.445**	0.631	0.257	**0.405**	**0.435**	0.143
DD3D [31]	C	V2-99*	0.477	0.418	**0.572**	**0.249**	**0.368**	1.014	**0.124**
DETR3D [47]	C	V2-99*	0.479	0.412	0.641	0.255	0.394	0.845	0.133
BEVFormer-S	C	V2-99*	0.495	0.435	0.589	0.254	0.402	0.842	0.131
BEVFormer	C	V2-99*	**0.569**	**0.481**	0.582	0.256	0.375	**0.378**	0.126

BEVFormer

- Experimental results

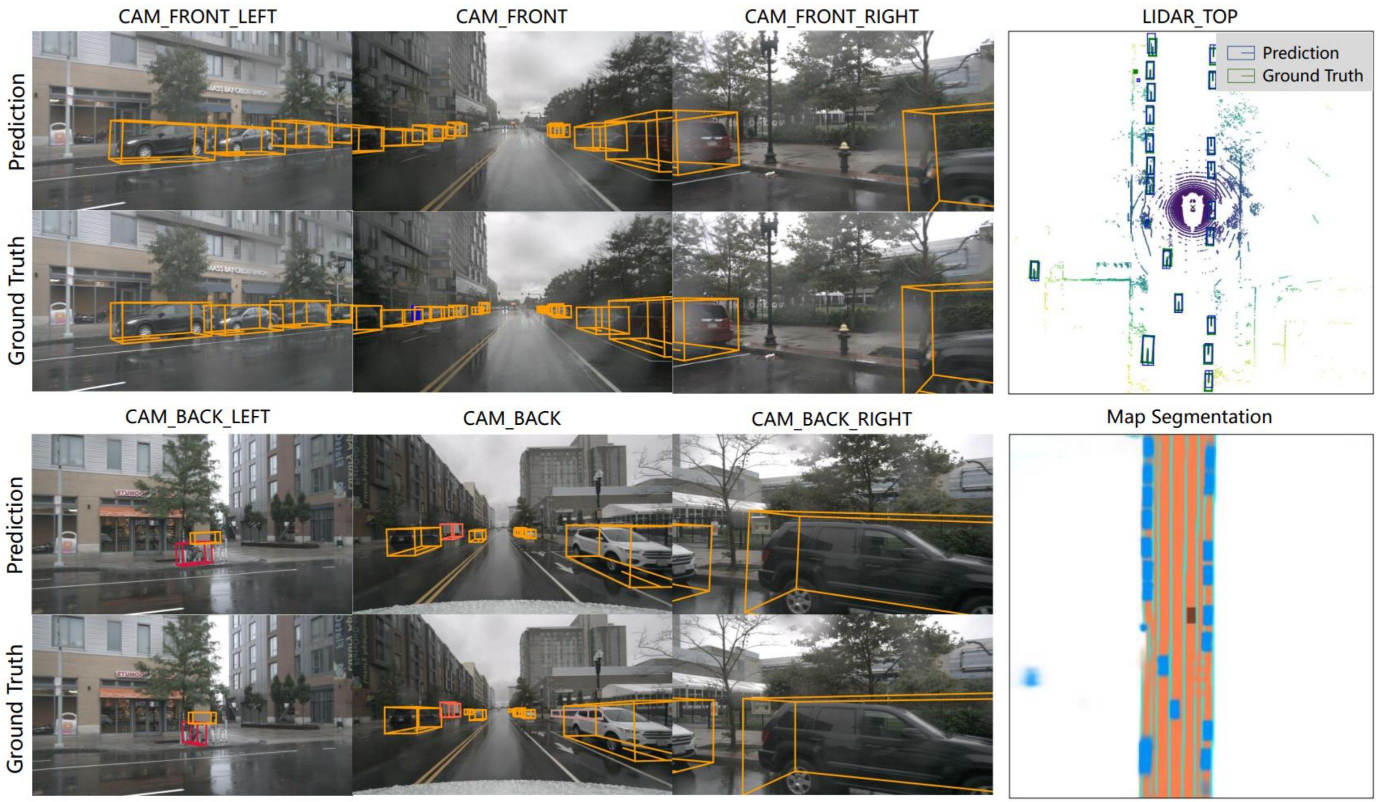


Figure 8: **Visualization results of both object detection and map segmentation tasks.** We show vehicle, road, and lane segmentation in blue, orange, and green, respectively.

