BEVFormer
looking back and looking forward

李志琦
南京大学

Outline

南京大學 NANJING UNIVERSITY

- **BEV研究现状**
- BEVFormer及相关改进
- 基于BEVFormer的端到端自动驾驶

BEV检测研究现状

相比DD3D(单目方法)，基于BEV的方法获得了超过20个点的提升这些提升来源于：

1、时序信息的使用 （短时序，长时序）
- BEVFormer
- SOlOFusion

2、更好的基础网络（backbone）
StreamPETR-Large引入预训练ViT

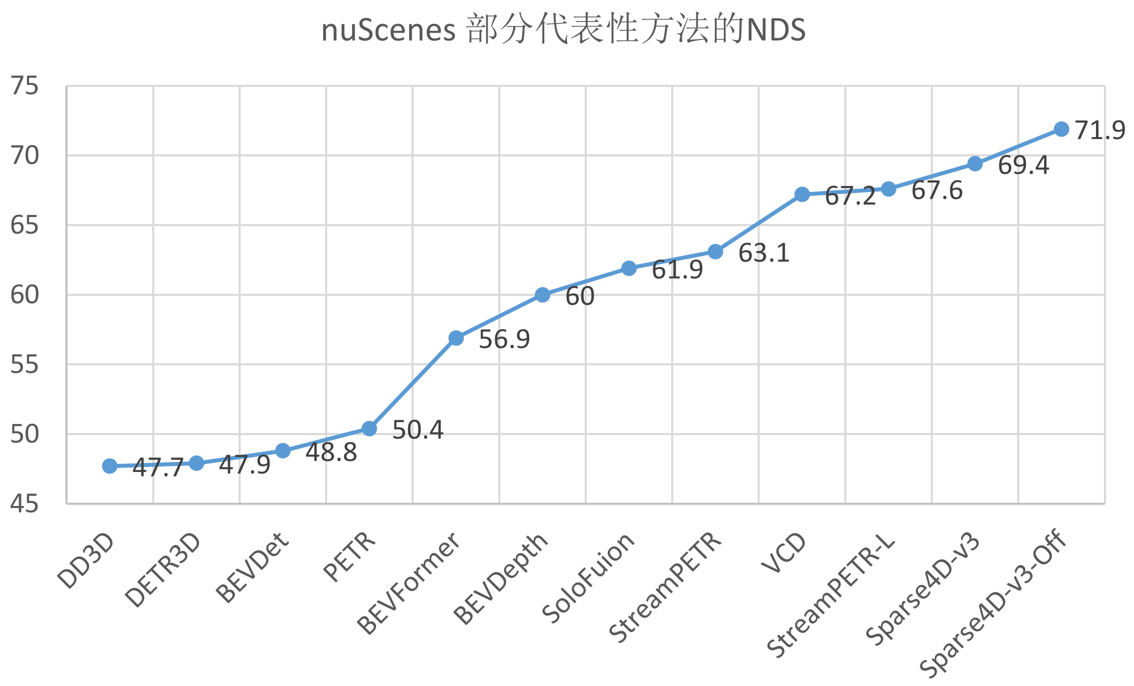
3、Dense BEV->Sparse BEV
StreamPETR, SparseBEV, Sparse4D

4、更好的深度估计
BEVDepth， BEVNeXt

5、未来帧的使用?
BEVFormerv2, Sparse4D-v3-Off

6、跨模态蒸馏
VCD, BEVDistll

…



BEV检测研究现状

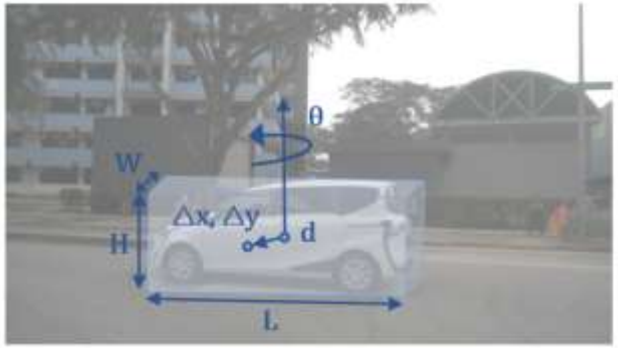
Method					Metrics									
Date	Name	Modalities	Map data	External data	mAP	mATE (m)	mASE (1-IOU)	mAOE (rad)	mAVE (m/s)	mAAE (1-acc)	NDS	PKL *	FPS (Hz)	Stats
		Camera	All	All										
2023-10-13	Sparse4D-v3-offline	Camera	no	yes	0.668	0.346	0.234	0.279	0.142	0.145	0.719	0.686	n/a	
2024-05-09	HENet_Sp	Camera	no	no	0.645	0.402	0.235	0.237	0.155	0.129	0.707	0.704	n/a	
2023-10-16	Sparse4D-v3	Camera	no	yes	0.630	0.379	0.235	0.281	0.184	0.127	0.694	0.751	n/a	
2024-02-01	HaomoAI Perceptic	Camera	no	no	0.624	0.405	0.238	0.288	0.188	0.119	0.688	0.808	n/a	
2023-08-01	Far3D	Camera	no	no	0.635	0.432	0.237	0.278	0.227	0.130	0.687	0.757	n/a	
2024-03-13	RayDN	Camera	no	no	0.631	0.437	0.235	0.283	0.220	0.120	0.686	0.793	n/a	
2023-04-05	HoP	Camera	no	no	0.624	0.367	0.249	0.353	0.171	0.131	0.685	0.875	n/a	
2023-08-29	Li	Camera	no	yes	0.623	0.433	0.238	0.287	0.221	0.129	0.681	0.788	n/a	
2023-05-03	StreamPETR-Large	Camera	no	no	0.620	0.470	0.241	0.258	0.236	0.134	0.676	0.880	n/a	
2023-08-17	SparseBEV	Camera	no	no	0.603	0.425	0.239	0.311	0.172	0.116	0.675	0.789	n/a	

前10名有~8个基于Sparse BEV的方法, 2024年新提交的方法仅有三个，均为基于去年方法的改进

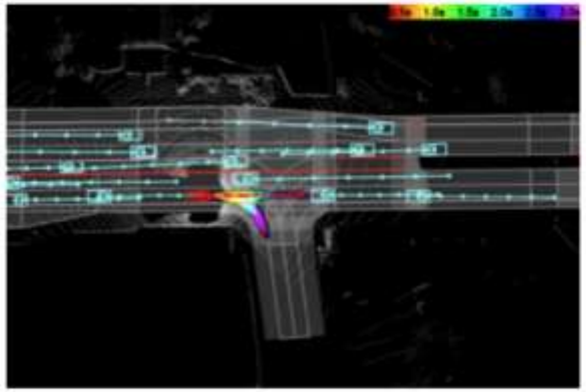
BEV的优越性

BEV Perception ~~is the future~~ has now come to pass of vision-centric perception[1]:

- Fuse multi-camera features in early stage.
- Straightforward to combine with other modalities.
  e.g. BEVFusion
- Readily consumable by downstream such as prediction and planning.
  e.g. BEVFormer->UniAD, VAD



3D Detection [2]

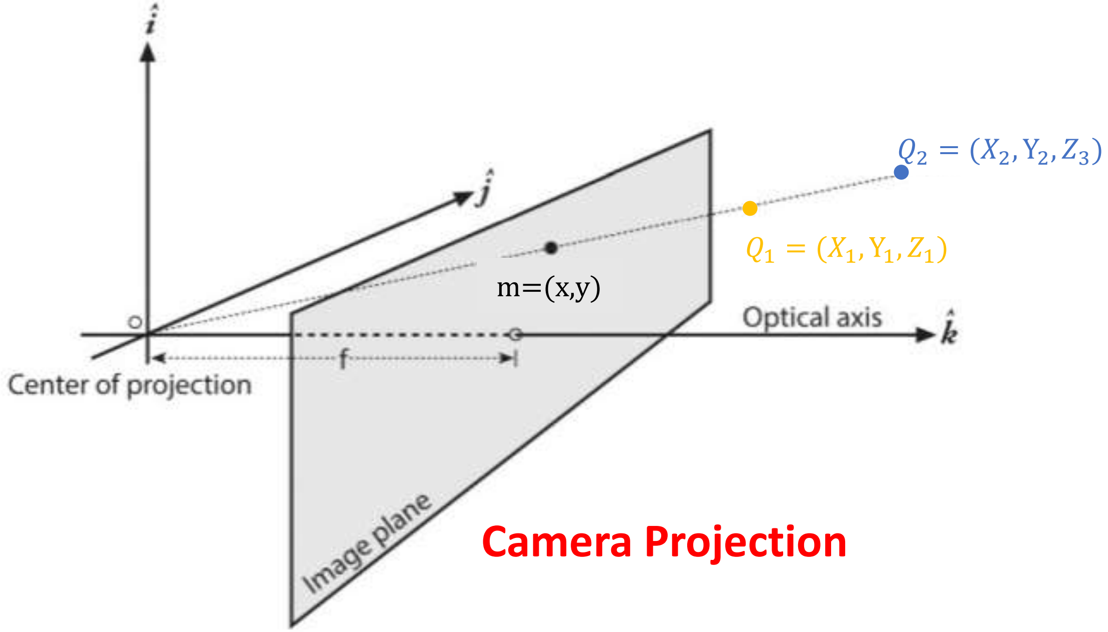


bird's-eye-view (BEV)[3]

