BEV-Guided Multi-Modality Fusion for Driving Perception

Yunze Man Liang-Yan Gui Yu-Xiong Wang
University of Illinois Urbana-Champaign

Check Our Website and GitHub

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

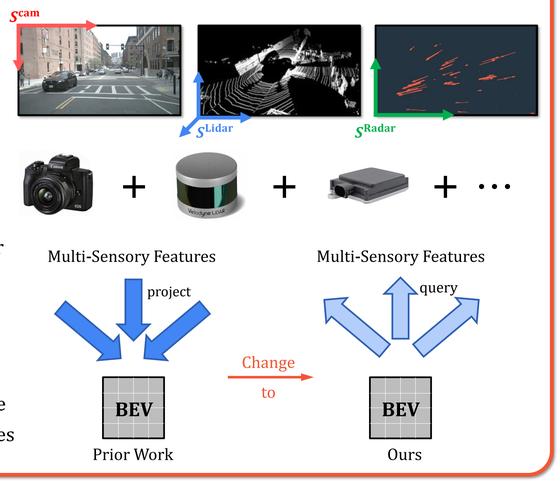
Motivation

Existing fusion strategies:

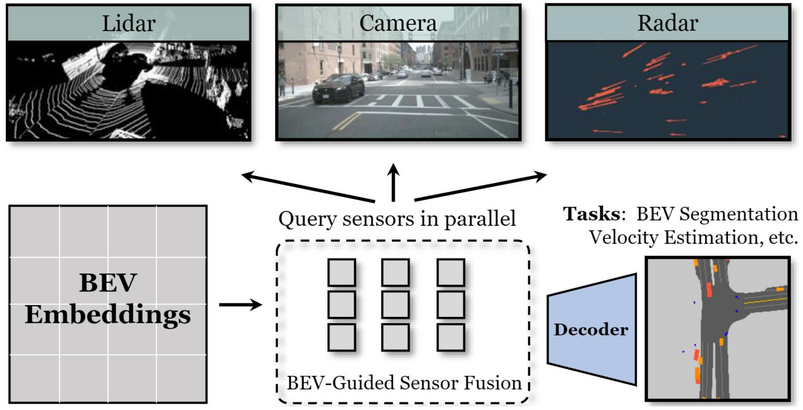
- Support a limited and fixed set of sensors
  - Not flexible with different sensor configurations
- Have ad-hoc sensor fusion designs and cannot adapt to diverse input samples
  - Cannot dynamically adjust weights of different sensors
- Overlook unique properties of Radar sensor

Our method BEVGuide:

- Addressed all the limitations
- Top-down design ⇒ Bottom-up design
- Sensor features project to unified BEV space ⇒ Unified BEV space queries sensor features



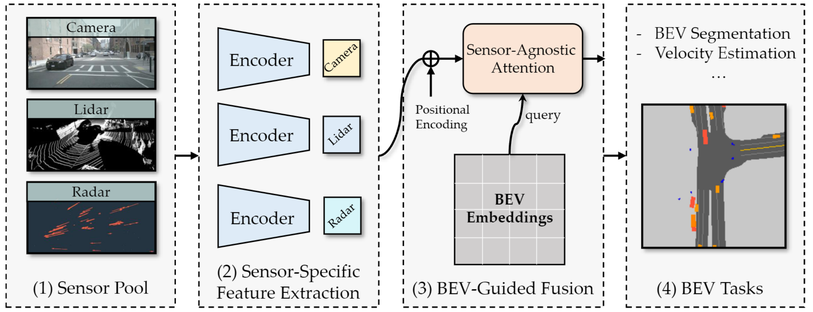
BEVGuide Formulation



Input: Sensory measurements (RGB, Lidar Points, Radar Points, *etc.*)

Output: Bird-eye's-view (BEV) Segmentation, Detection, Velocity Estimation, Prediction, *etc.*

