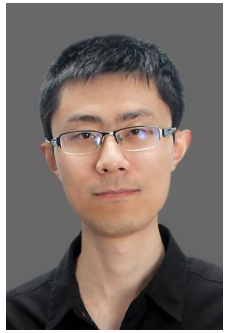
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

BEV-Guided Multi-Modality Fusion for Driving Perception

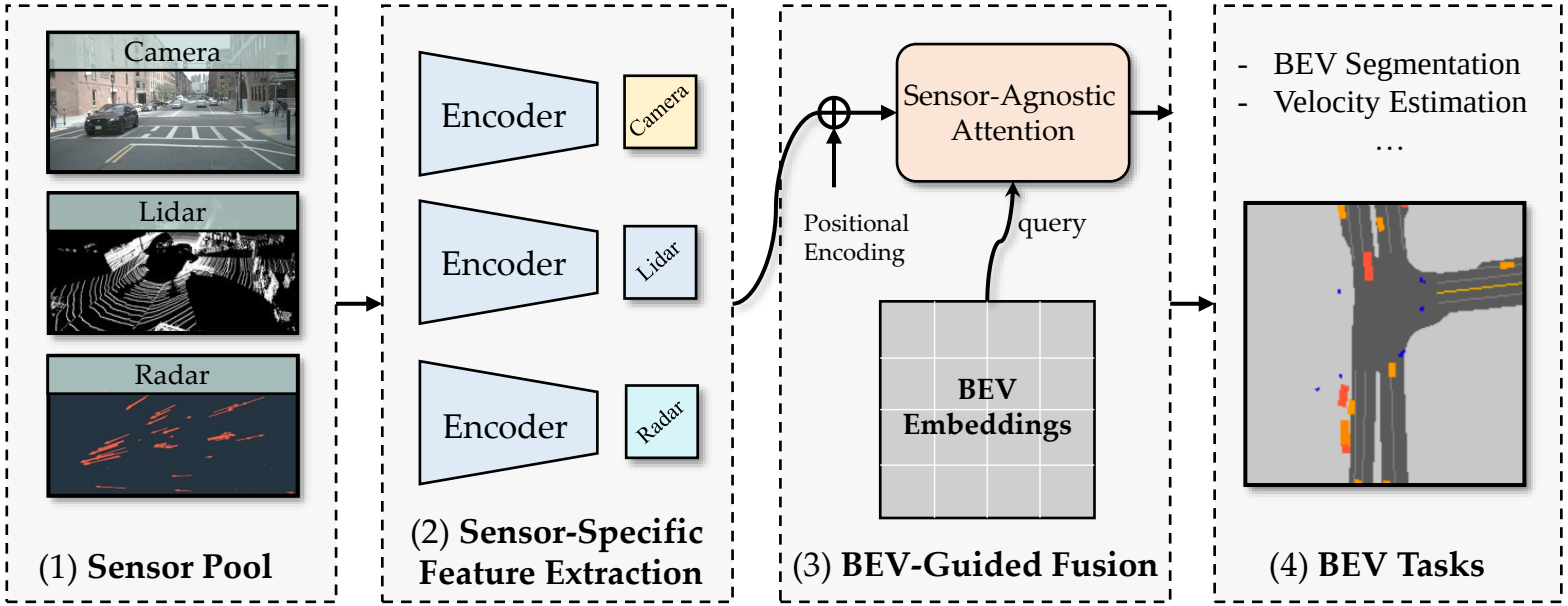
Yunze Man Liang-Yan Gui Yu-Xiong Wang



UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

THU-PM-129