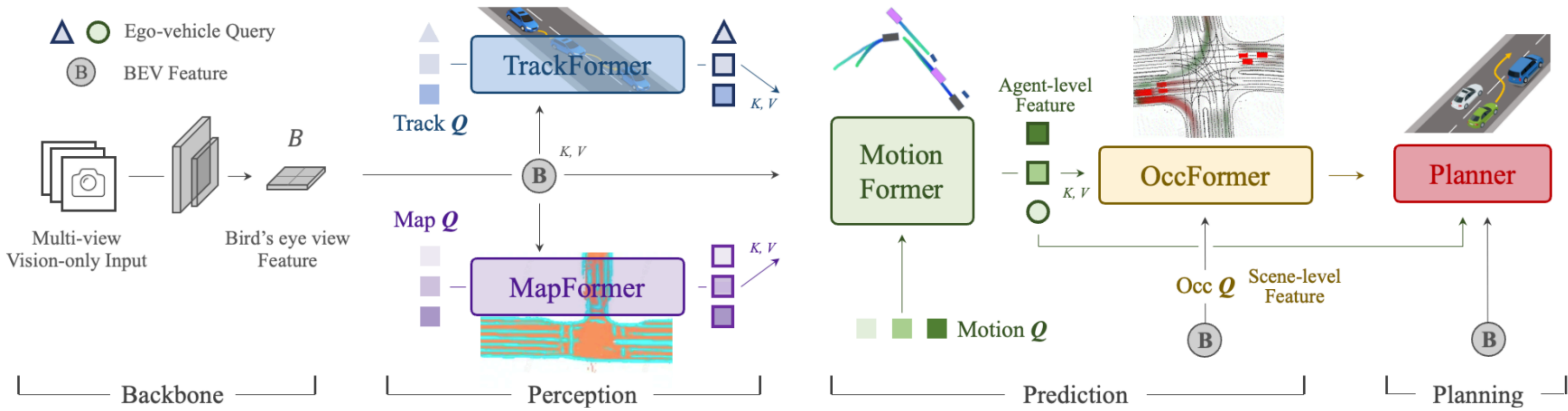
BEVFormer

- Limitations
  - BEVFormer only adopts 3D detection task
    - Not enough metrics for autonomous driving
  - Still has a low FPS due to high latency

Table 6: **Latency and performance of different model configurations on nuScenes val set.** The latency is measured on a V100 GPU, and the backbone is R101-DCN. The input image shape is 900×1600. "MS" notes multi-scale view features.

Method	Scale of BEVFormer			Latency (ms)			FPS	NDS↑	mAP↑
	MS	BEV	#Layer	Backbone	BEVFormer	Head			
BEVFormer	✓	200×200	6	391	130	19	1.7	**0.517**	**0.416**
A	✗	200×200	6	387	87	19	1.9	0.511	0.406
B	✓	100×100	6	391	53	18	2.0	0.504	0.402
C	✓	200×200	1	391	25	19	2.1	0.501	0.396
D	✗	100×100	1	387	**7**	18	**2.3**	0.478	0.374