Model Architecture



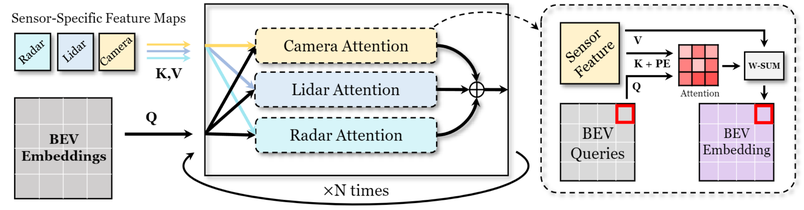
(1) **Input**: A sensor pool – flexible configurations of sensors

(2) **Encoders**: Sensor-specific, frozen or learnable

(3) **Fusion**: Queries from BEV, a sensor-agnostic attention module

(4) **Output**: BEV representation, various perception and prediction tasks

BEV-Guided Sensor Agnostic Attention



Geometric correspondence is encoded through positional encoding

For camera: $x^{(\mathrm{im})} \simeq KMx^{(\mathrm{w})}$, $M^{-1}K^{-1}x^{(\mathrm{im})} \simeq x^{(\mathrm{w})}$ (Image Key, BEV Query)

For Lidar/Radar: $x^{(\mathrm{L/R})} \simeq SX^{(\mathrm{w})}$, $S^{-1}x^{(\mathrm{L/R})} \simeq X^{(\mathrm{w})}$ (Lidar/Radar Key, BEV Query)

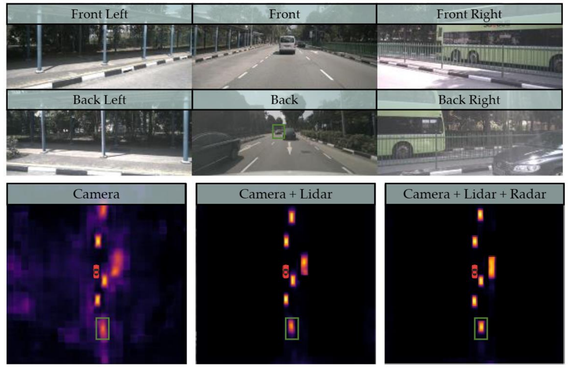
Conclusions

BEVGuide is a comprehensive and versatile multi-modality fusion architecture.

BEVGuide easily adapts to different sensor combinations and is robust to sensor failures.

BEVGuide achieves state-of-the-art performance on various driving tasks

Qualitative & Quantitative Results



Bright regions represent high probability of being vehicle

Segmentation (IoU)	C	R	L	Vehicles	Roads
Cross-view	✓			36.0	74.3
FUTR3D	✓	✓		46.6	-
Simple-BEV	✓	✓		60.8	-
BEVFusion	✓		✓	-	85.5
X-Align	✓		✓	-	86.8
BEVGuide	✓	✓	✓	**79.0**	**86.9**

Detection	C	R	L	mAP	NDS
FUTR3D	✓	✓		35.0	45.9
BEVGuide*	✓	✓		42.1	53.7
BEVFusion	✓		✓	68.5	71.4
BEVGuide*	✓		✓	68.9	71.4
BEVGuide	✓	✓	✓	**69.3**	**71.5**

Prediction	C	R	L	P-IoU
PointPillar	✓		✓	25.3
CenterFusion	✓	✓		55.9
BEVGuide*	✓			21.7
BEVGuide*	✓		✓	32.6
BEVGuide	✓	✓	✓	**67.4**

Velo. Est.	C	R	L	P-AVE
Cross-view	✓			2.13
PointPainting	✓	✓		1.90
BEVGuide	✓	✓	✓	**0.81**

BEVGuide achieves leading performance on diverse driving tasks.

Day → Night	C	R	L	Day	Night	Gap
Cross-view	✓			40.4	18.8	21.6
BEVGuide*	✓		✓	76.7	58.8	17.9
BEVGuide	✓	✓	✓	**79.5**	**64.2**	**15.3**

Sunny → Rainy	C	R	L	Day	Night	Gap
Cross-view	✓			37.3	28.1	9.2
BEVGuide*	✓		✓	77.0	69.9	7.1
BEVGuide	✓	✓	✓	**80.7**	**74.6**	**6.1**

Multi-modality reduces domain gap, increases robustness