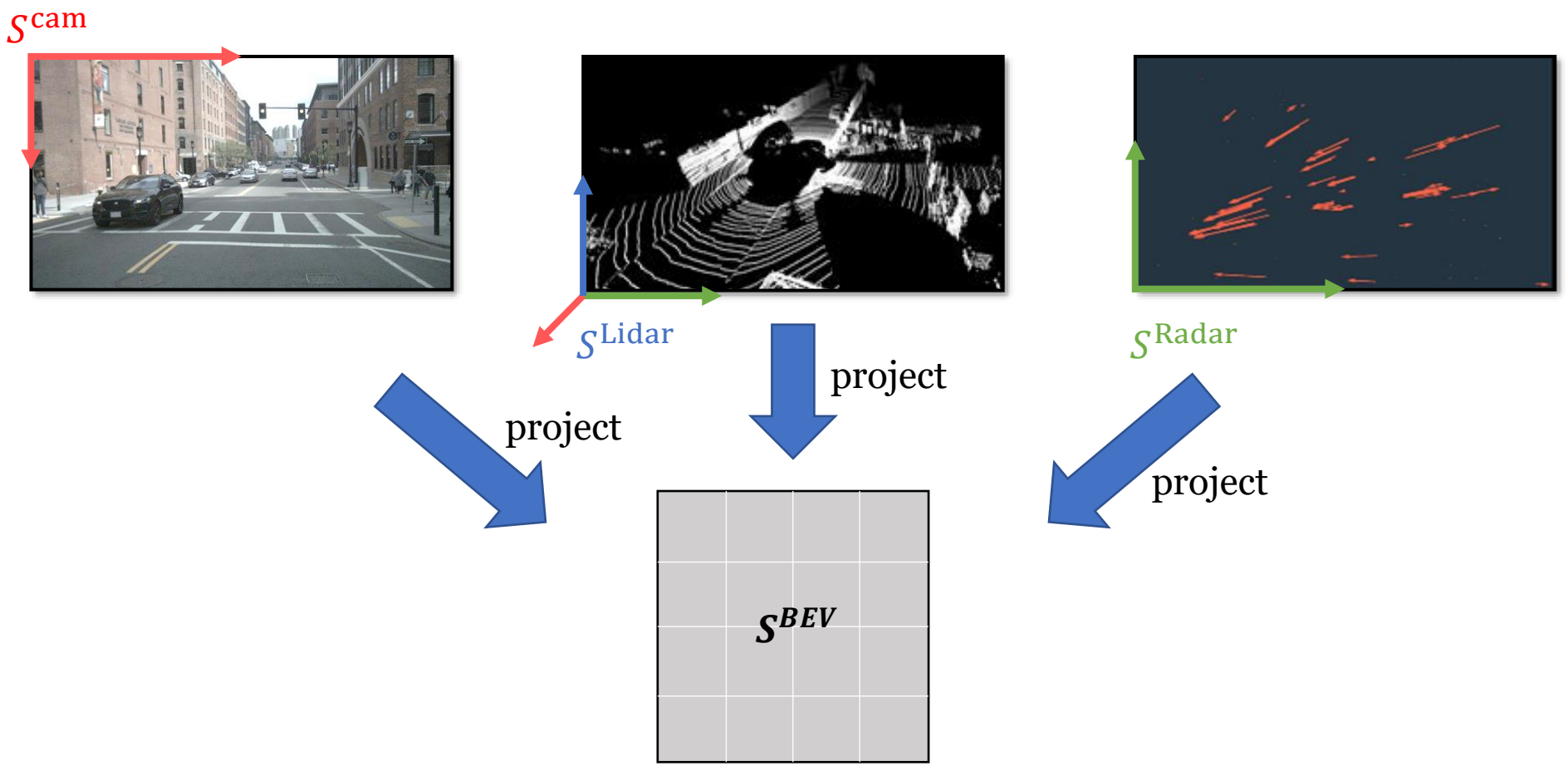
Overview



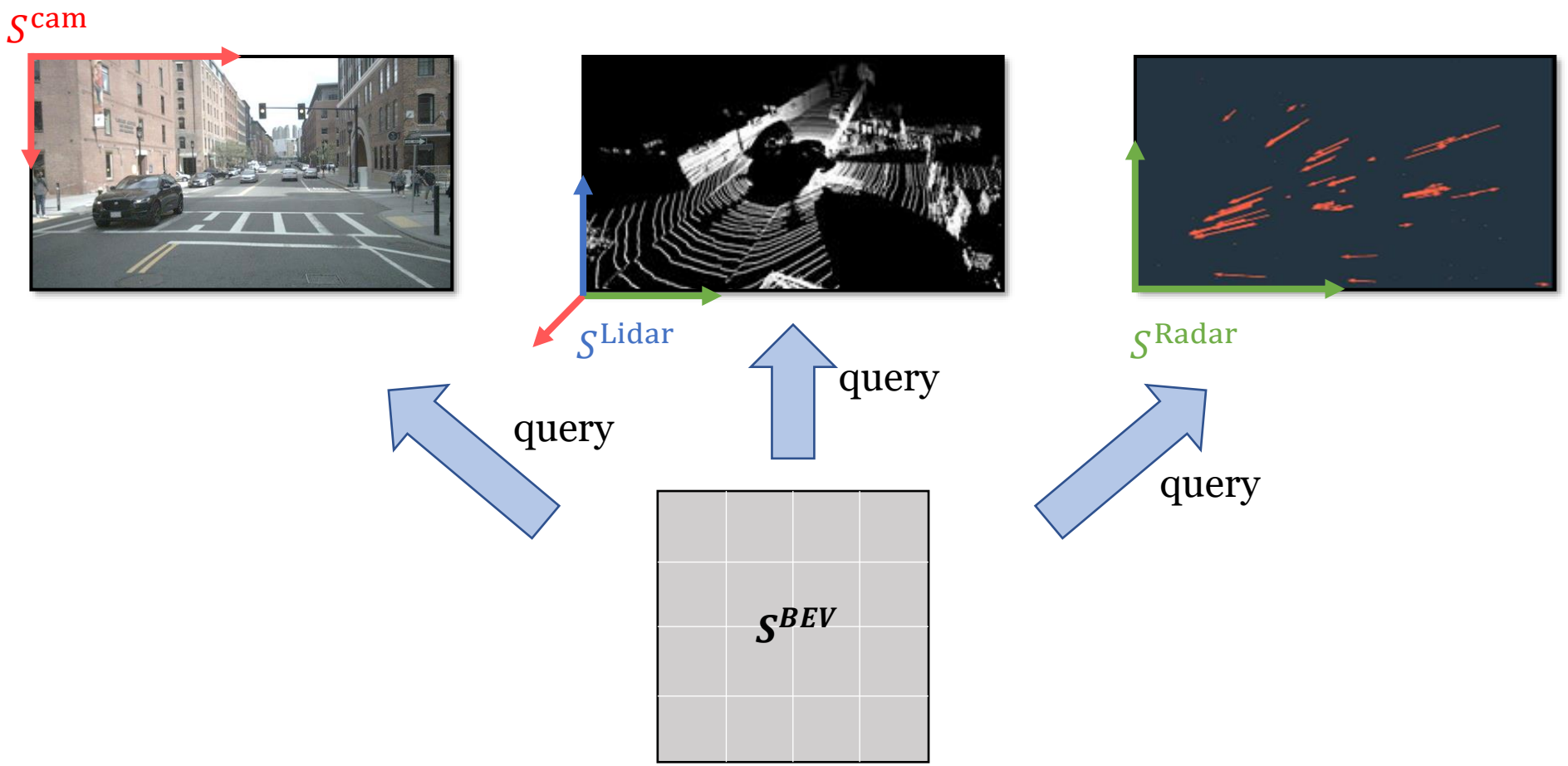
BEVGuide

Comprehensive and versatile multi-modality fusion architecture for driving tasks

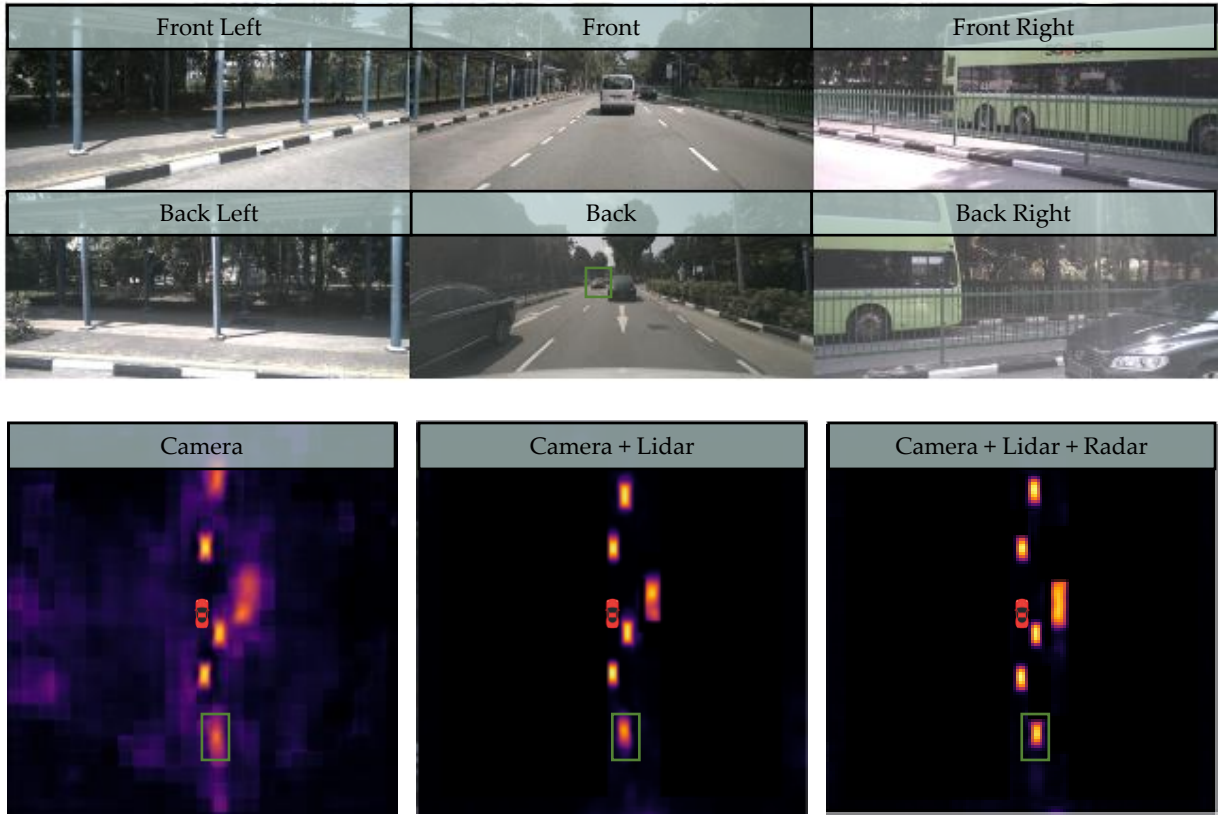
Previous Fusion Methods: Project Sensors into the Unified Space