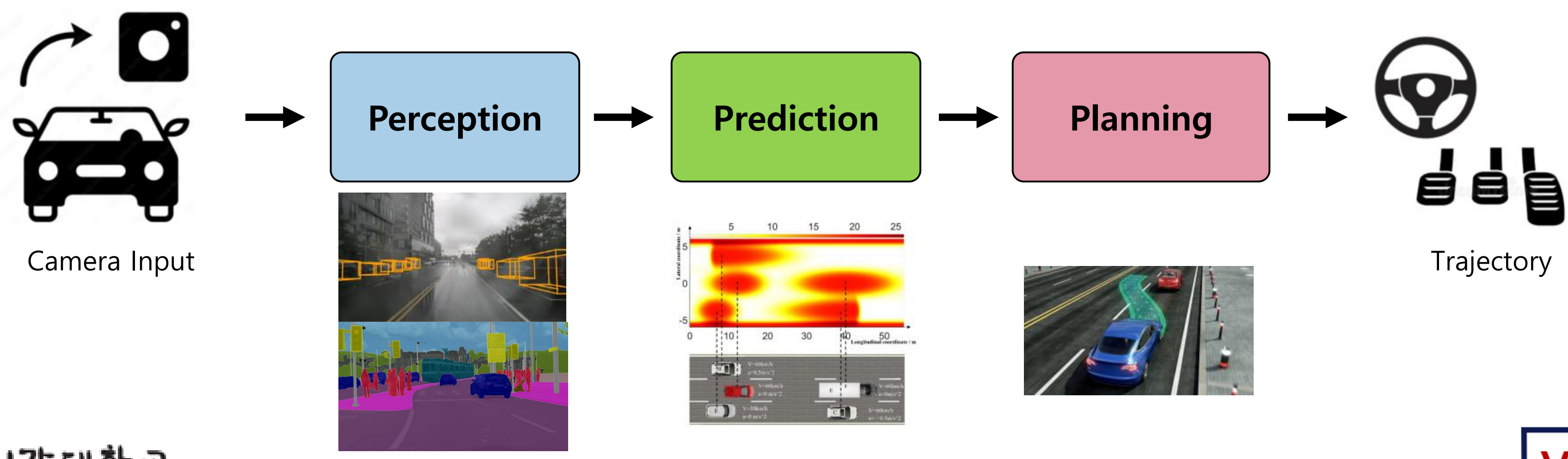
uniAD



- ***Best Paper:*** **Visual Programming: Compositional visual reasoning without training**
  Authors: Tanmay Gupta, Aniruddha Kembhavi (*Author Q&A*)
- ***Best Paper:*** **Planning-oriented Autonomous Driving**
  Authors: Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, Hongyang Li (*Author Q&A*)
- ***Best Paper Honorable Mention:*** **DynIBaR: Neural Dynamic Image-Based Rendering**
  Authors: Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, Noah Snavely
- ***Best Student Paper:*** **3D Registration with Maximal Cliques**
  Authors: Xiyu Zhang, Jiaqi Yang, Shikun Zhang, Yanning Zhang
- ***Best Student Paper Honorable Mention:*** **DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation**
  Authors: Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, Kfir Aberman

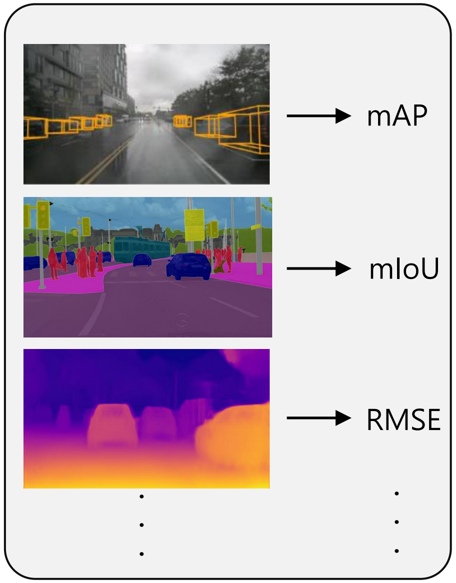
uniAD

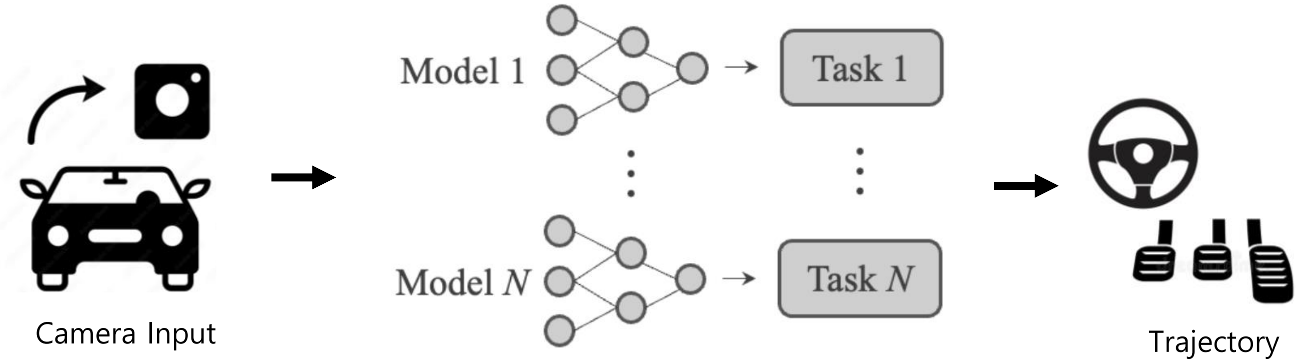
- Autonomous driving systems
  - Perception
    - Bounding boxes, map segmentation
  - Prediction
    - Predicts other object's occupancies
  - Planning
    - Plan the way where we go