[1] Monocular BEV Perception with Transformers in Autonomous Driving, Patrick Langechuan Liu
[2] FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection
[3] Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D

5

View Transformation



Camera Projection

Geometry-Base

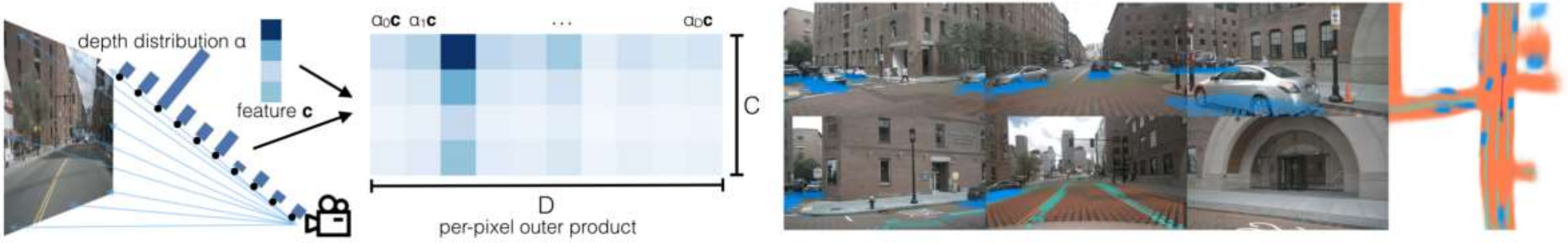
- From 3D to 2D (issue: Multiple 3D points will hit the same 2D pixel.)
- From 2D to 3D (issue: Depth is unknown)

Learning-Base

- Attention is all you need (issue: Not as efficient as geometry-based methods)

No matter what, the transformation is ill-posed

2D-3D: 以LSS为例



LSS[1]

Using categorical distribution over depth instead of depth estimates.

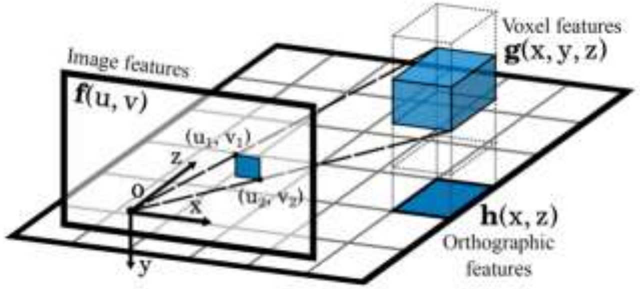
- Strength:
  - Generate representation as all possible depths for each pixel.
- Weakness:
  - The generated BEV is discontinuous and sparse.
  - The fusion process is inefficient.

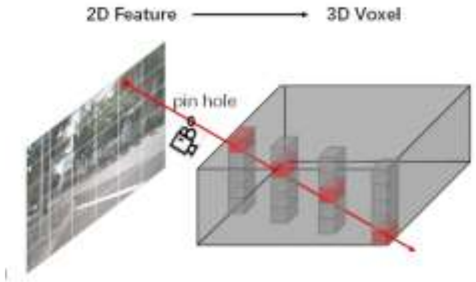
Following works:
- CADDN[2]
- FIERY[3]
- BEVDet[4]

[1] Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D
[2] Categorical Depth Distribution Network for Monocular 3D Object Detection
[3] FIERY: Future Instance Prediction in Bird's-Eye View From Surround Monocular Cameras
[4] BEVDet: High-performance Multi-camera 3D Object Detection in Bird-Eye-View

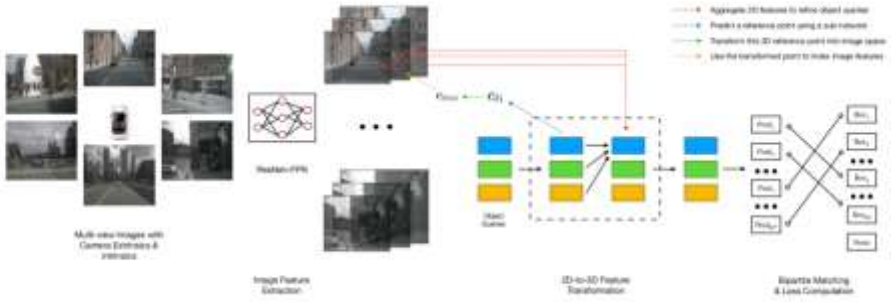
3D-2D: 以OFT为例

NANJING UNIVERSITY





OFT(left)[1], M2BEV(right)[2]



DETR3D[3]

Obtain image features that corresponds to predefined 3D anchors.

- Strength:
  - Dense or Sparse BEV feature maps.
  - Efficient compared to 2D to 3D.
- Weakness:
  - False positive BEV features.

Related works:

- ImVoxelNet[5]
- DETR3D[3]
- BEVFormer[4]

[1] Orthographic Feature Transform for Monocular 3D Object Detection
[2] M2BEV: Multi-Camera Joint 3D Detection and Segmentation with Unified Bird's-Eye View
[3] DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries
[4] BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers
[5] ImVoxelNet: Image to Voxels Projection for Monocular and Multi-View General-Purpose 3D Object Detection

Learning-Base: 以StreamPETR为例

除了2D-3D, 3D-2D的显式转换方法外，还有一些方法借助attention隐式完成视角转换
但是根据现有方法来看，显式转换具有效果更好，速度更快，显存消耗更低等优势。

Deformable Attention 这一在BEVFormer, RepDETR3D等方法中广泛使用的稀疏注意力机制

1. 总体时间占head的比例并不显著
2. 提速~2x的Deformable Attention OP即将发布
3. Flash Attn 在长序列上相比Deformable Attention并不占优势

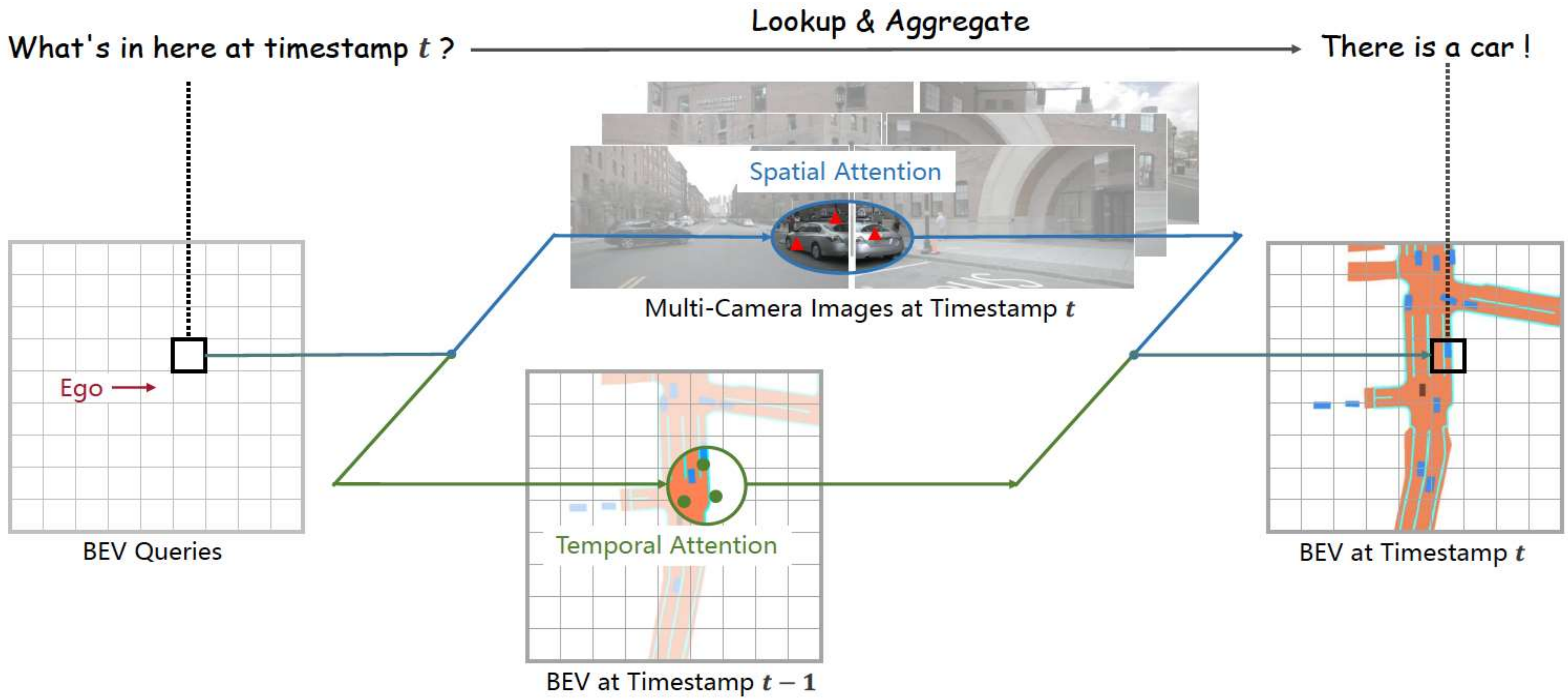
Model	Setting	Pretrain	Lr Schd	Training Time	NDS	mAP	FPS-pytorch	Config	Download
StreamPETR	V2-99 - 900q	FCOS3D	24ep	13 hours	57.1	48.2	12.5	config	model/log
RepDETR3D	V2-99 - 900q	FCOS3D	24ep	13 hours	58.4	50.1	13.1	config	model/log