Our Key Insight: Query Sensors from the Unified Space



Superior Performance on Segmentation



Superior Performance on Diverse Driving Tasks: Segmentation, Detection, and Velocity Estimation

Segmentation	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**

Velo. Estimation	C	R	L	P-AVE ↓
Cross-view	✓			2.13
PointPainting	✓	✓		1.90
BEVGuide*	✓		✓	1.63
BEVGuide	✓	✓	✓	**0.81**

Summary

- BEVGuide, a comprehensive and versatile multi-modality fusion architecture
- Easily adapt to different sensor combinations
- Achieved state-of-the-art performance on various driving tasks

Check Our Project Website

https://yunzeman.github.io/BEVGuide/

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Multi-Modality is Everywhere

Music videos: Video + Audio

Communication : Video + Language + Audio + *etc.*

Reading: Language + Video + *etc.*

Multi-Modality is Everywhere

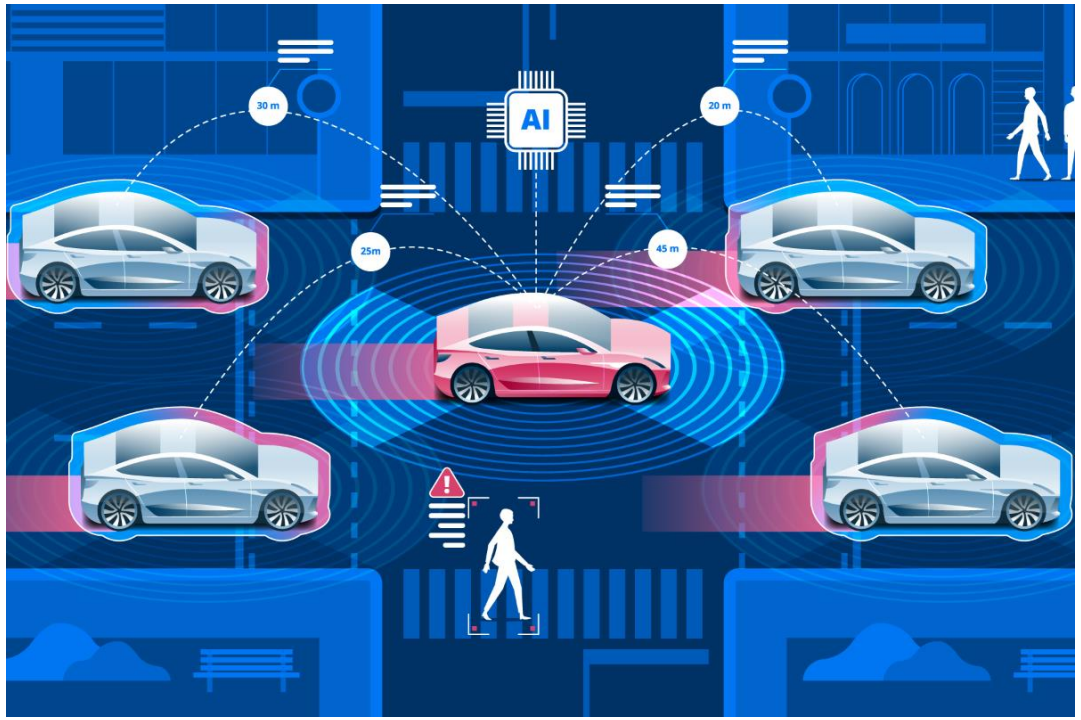
Music videos: Video + Audio

Communication : Video + Language + Audio + *etc.*

Reading: Language + Video + *etc.*

Multi-Modality:
A more complete and diverse experience

Multi-Modality in Driving Scenarios



Camera

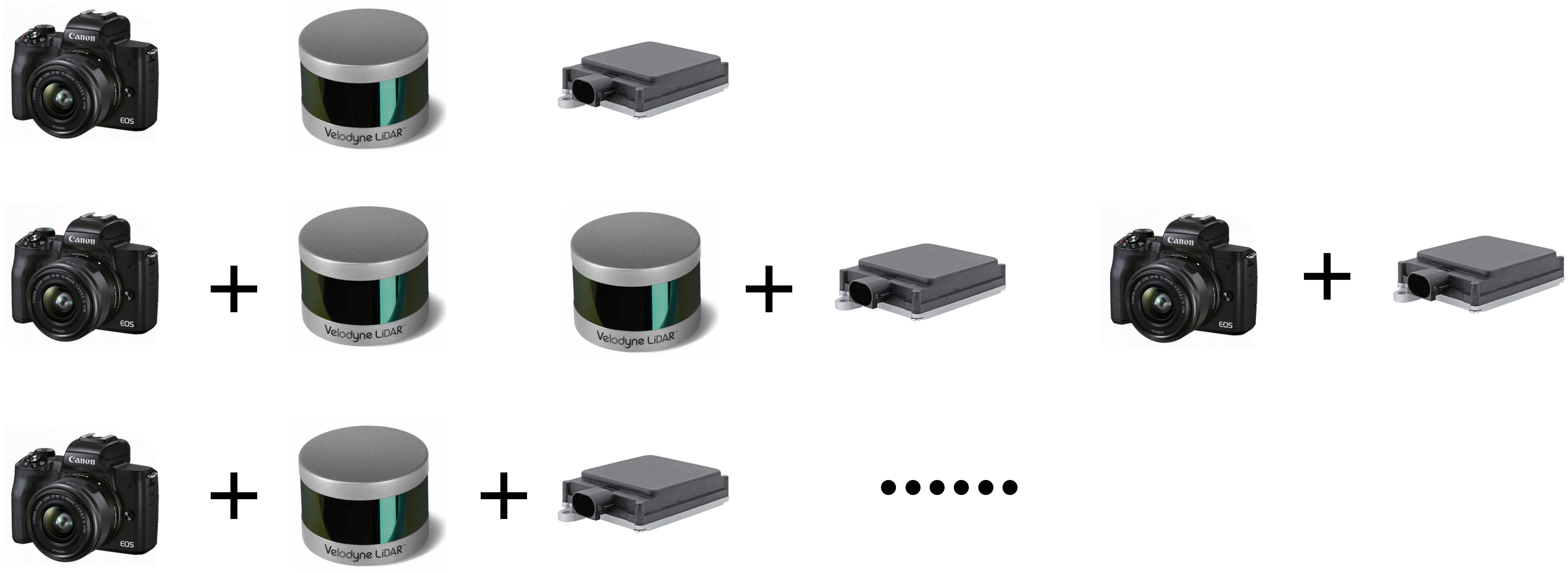
Lidar

Radar

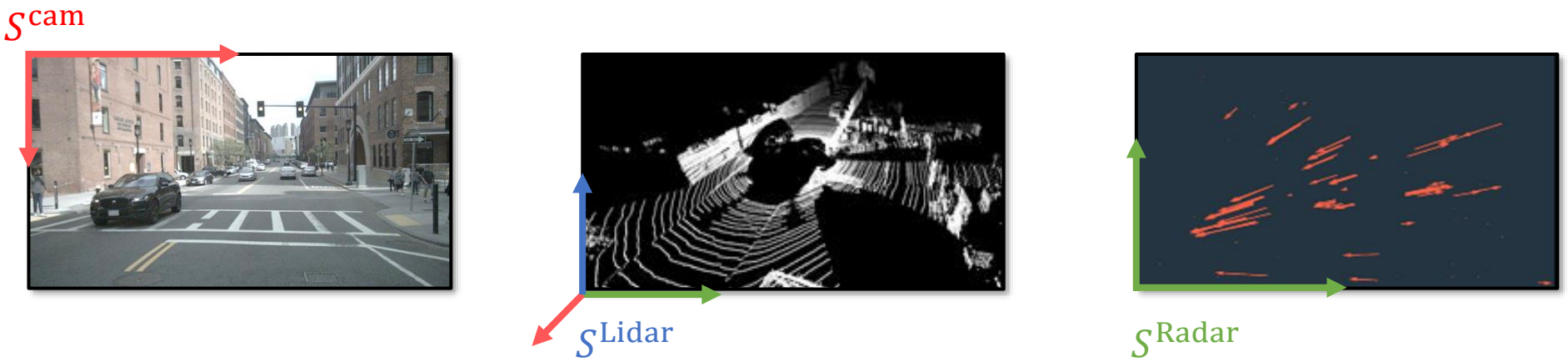
Motivation of BEVGuide (Limitation of Prior Work)

- Inflexible design for different sensors (Limited and fixed)
- Do not dynamically change weights of different sensors based on diverse input sample
- Overlook Radar sensor and its unique properties

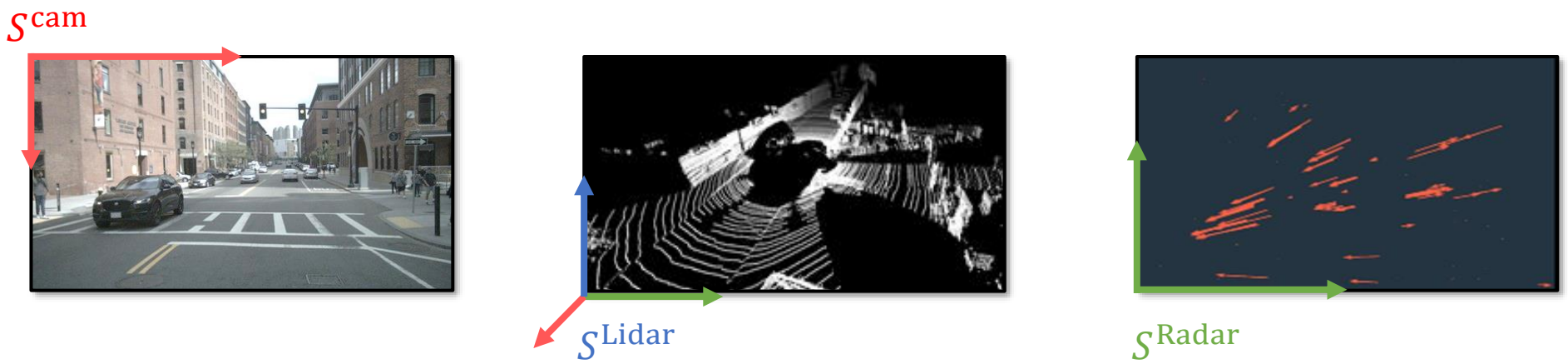
Flexible Sensor Combination



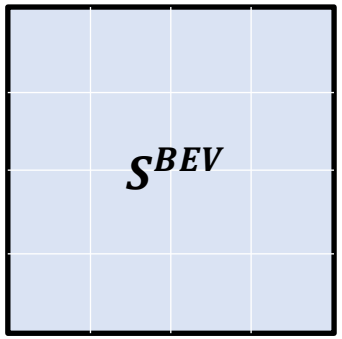
Different sensors are natively represented in different coordinates