uniAD

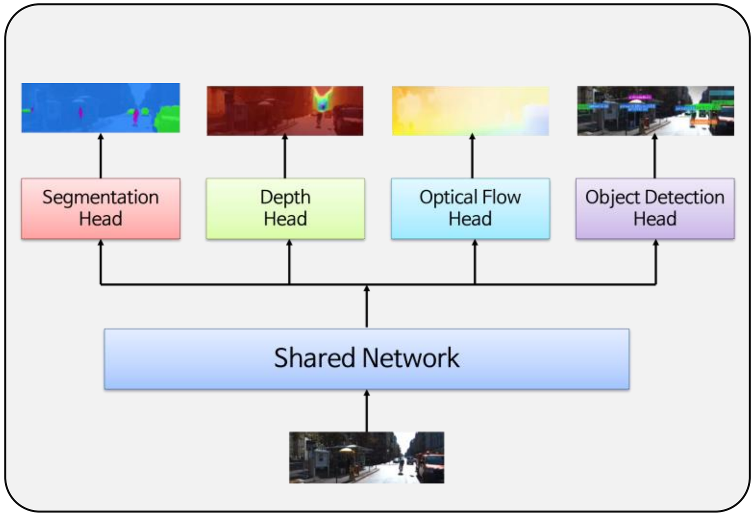
- Standalone models
  - Typical industry solutions
  - Pros
    - Independent teams for modules developments
      - Segmentation, object detection, depth estimation…
  - Cons
    - Severe error accumulation