Outline

南京大学 NANJING UNIVERSITY

- BEV研究现状
- **BEVFormer及相关改进**
- 基于BEVFormer的端到端自动驾驶

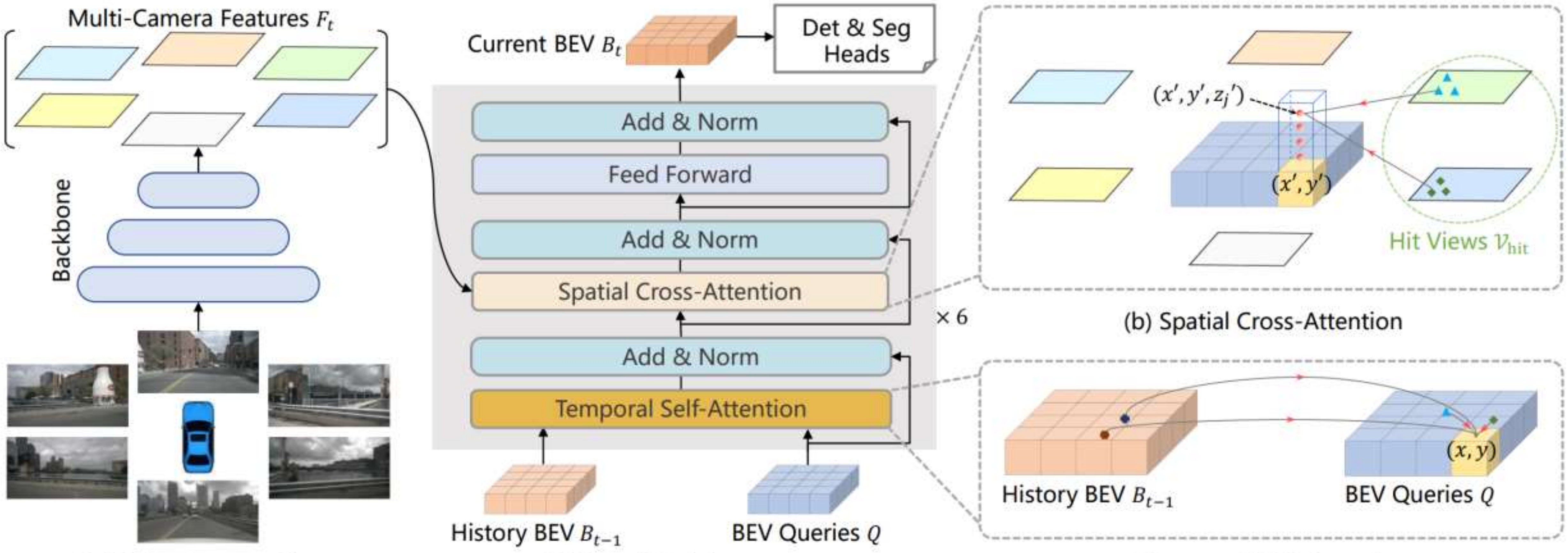
Sum-up: BEVFormer



Key Points

1. Using *learnable queries* to represent real world from BEV view.
2. Lookup *spatial features in images* and *temporal features in previous BEV map*

Overall Architecture: BEVFormer



(a) Overall Architecture

(b) Spatial Cross-Attention

(c) Temporal Self-Attention

Spatial Cross-Attention

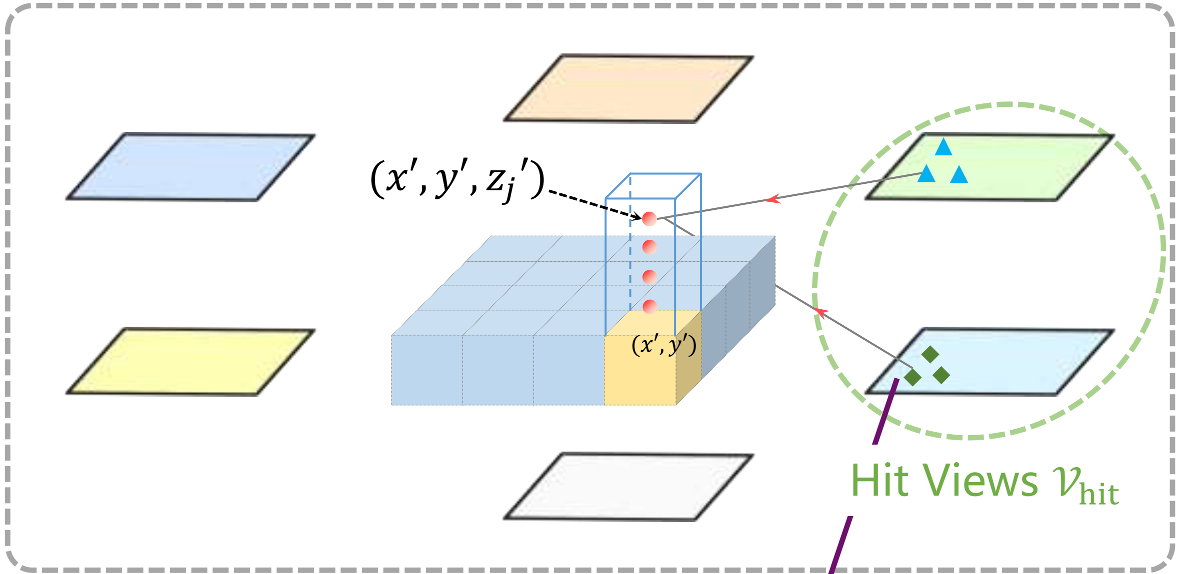
南京大學 NANJING UNIVERSITY

- **Lookup and aggregate the spatial information**

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

Key Steps:

1. Lift each BEV query to be a ***pillar***
2. Project the ***3D points*** in pillar to ***2D points*** in views
3. Sample features from ***RoIs in hit views***
4. Fuse by weight



Spatial Cross-Attention

Sparse Attention, e.g., Deformable Attention [1]

[1] Zhu, Xizhou, et al. "Deformable detr: Deformable transformers for end-to-end object detection." ICLR (2020).

NANJING UNIVERSITY

- **Lookup and Aggregate the Temporal information**

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

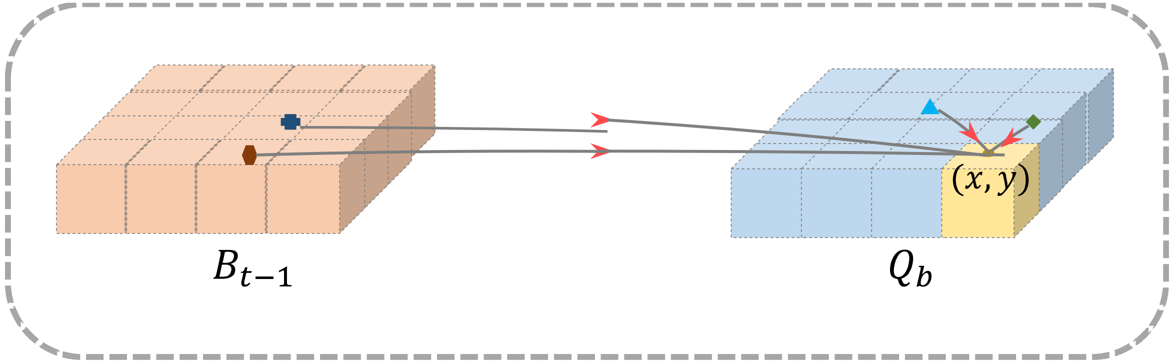
Key Steps:

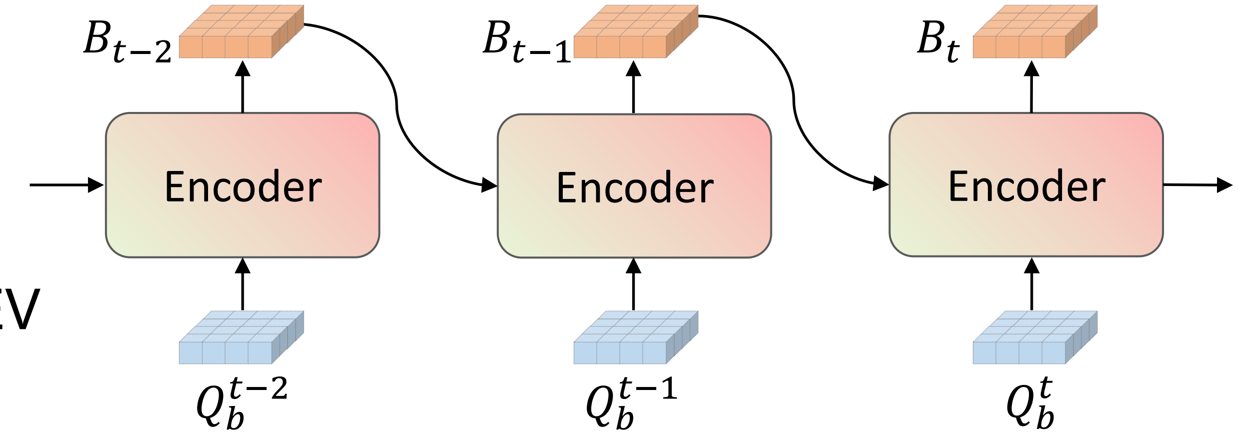
1. ***Align two BEV maps*** according to the ego motion.