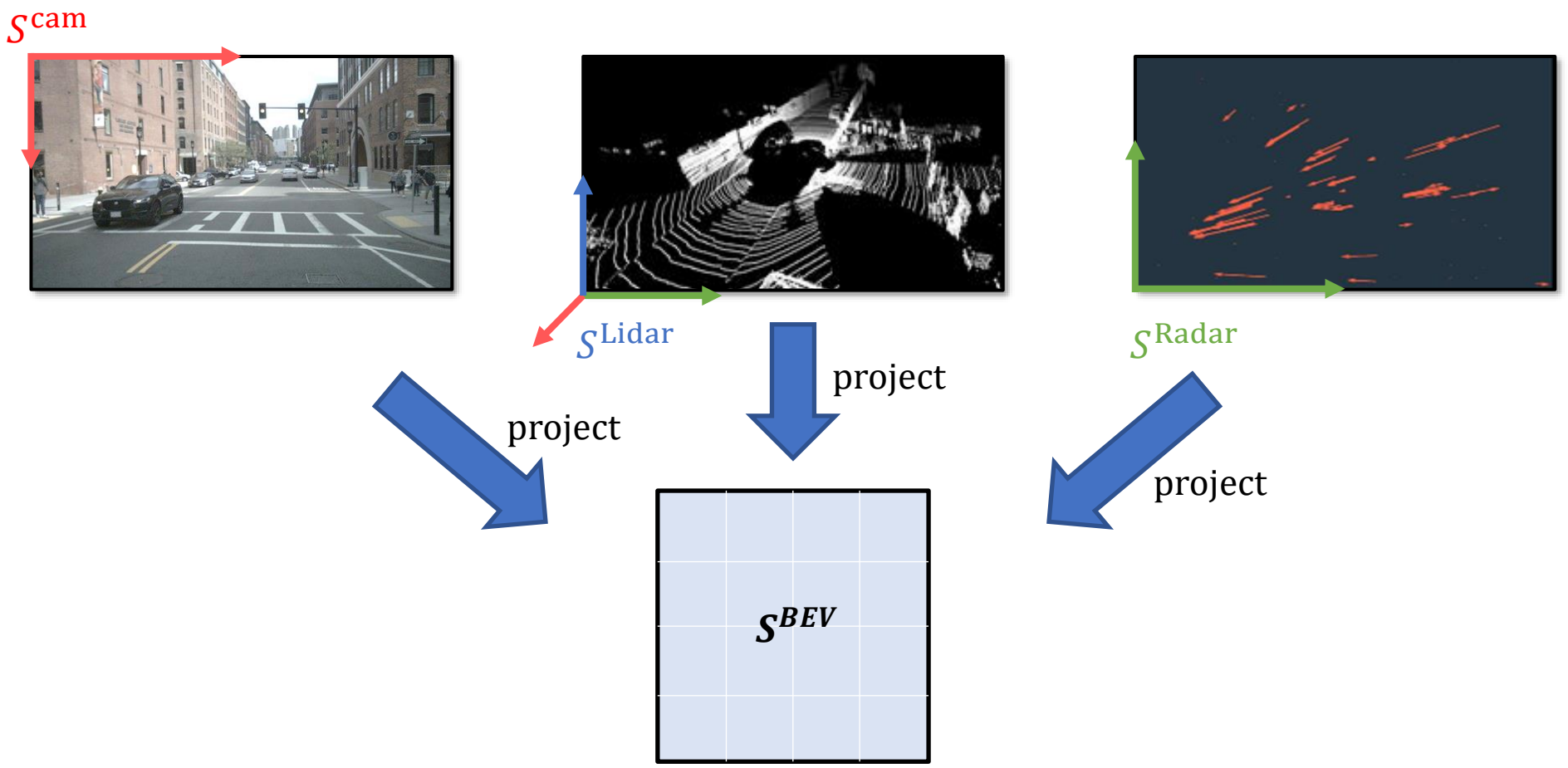
Bridge coordinate systems with BEV



- Almost everything stands and moves on the ground plane
- Simplify 3D into 2D
- BEV is an expressive and concise coordinate system



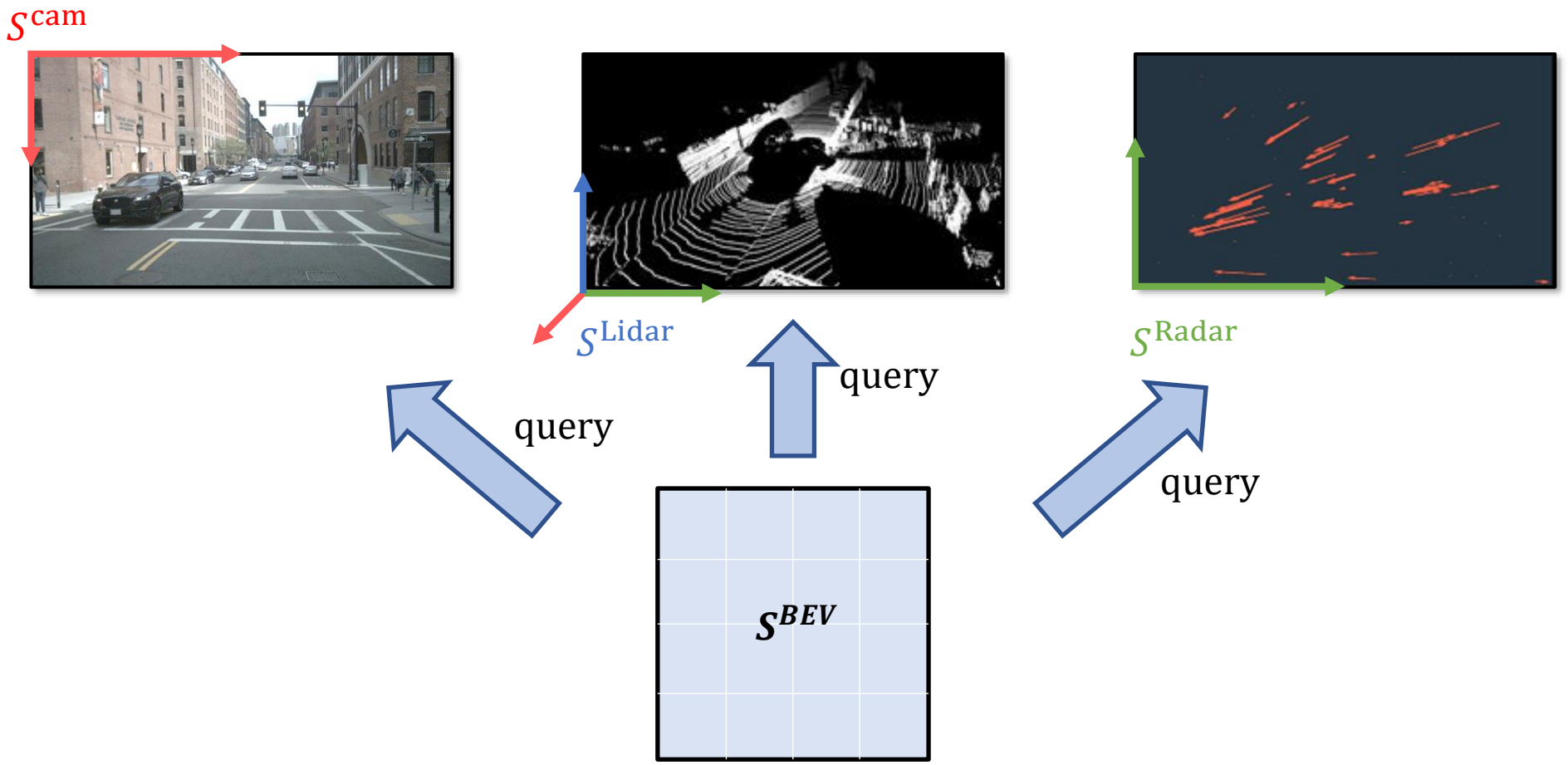
Previous Fusion Methods: Project Sensors into the Unified Space