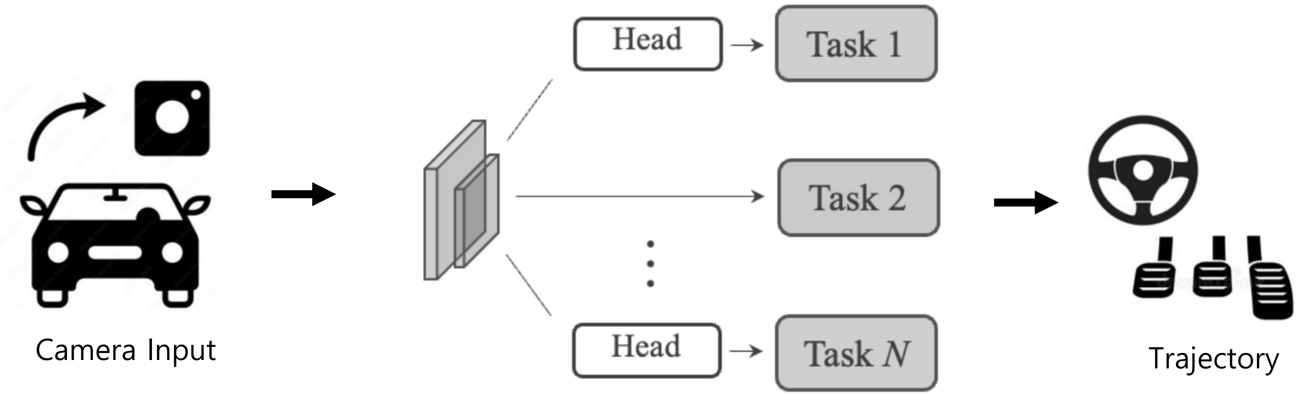




uniAD

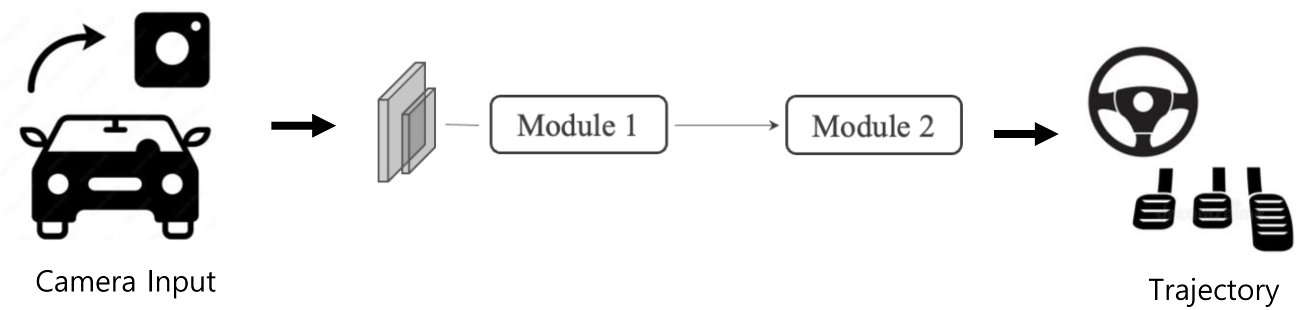
- Multi-task frameworks
  - Shared features for multiple tasks
  - Pros
    - Easily extend to multiple tasks
    - Efficient architecture for compute
  - Cons
    - Lack of tasks coordination





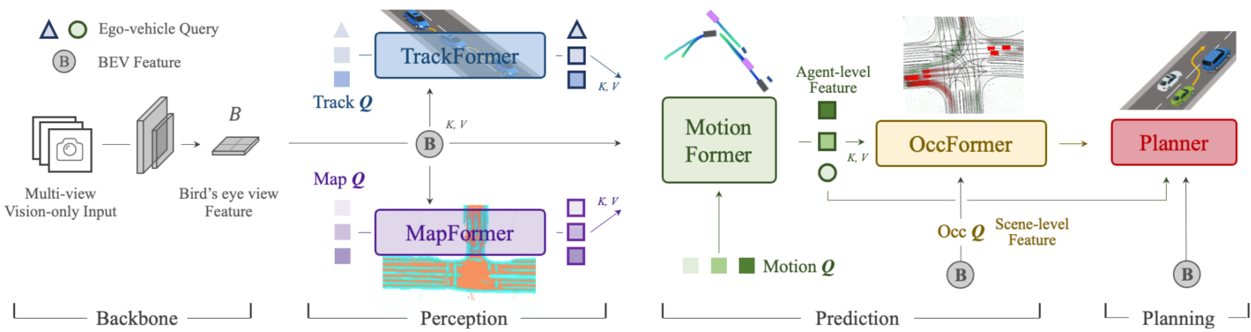
uniAD

- Previous end-to-end frameworks
  - Introduced multiple tasks to assist planning
  - Pros
    - Better interpretability with multiple tasks
  - Cons
    - Lack some crucial components
      - Occupancy module, prediction module …

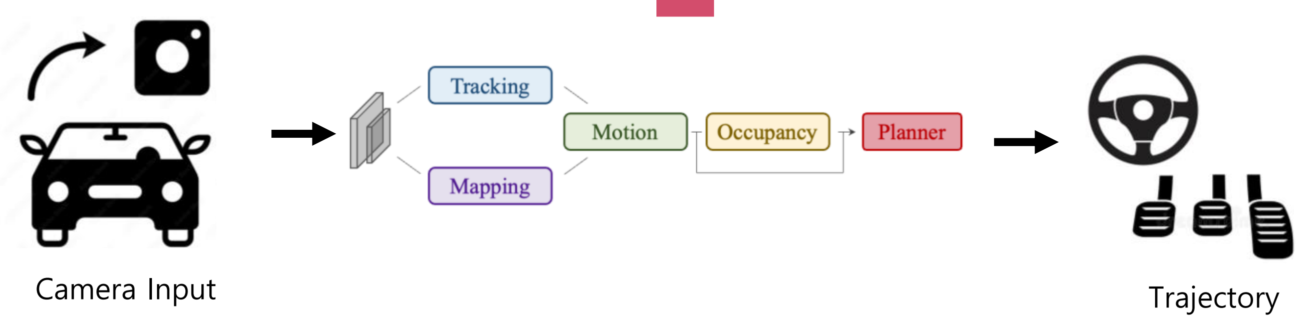


uniAD

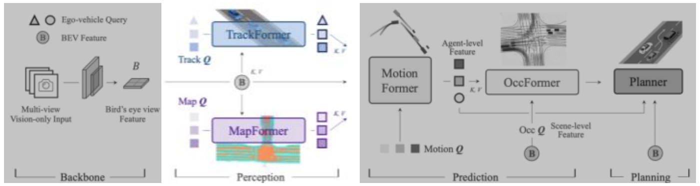
- Overall architecture
  - Unify full-stack Autonomous driving tasks
  - Coordinate all task towards safe planning



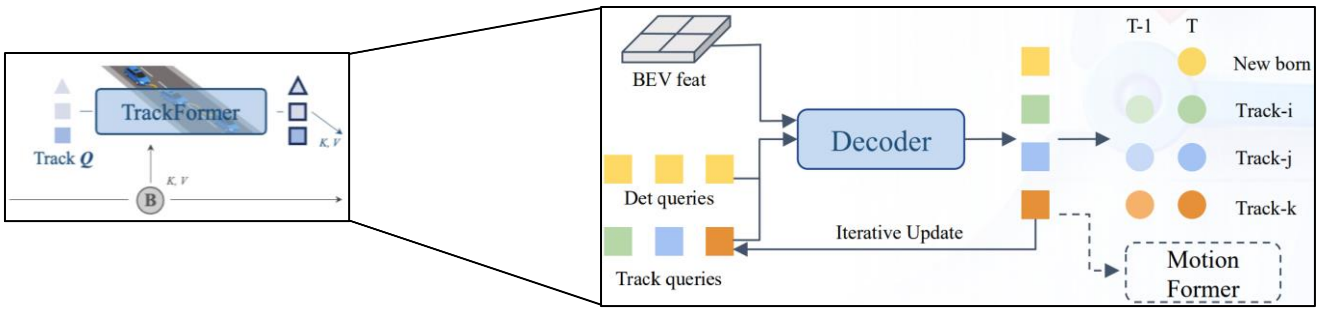
- ✓ Entire pipeline connected by queries
- ✓ Task coordinated with queries
- ✓ Interactions modeled by attention



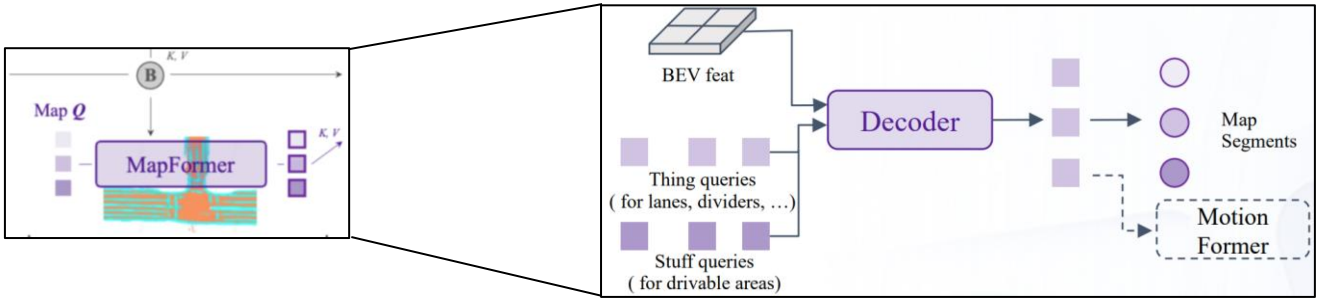
uniAD



- Perception
  - TrackFormer – MOTR (ECCV 2022)
    - End-to-end trainable tracking agents across time



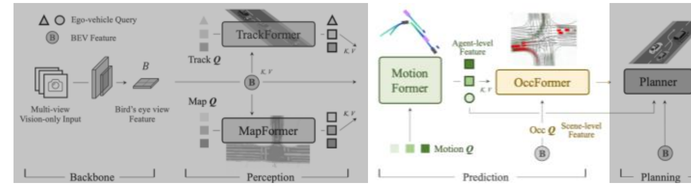
  - MapFormer – Panoptic SegFormer (CVPR 2022)
    - Each query represents a map element



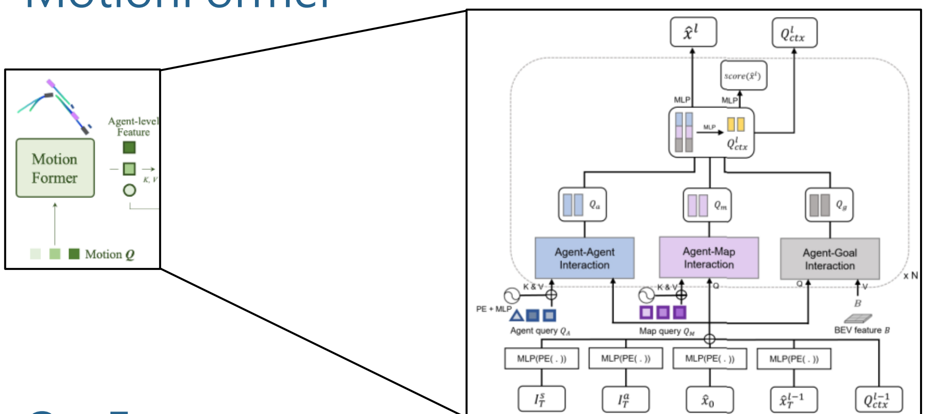
서강대학교 SOGANG UNIVERSITY

VDS LAB

uniAD

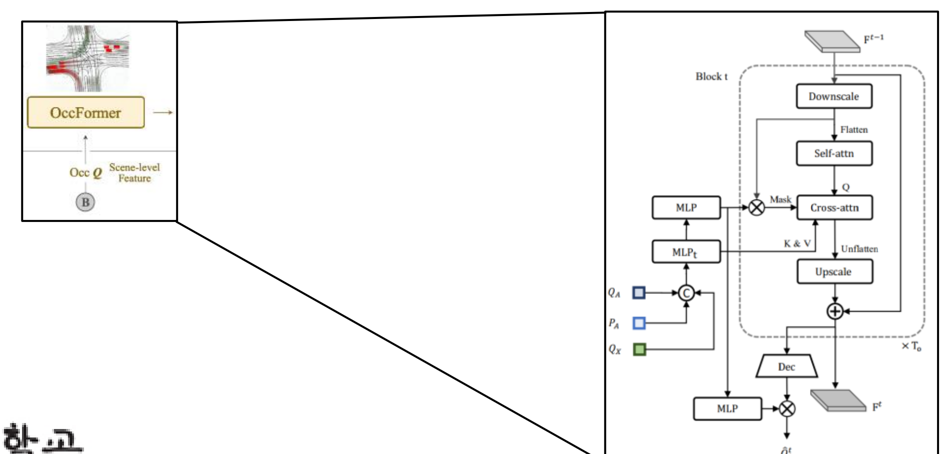


- Prediction
  - MotionFormer



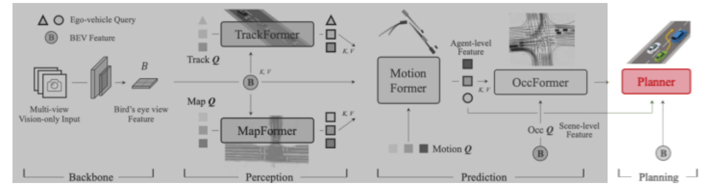
✓ Relation modelings via attention
- Agent-agent : TrackFormer K,V
- Agent-map : MapFormer K,V
- Agent-goal : BEV feature B

  - OccFormer



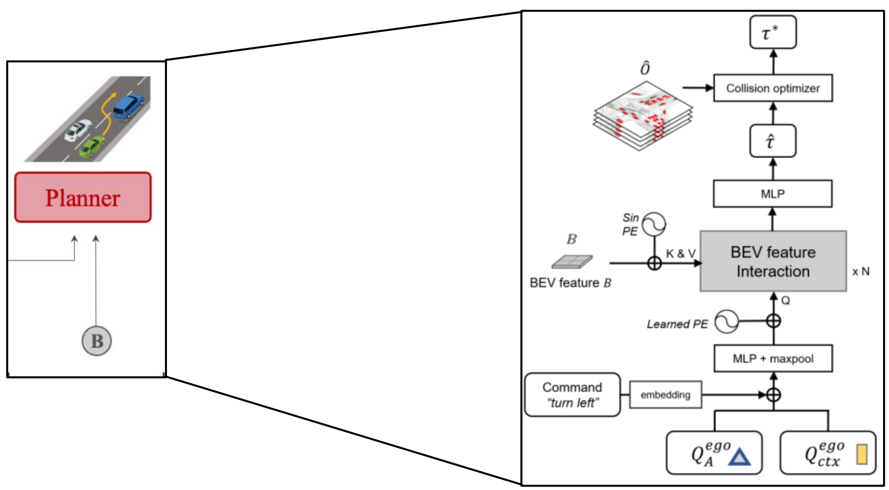
✓ Predict occupancy as attention mask

✓ Cross-attention for interaction with agent and environment from BEV features



uniAD

- Planning
  - Planner
    - Using ego-vehicle query from MotionFormer
      - Interaction with other agents
    - Collision optimization
      - Steer the predicted trajectories clear of predicted occupancy



uniAD

- Experimental results

Method	L2(m)↓				Col. Rate(%)↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
NMP† [101]	-	-	2.31	-	-	-	1.92	-
SA-NMP† [101]	-	-	2.05	-	-	-	1.59	-
FF† [37]	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
EO† [47]	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
ST-P3 [38]	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD	**0.48**	**0.96**	**1.65**	**1.03**	**0.05**	**0.17**	**0.71**	**0.31**

ID	Det.	Track	Map	Motion	Occ.	Plan	#Params	FLOPs	FPS
0 [105]	✓		✓		✓		102.5M	1921G	-
1	✓						65.9M	1324G	4.2
2	✓	✓					68.2M	1326G	2.7
3	✓	✓	✓				95.8M	1520G	2.2
4	✓	✓	✓	✓			108.6M	1535G	2.1
5	✓	✓	✓	✓	✓		122.5M	1701G	2.0
6	✓	✓	✓	✓	✓	✓	125.0M	1709G	1.8

ID	Modules					Tracking			Mapping		Motion Forecasting			Occupancy Prediction				Planning	
	Track	Map	Motion	Occ.	Plan	AMOTA↑	AMOTP↓	IDS↓	IoU-lane↑	IoU-road↑	minADE↓	minFDE↓	MR↓	IoU-n.↑	IoU-f.↑	VPQ-n.↑	VPQ-f.↑	avg.L2↓	avg.Col.↓
0*	✓	✓	✓	✓	✓	0.356	1.328	893	0.302	0.675	0.858	1.270	0.186	55.9	34.6	47.8	26.4	1.154	0.941
1	✓					0.348	1.333	791	-	-	-	-	-	-	-	-	-	-	-
2		✓				-	-	-	**0.305**	0.674	-	-	-	-	-	-	-	-	-
3	✓	✓				0.355	1.336	785	0.301	0.671	-	-	-	-	-	-	-	-	-
4			✓			-	-	-	-	-	0.815	1.224	0.182	-	-	-	-	-	-
5	✓		✓			0.360	1.350	919	-	-	0.751	1.109	0.162	-	-	-	-	-	-
6	✓	✓	✓			0.354	1.339	820	0.303	0.672	0.736(-9.7%)	1.066(-12.9%)	0.158	-	-	-	-	-	-
7				✓		-	-	-	-	-	-	-	-	60.5	37.0	52.4	29.8	-	-
8	✓			✓		0.360	**1.322**	809	-	-	-	-	-	62.1	38.4	52.2	32.1	-	-
9	✓	✓	✓	✓		0.359	1.359	1057	0.304	**0.675**	**0.710**(-3.5%)	**1.005**(-5.8%)	**0.146**	**62.3**	39.4	**53.1**	32.2	-	-
10					✓		-	-	-	-	-	-	-	-	-	-	-	1.131	0.773
11	✓	✓	✓		✓	**0.366**	1.337	889	0.303	0.672	0.741	1.077	0.157	-	-	-	-	1.014	0.717
12	✓	✓	✓	✓	✓	0.358	1.334	**641**	0.302	0.672	0.728	1.054	0.154	**62.3**	**39.5**	52.8	**32.3**	**1.004**	**0.430**

uniAD

- Experimental results
  - Cruising around urban scene

서강대학교 SOGANG UNIVERSITY

VDS LAB

uniAD

- Experimental results
  - Obstacle avoidance visualizations

Conclusion

- BEVFormer
  - Achieved comparable performance to Lidar-based models
  - Only 3D object detection output and high latency
- UniAD
  - An end-to-end autonomous driving framework
    - Pursuit of safe planning
  - State-of-the-art (SOTA) performance with vision-only input

Thank you for listening