2. Sample features from ***both past and current***.

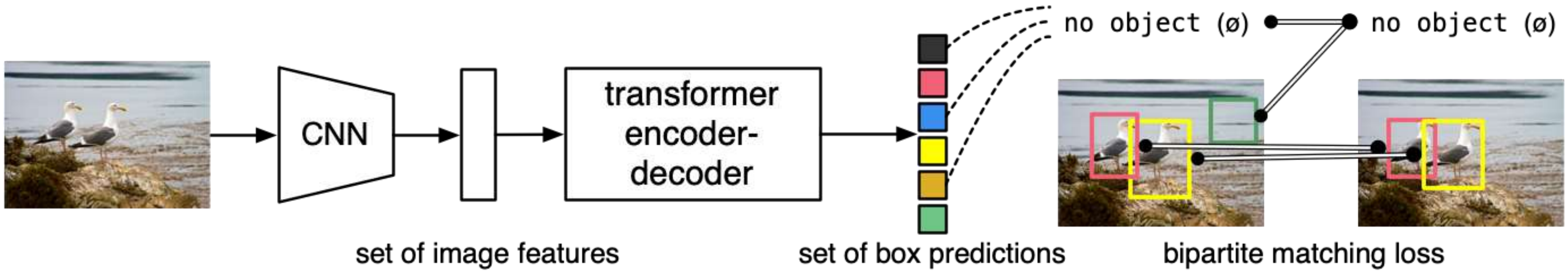
3. ***Weighted summation of sampled features*** from past and current BEV maps.

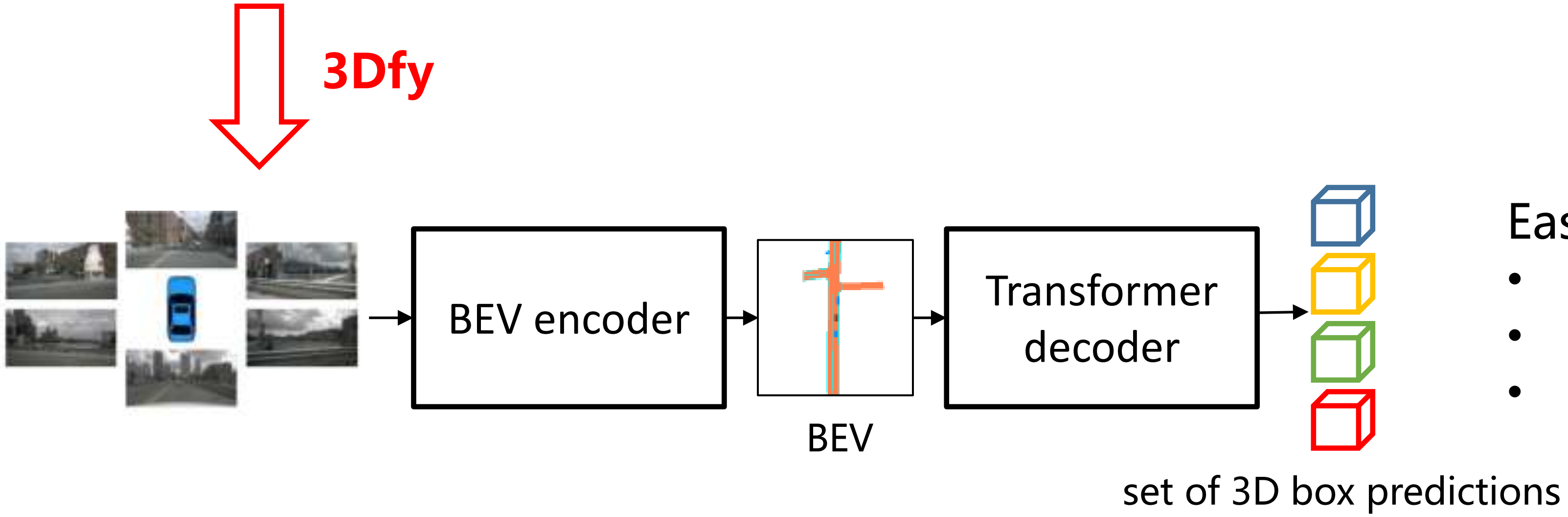
4. Use ***RNN-style*** to literately collect history BEV features





3Dfy DETR Detector

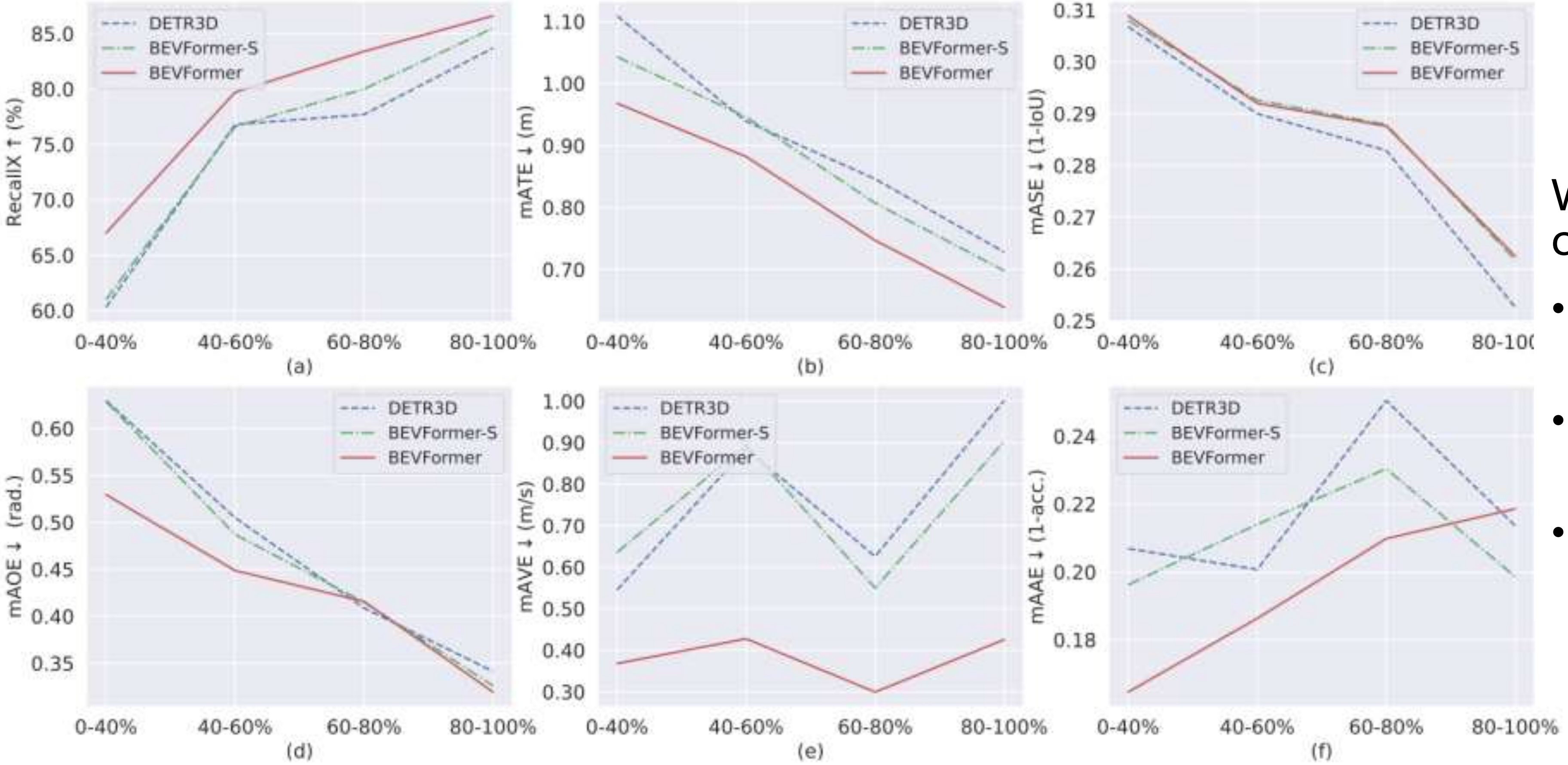




Easy to leverage 2D detectors:

- DETR
- Deformable DETR
- DN-DETR, DINO

Temporal clues matters

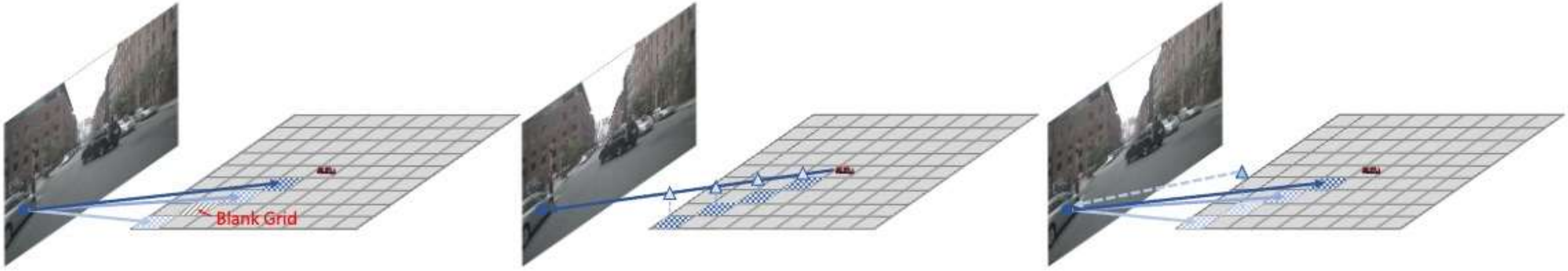


With temporal clues, we obtain:

- Higher recall, especially for *low-visible objects*
- More accurate *location estimation*
- Very accurate estimation of *velocity*

FB-BEV: Forward-Backward Projection

南京大學 NANJING UNIVERSITY



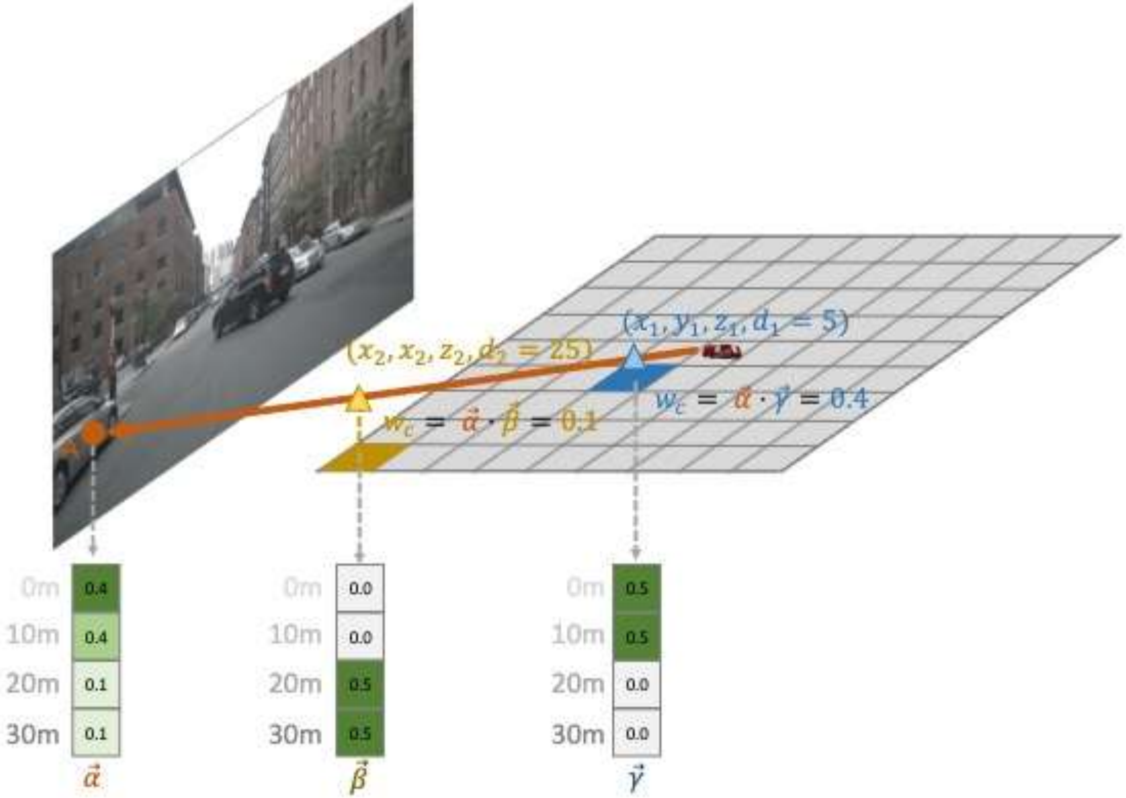
(a) Forward Projection (b) Backward Projection (c) Forward-backward Projection (ours)

Forward Projection
- Weakness: Blank Grid
- Solution: fill the blank grid with the backward projection

Backward Projection
- Weakness: Unable to utilize depth information
- Solution: Propose Depth-aware Backward Projection

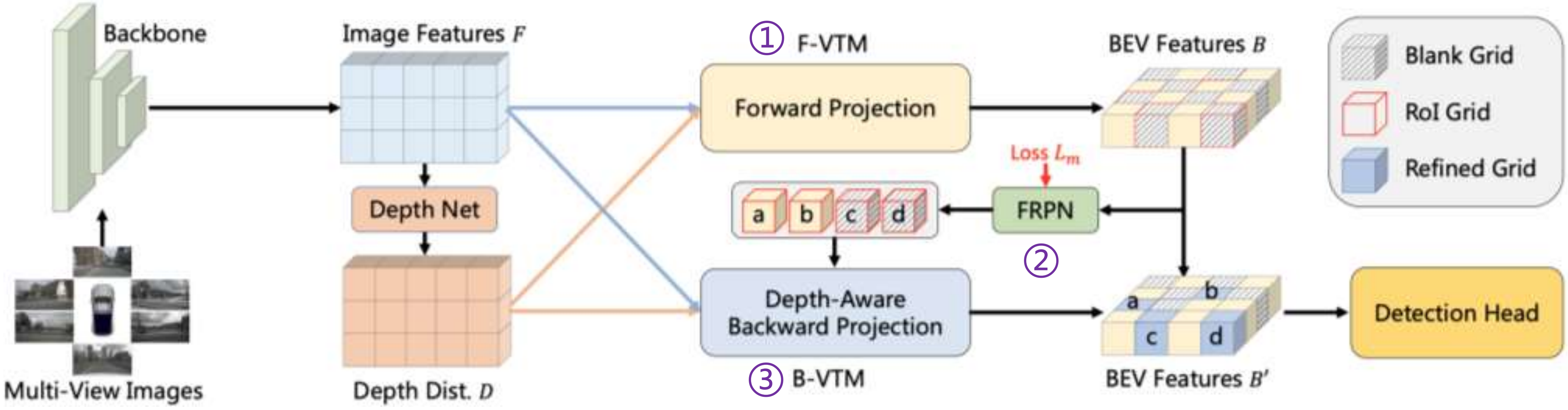
Neither Forward Projection or Backward Projection is perfect, but they are basically **complementary**.

FB-BEV: Depth-Aware Backward Projection



Backward projection can also model more accurate projection relationship based on depth distribution.

FB-BEV

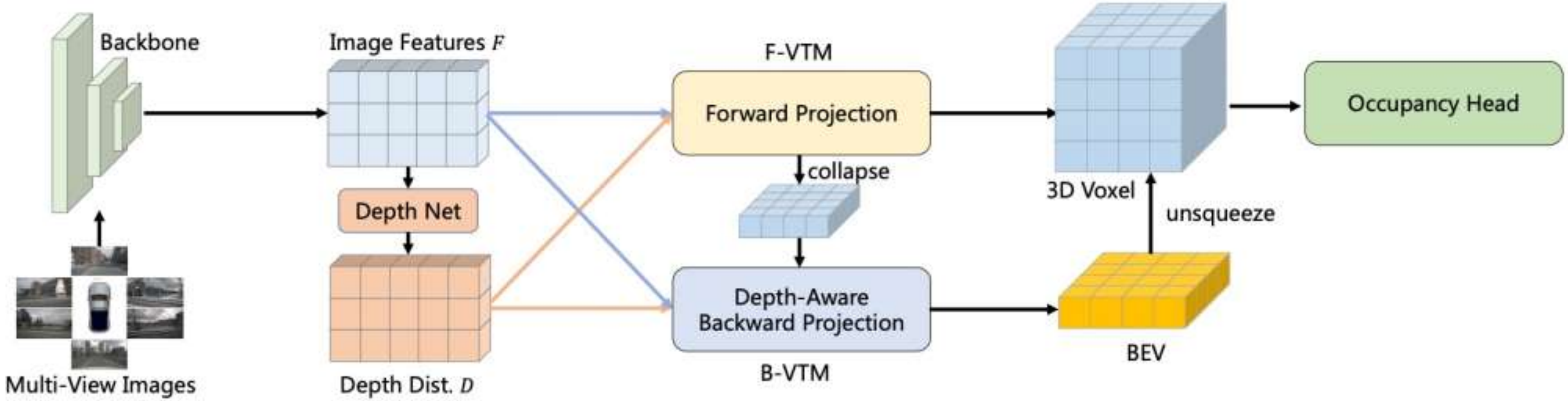


① Forward Projection provides initial sparse BEV features

② FRPN extract foreground BEV features

③ Depth-aware Backward Projection optimizes the foreground features

FB-BEV->FB-OCC



Joint Voxel and BEV representation

Joint Forward and Backward Projection

Sparse BEV vs Dense BEV

南京大学 NANJING UNIVERSITY

	Dense BEV	Sparse BEV
计算效率	特定优化，可以与Sparse BEV方法比较	高
感知范围	计算量正比于Dense BEV感知范围的平方	适合远距离
多任务支持	友好	不适合OCC等密集预测任务
感知性能	可以逼近Sparse BEV的性能	高
算子	Deformable/BEV poolv2/Conv	Vanilla Attn/Deformable

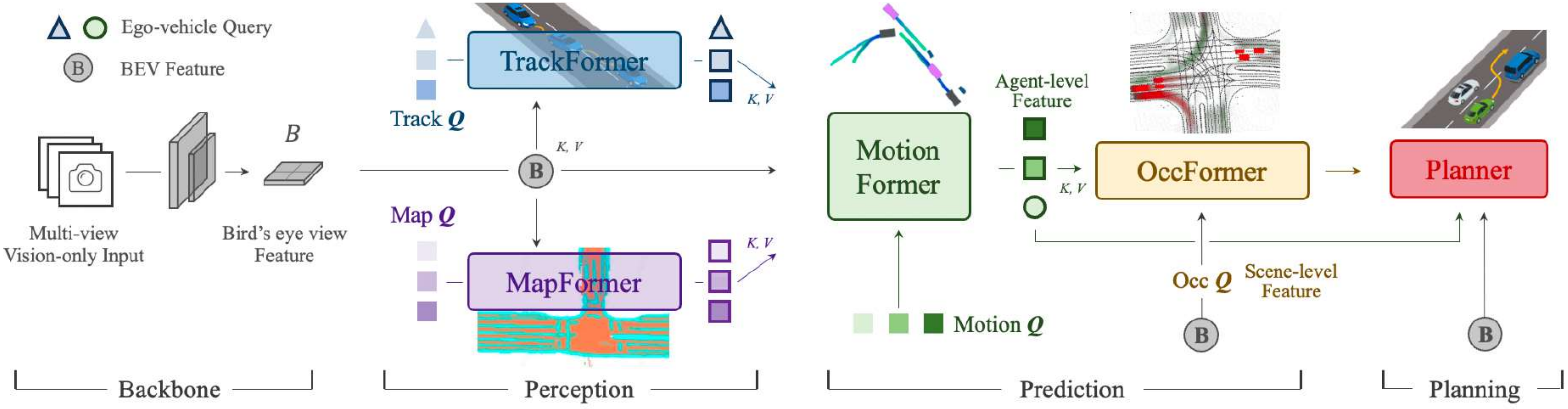
Dense BEV和Sparse BEV, 各有优劣，难以说那种方式更优

Outline

南京大學 NANJING UNIVERSITY

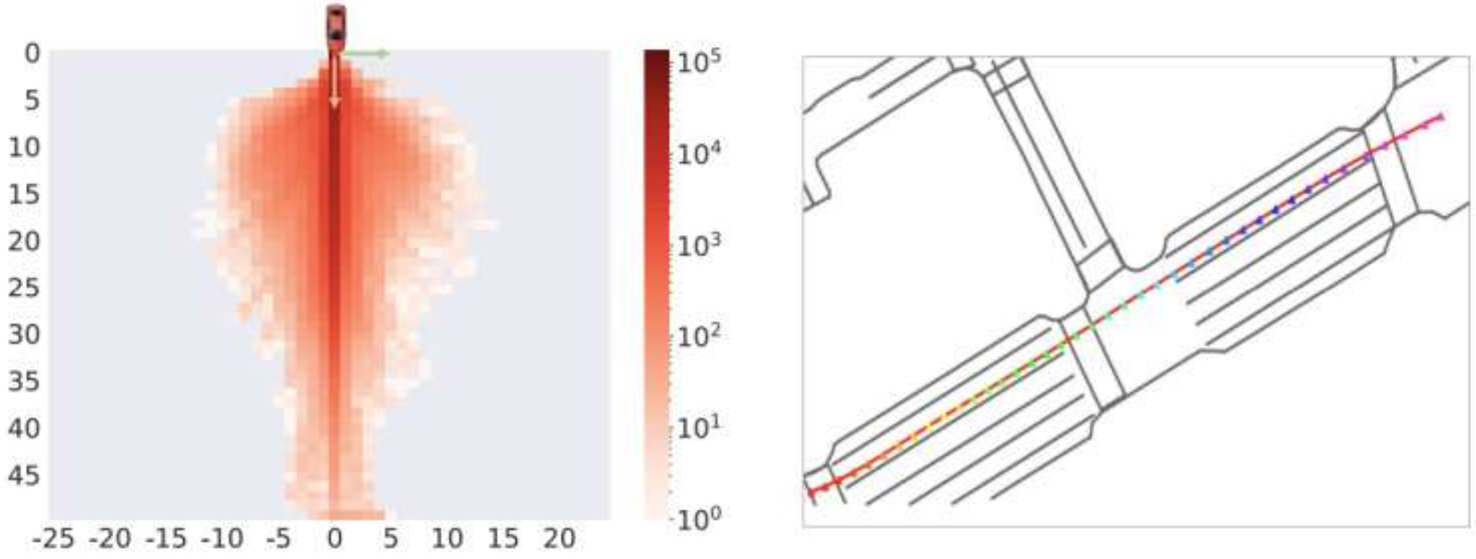
- BEV研究现状
- BEVFormer及相关改进
- **基于BEVFormer的端到端自动驾驶**

Open Loop End-to-End Autonomous Driving: UniAD



UniAD validates open-loop end-to-end autonomous driving on nuScenes

NuScenes is an imbalance dataset for planning task



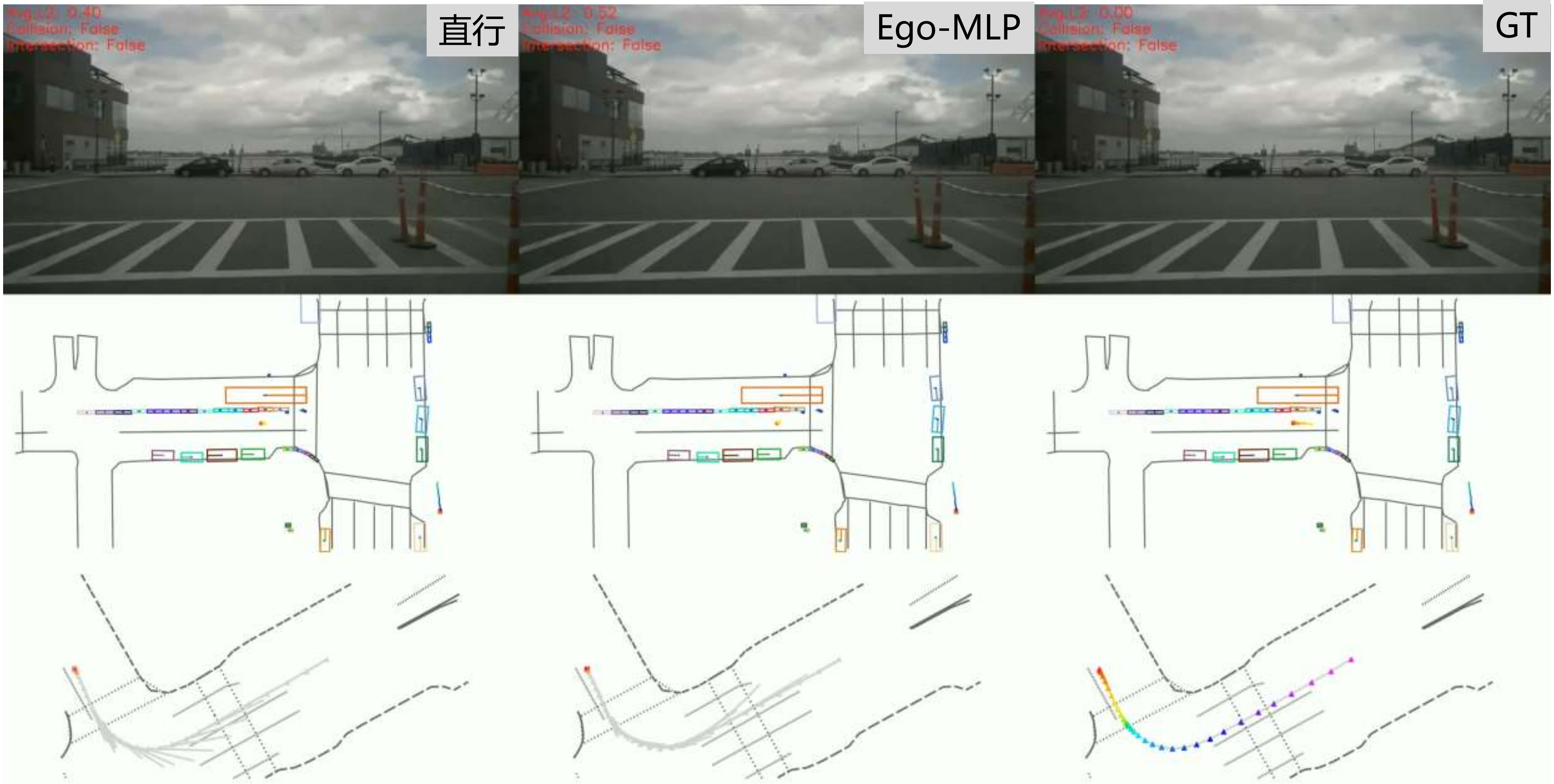
(a) Trajectory Heatmap

(b) Typical Scene of nuScenes

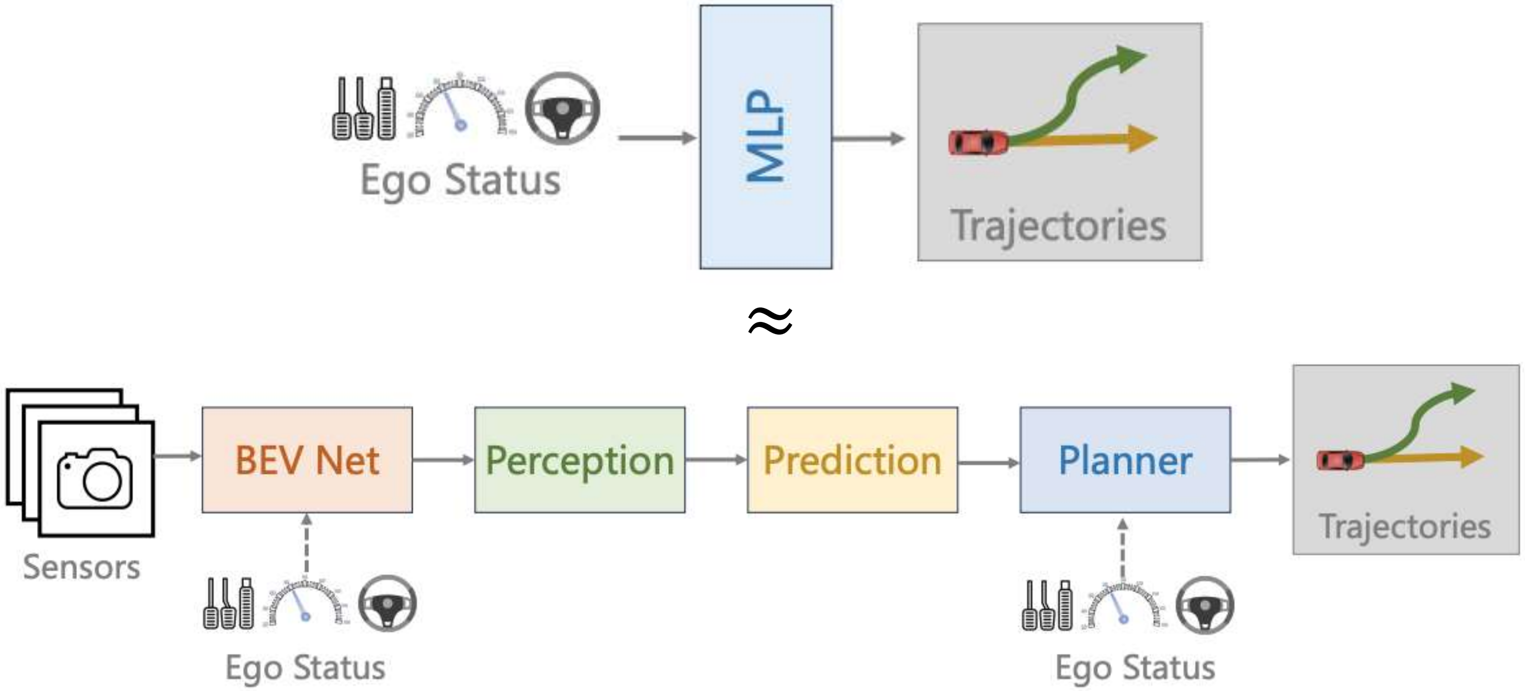
Figure 2. (a) The ego car trajectory heatmap on nuScenes dataset. (b) The majority of the scenes within the nuScenes dataset consist of straightforward driving situations.

NuScenes trajectory distribution is unbalanced, and most straight scenes are too simple.

DEMO



A+B≈A -> B≈ 0

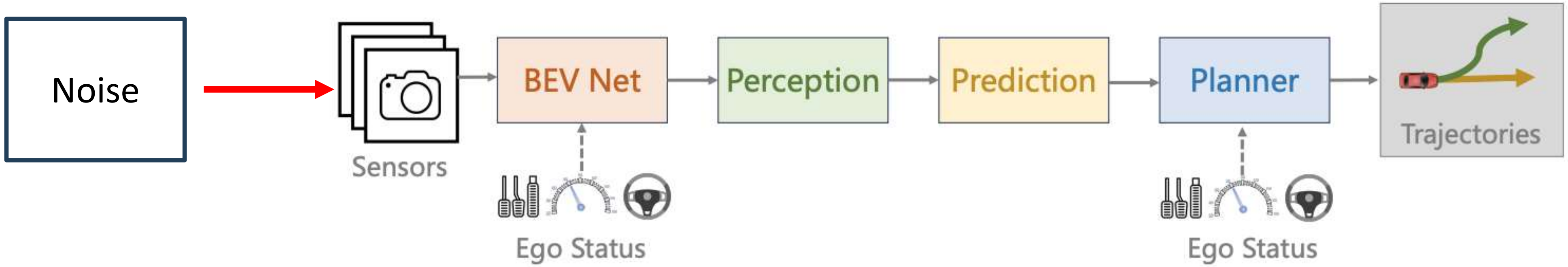


(b) Commonly Used Pipeline of End-to-End Autonomous Driving Model

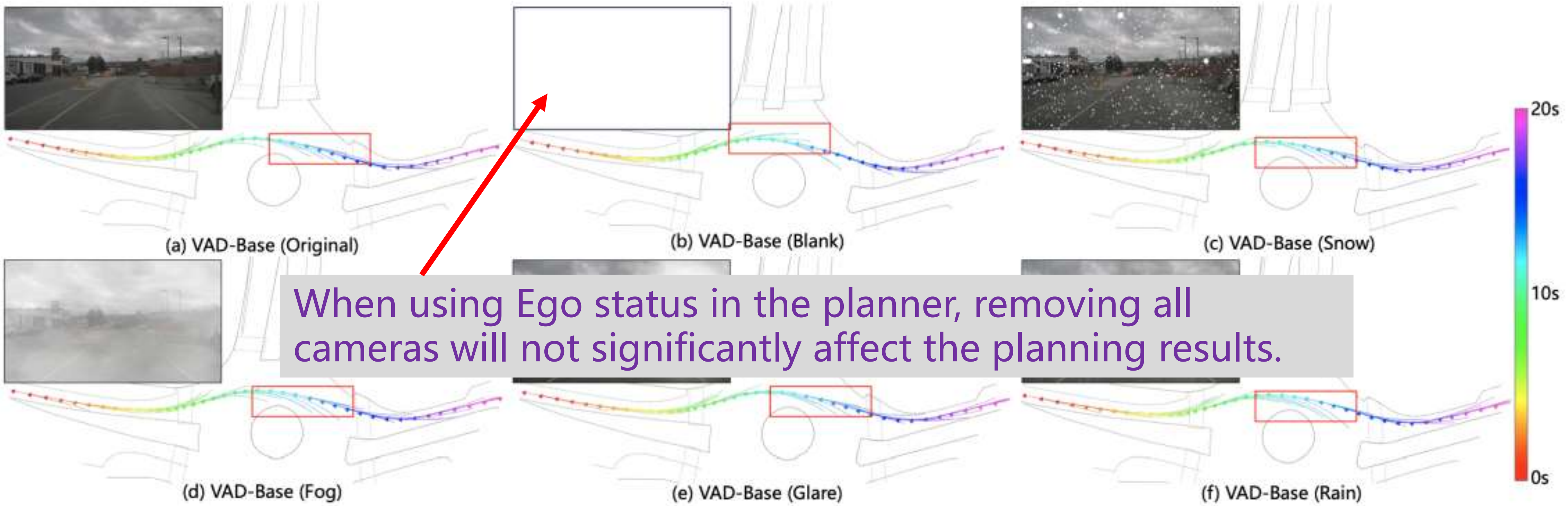
The effect of “Perception+Ego Status is approximately equal to Ego Status，

So what is the role of Perception?

Camera Sensor provides minor valid info

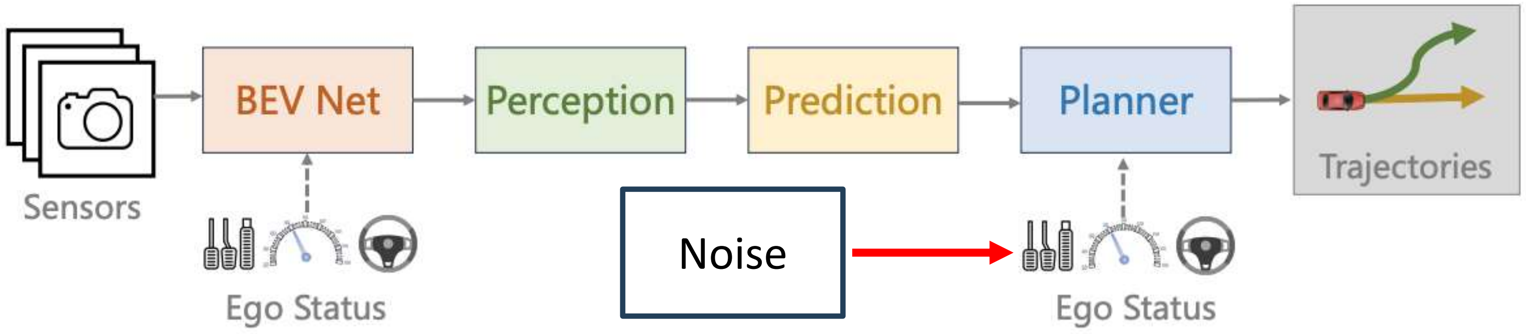


(b) Commonly Used Pipeline of End-to-End Autonomous Driving Model

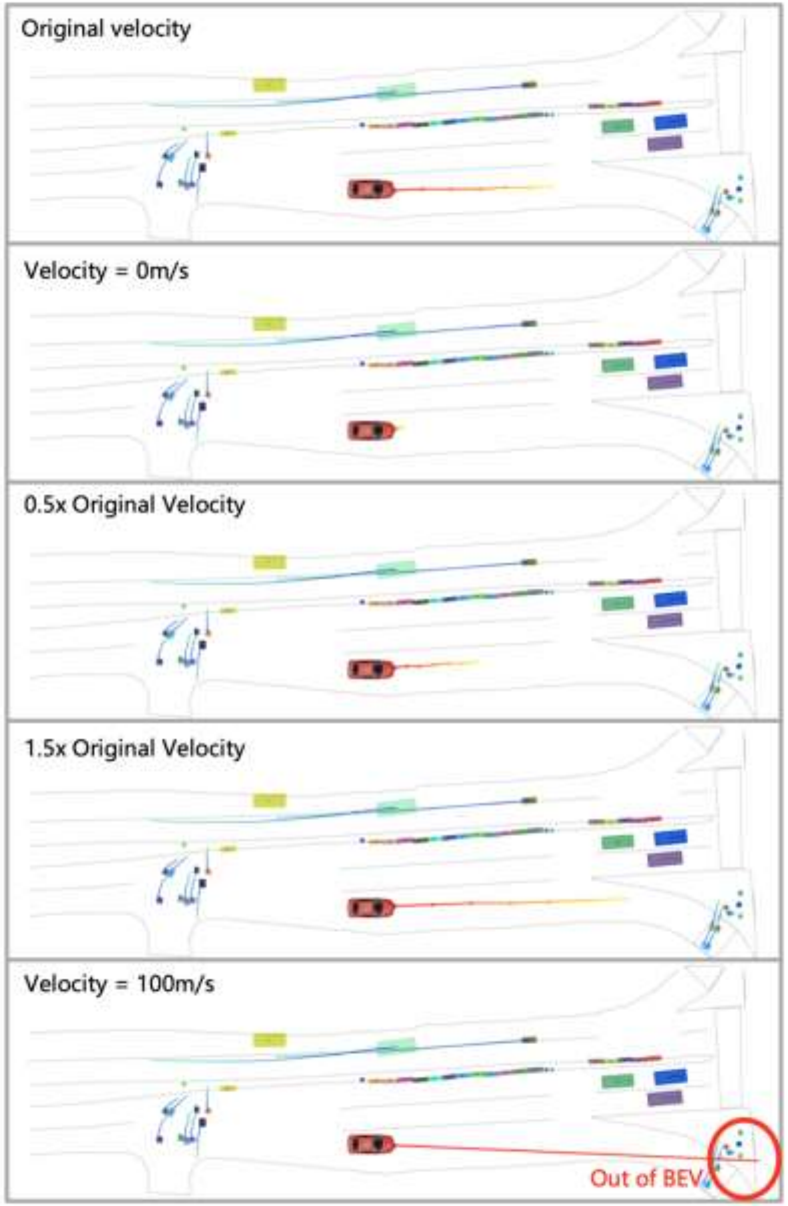


When using Ego status in the planner, removing all cameras will not significantly affect the planning results.

Ego Status Dominates the Plannning



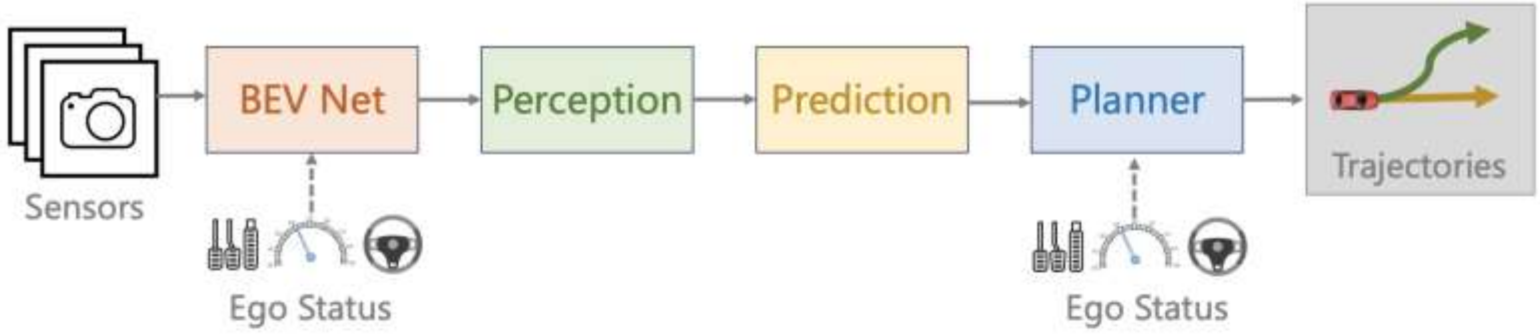
(b) Commonly Used Pipeline of End-to-End Autonomous Driving Model



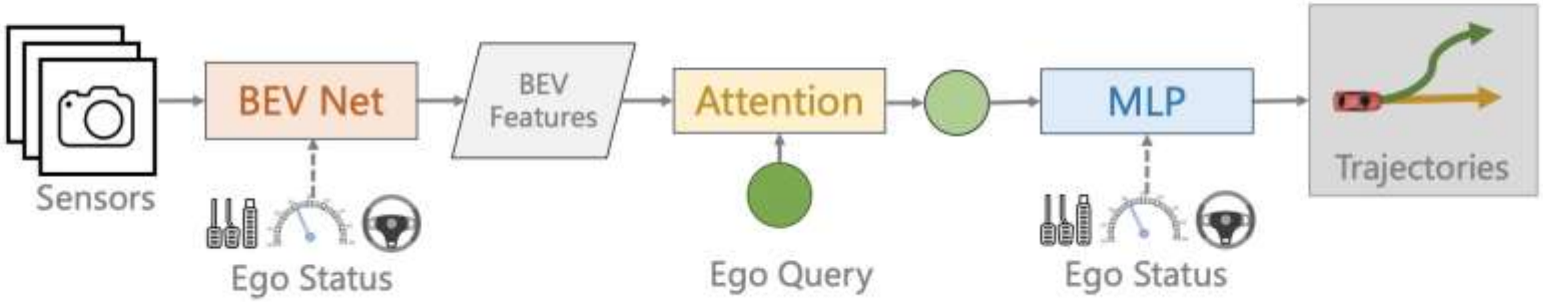
Adding noise to the input velocity will significantly affect the predicted trajectory

BEV-Planner

南京大學
NANJING UNIVERSITY



(b) Commonly Used Pipeline of End-to-End Autonomous Driving Model



(c) Pipeline of Our BEV-Planner

We proposed a very simple BEV-Planner to verify different settings

- Only use one L2 loss
- No using depth, detection, tracking, HD map info.

The Next-Step in Open Loop End2End AD

- 在现有框架上继续追求更好的指标没有意义
- 需要更适合的数据集 （更多样，更复杂）和更全面的评测指标
- 从开环走向闭环： 基于神经渲染的模拟器，世界模型

Thanks