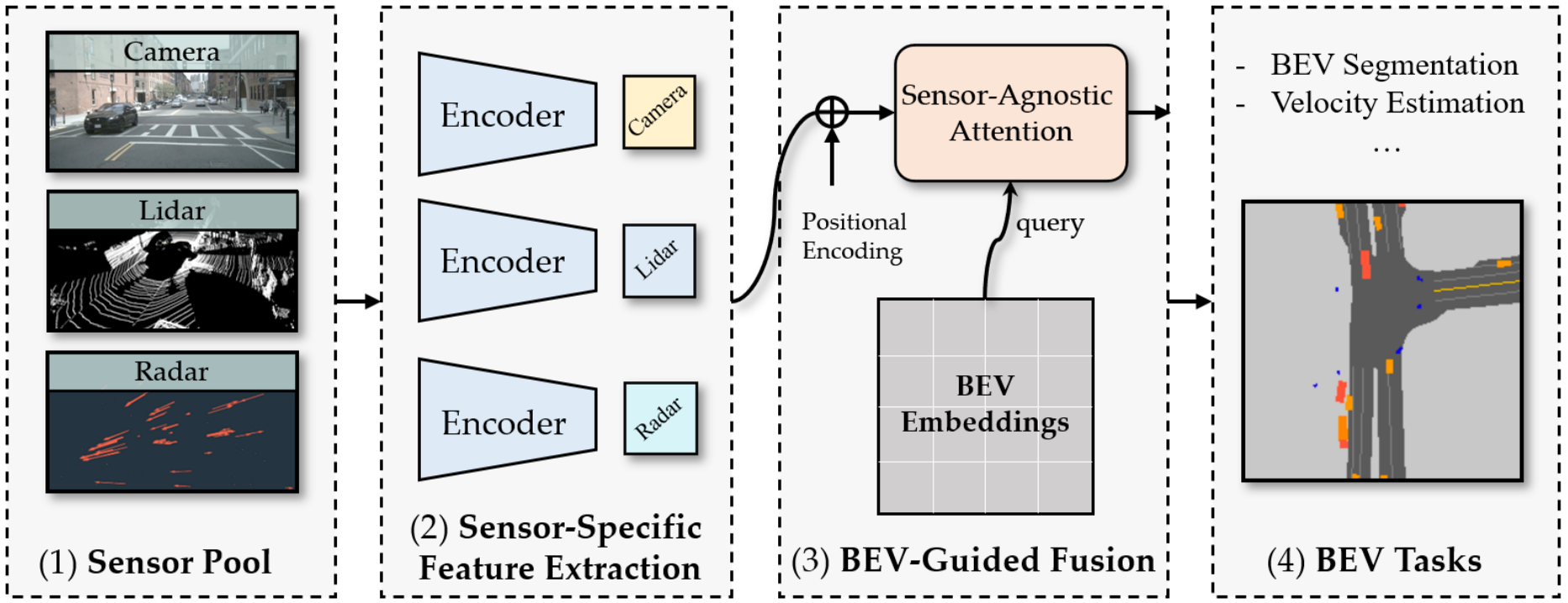
Our Key Insight: Query Sensors from the Unified Space

Allow model to decide the weights of different sensors

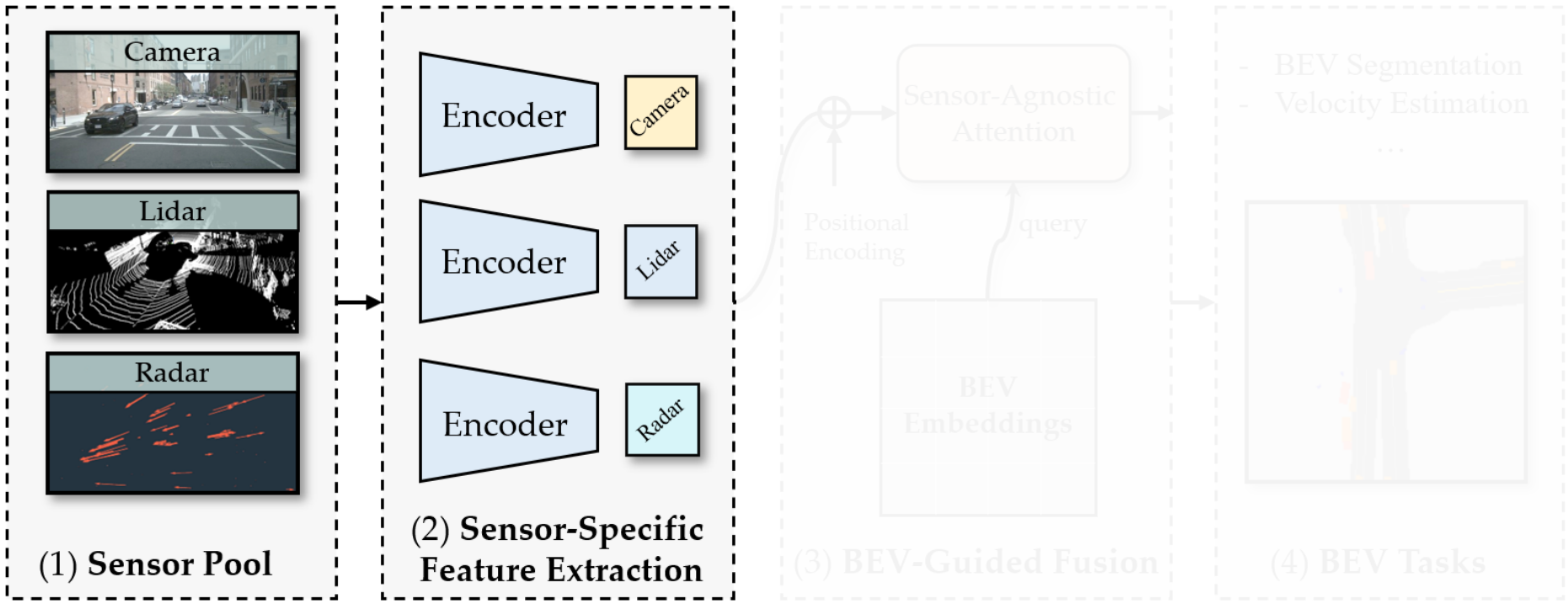
Build the representation bottom-up from BEV



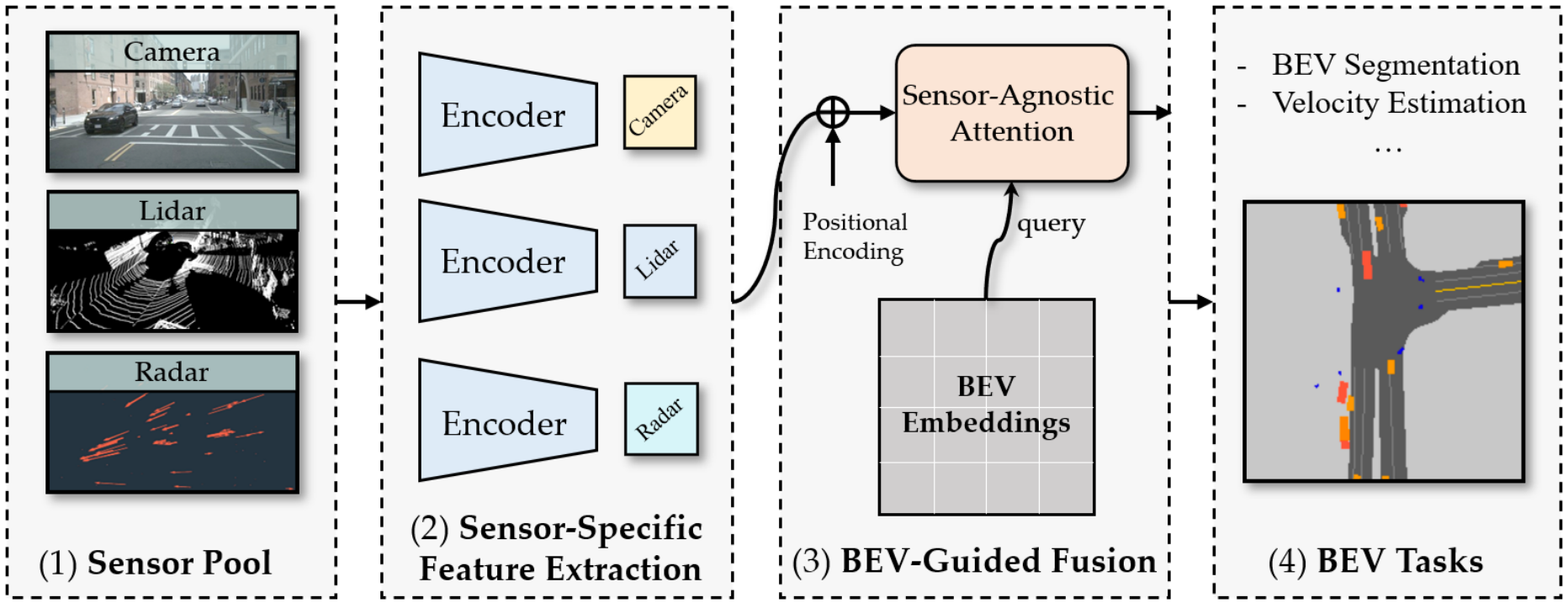
Pipeline of BEVGuide



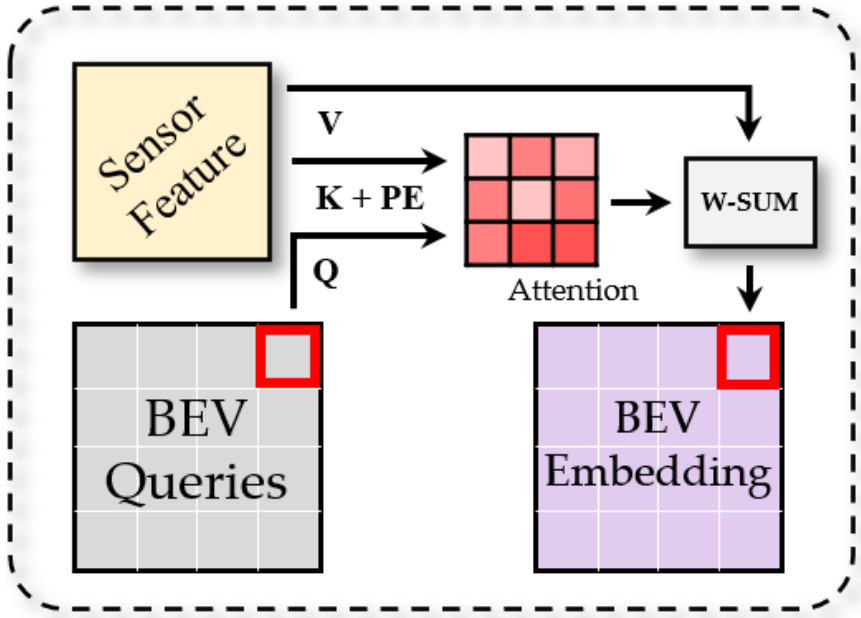
Sensor Pool and Sensor-Specific Encoders



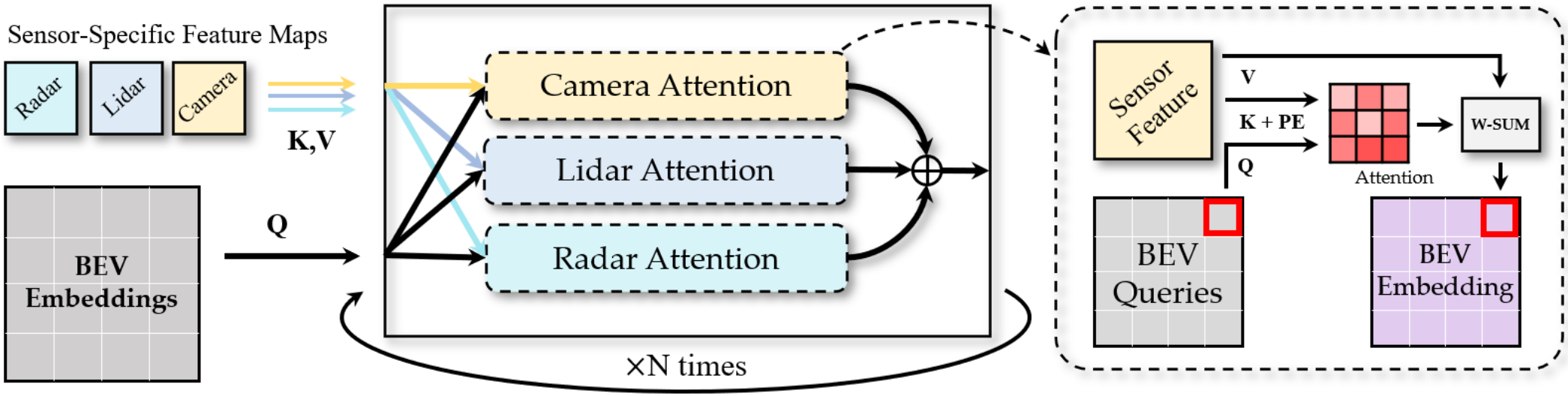
BEV-Guided Sensor-Agnostic Attention



BEV-Guided Sensor Agnostic Attention



BEV-Guided Sensor Agnostic Attention



Geometric Correspondence as Positional Embedding

- For Camera,

$$x^{(\mathrm{im})} \simeq KMX^{(\mathrm{w})}$$

$$M^{-1}K^{-1}x^{(\mathrm{im})} \simeq X^{(\mathrm{w})}$$

Image Key BEV Query

K: intrinsics

M: extrinsics (pose)

Source: Cross-View-Transformer (CVT)

Geometric Correspondence as Positional Embedding

- 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

$$x^{(\mathrm{L/R})} \simeq SX^{(\mathrm{w})}$$

$$S^{-1}x^{(\mathrm{L/R})} \simeq X^{(\mathrm{w})}$$

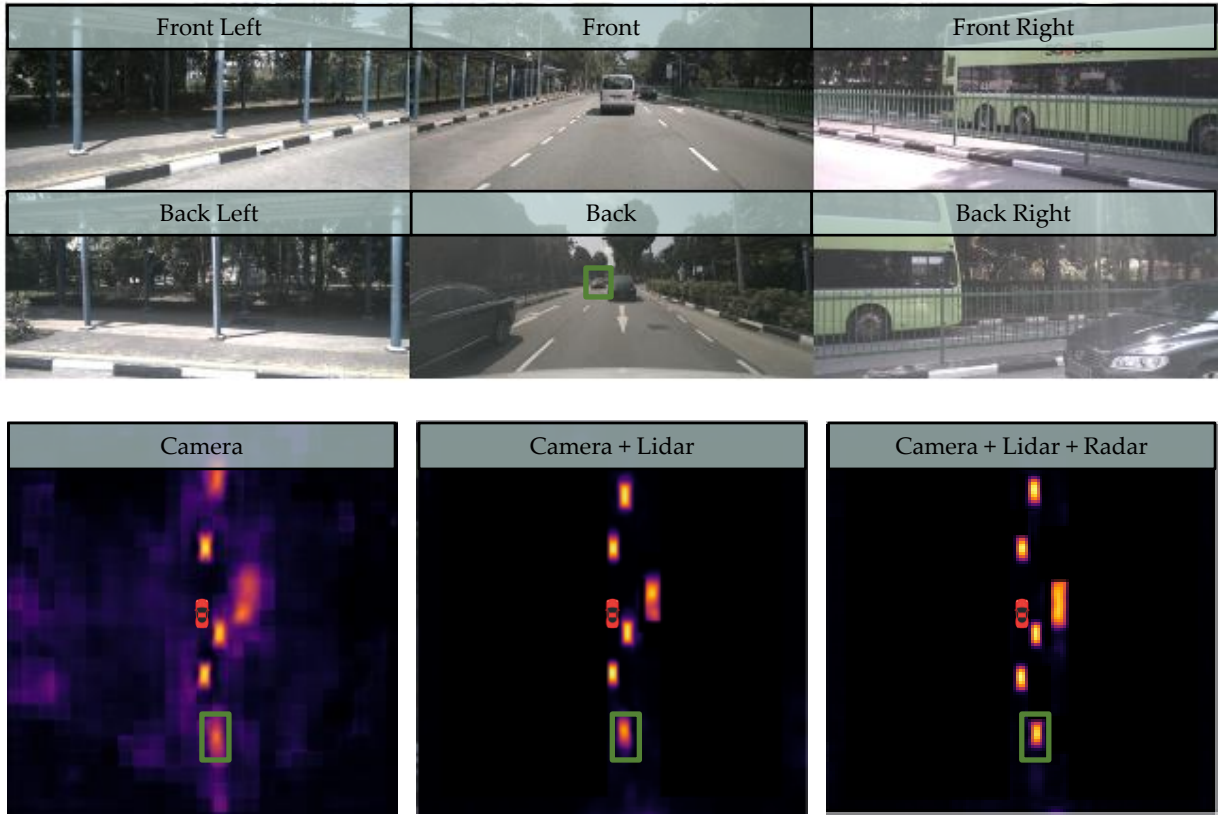
Lidar/Radar Key BEV Query

K: intrinsics

M: extrinsics (pose)

S: scale & translate matrix

Superior Performance on Segmentation



Superior Performance on Diverse Driving Tasks: Segmentation, Detection, and Velocity Estimation

Segmentation	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**

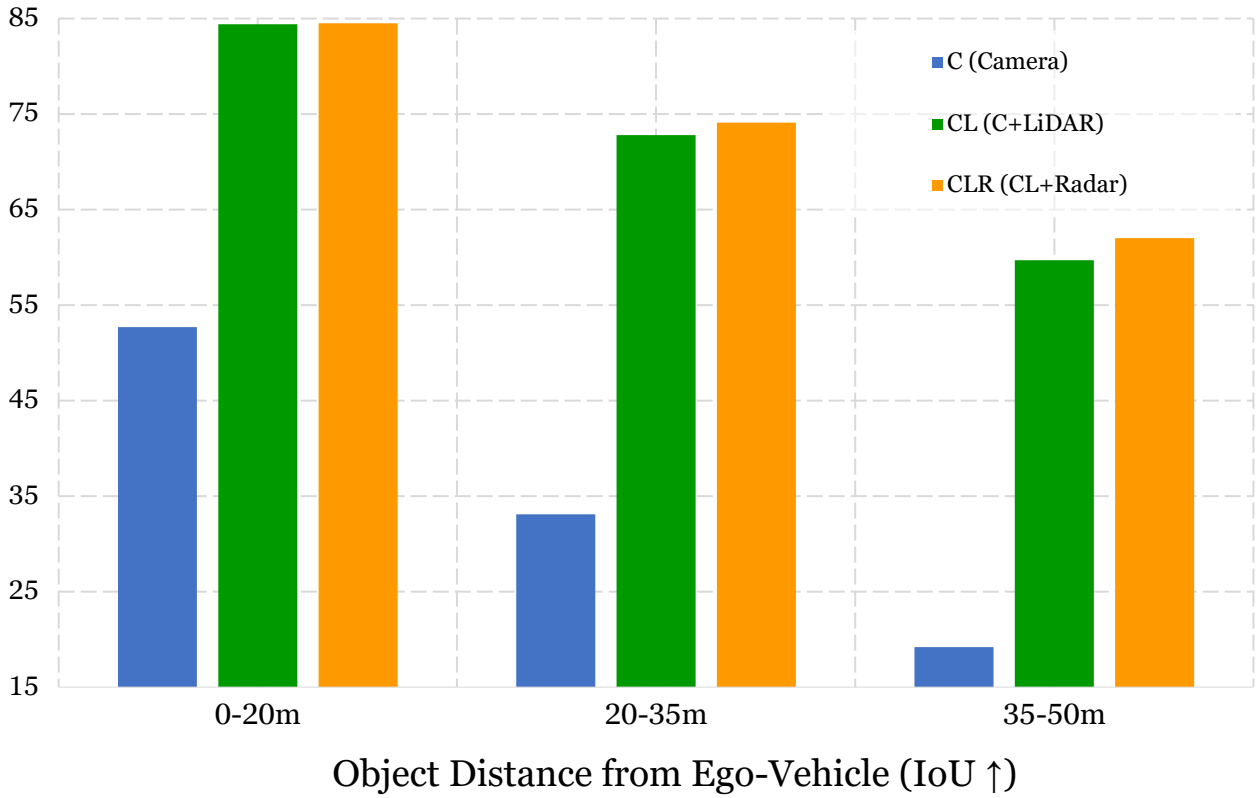
Velo. Estimation	C	R	L	P-AVE
Cross-view	✓			2.13
PointPainting	✓	✓		1.90
BEVGuide*	✓		✓	1.63
BEVGuide	✓	✓	✓	**0.81**

Multi-Modality Reduces Domain Gap, and Increases Robustness

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**

Radar Sensor Helps Improve the Perception of More Distant Objects



Summary

- BEVGuide, a comprehensive and versatile multi-modality fusion architecture
- Easily adapt to different sensor combinations
- Achieve state-of-the-art performance on various driving tasks

Check Our Project Website

https://yunzeman.github.io/BEVGuide/

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN