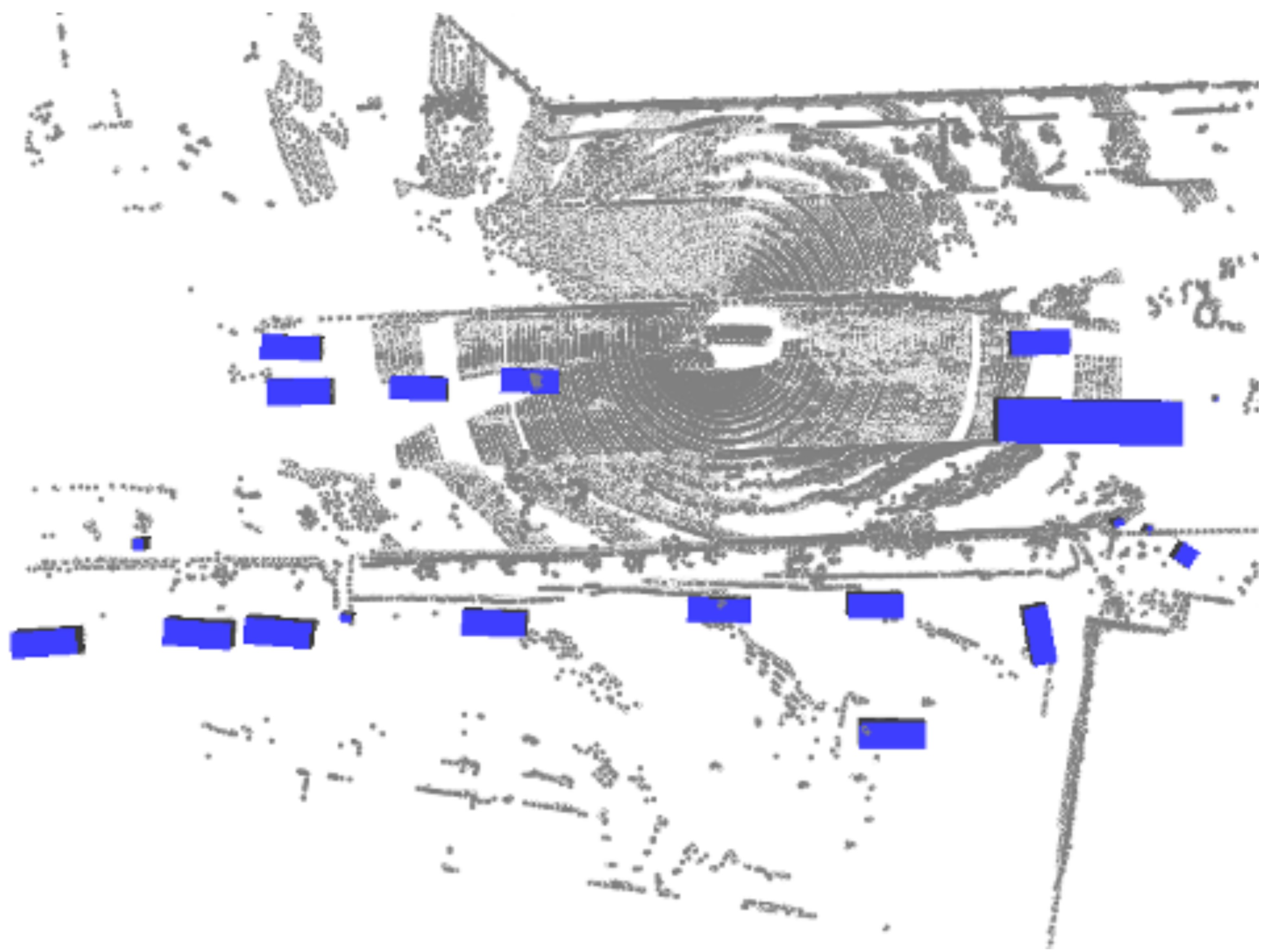
BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework

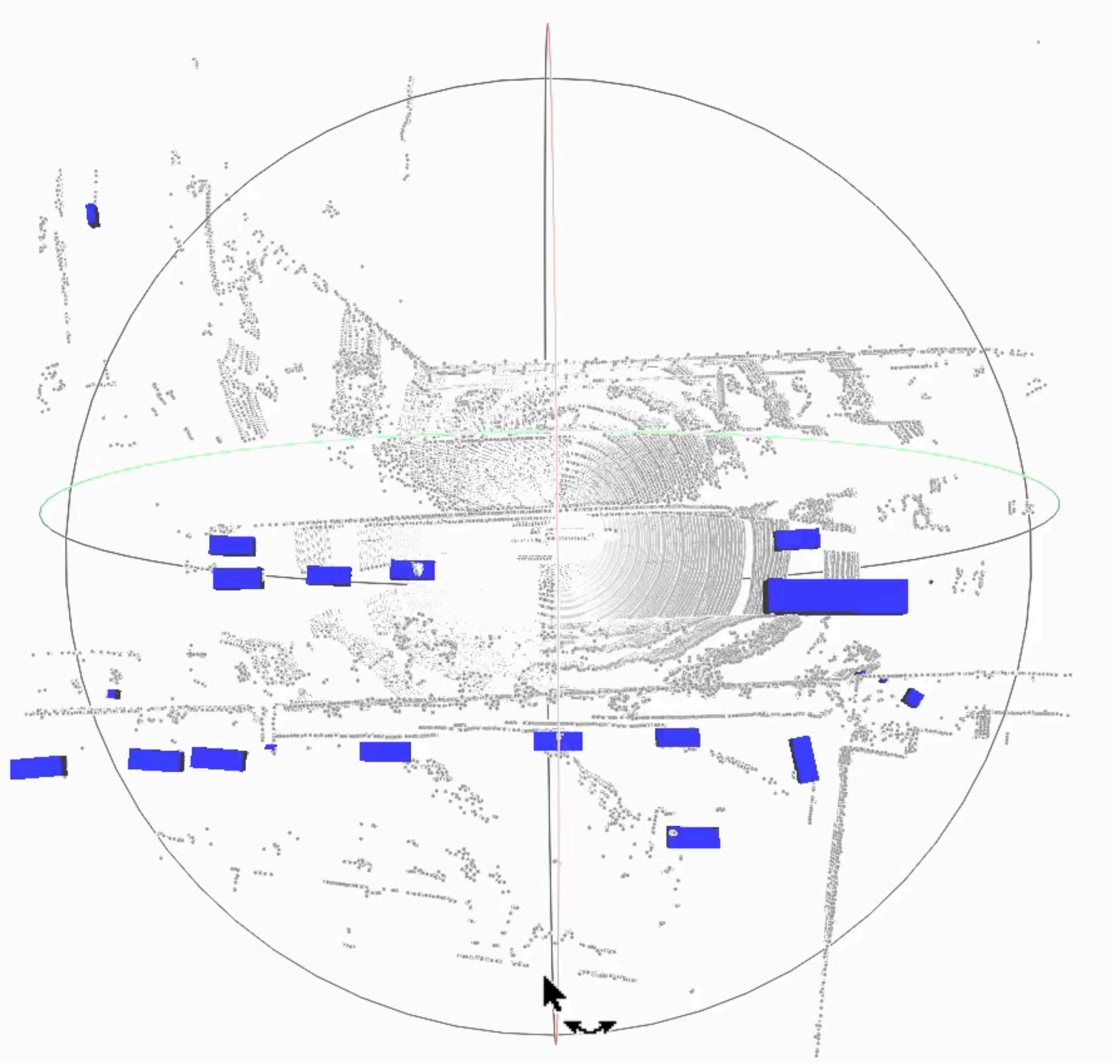
Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, Zhi Tang
NeurIPS 2022 | Peking University, DAMO Academy,

3D Object Detection

In autonomous driving



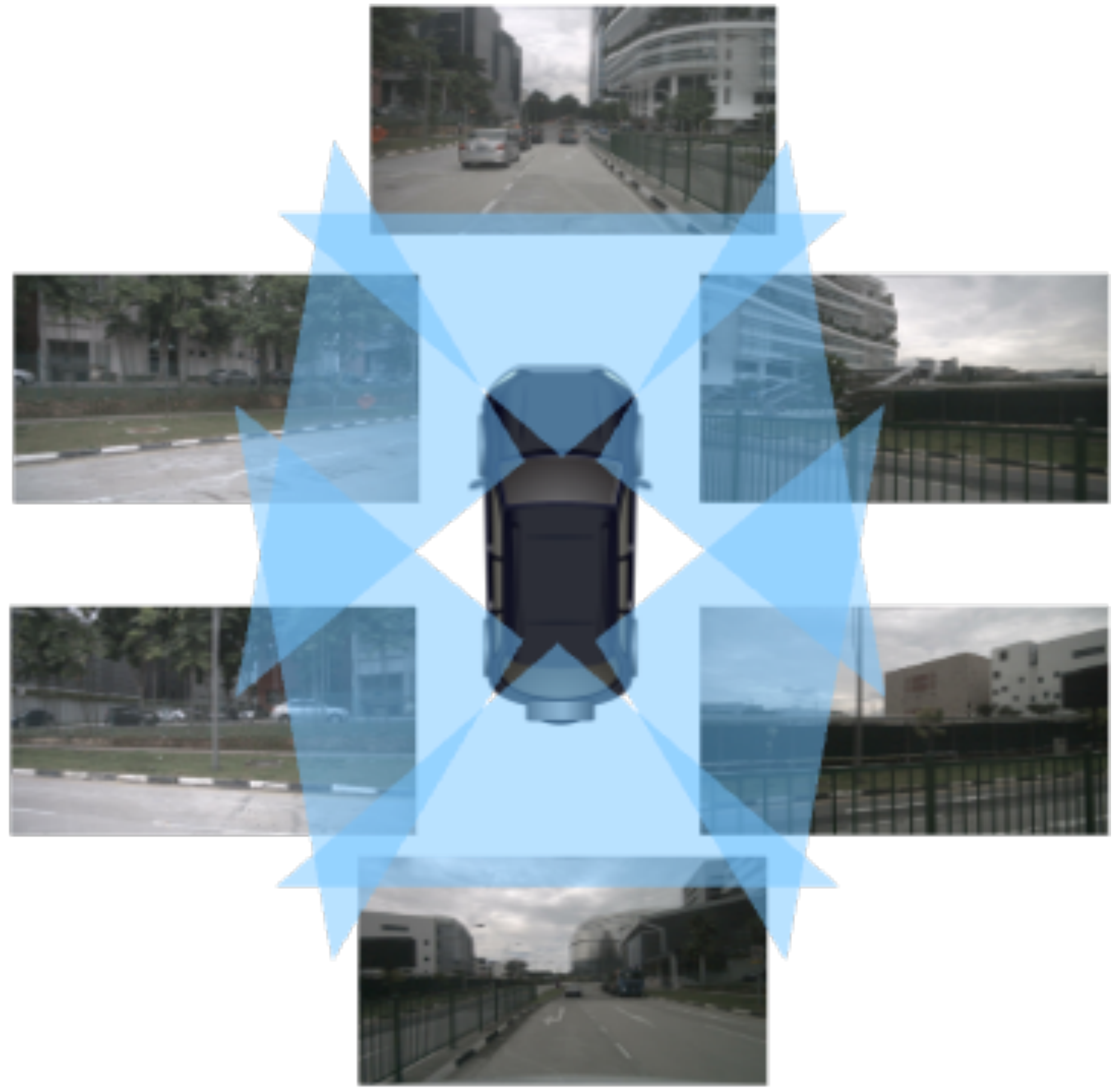
Bird's-Eye-View
(BEV)



Necessary of Fusing Camera & LiDAR

Camera-only:

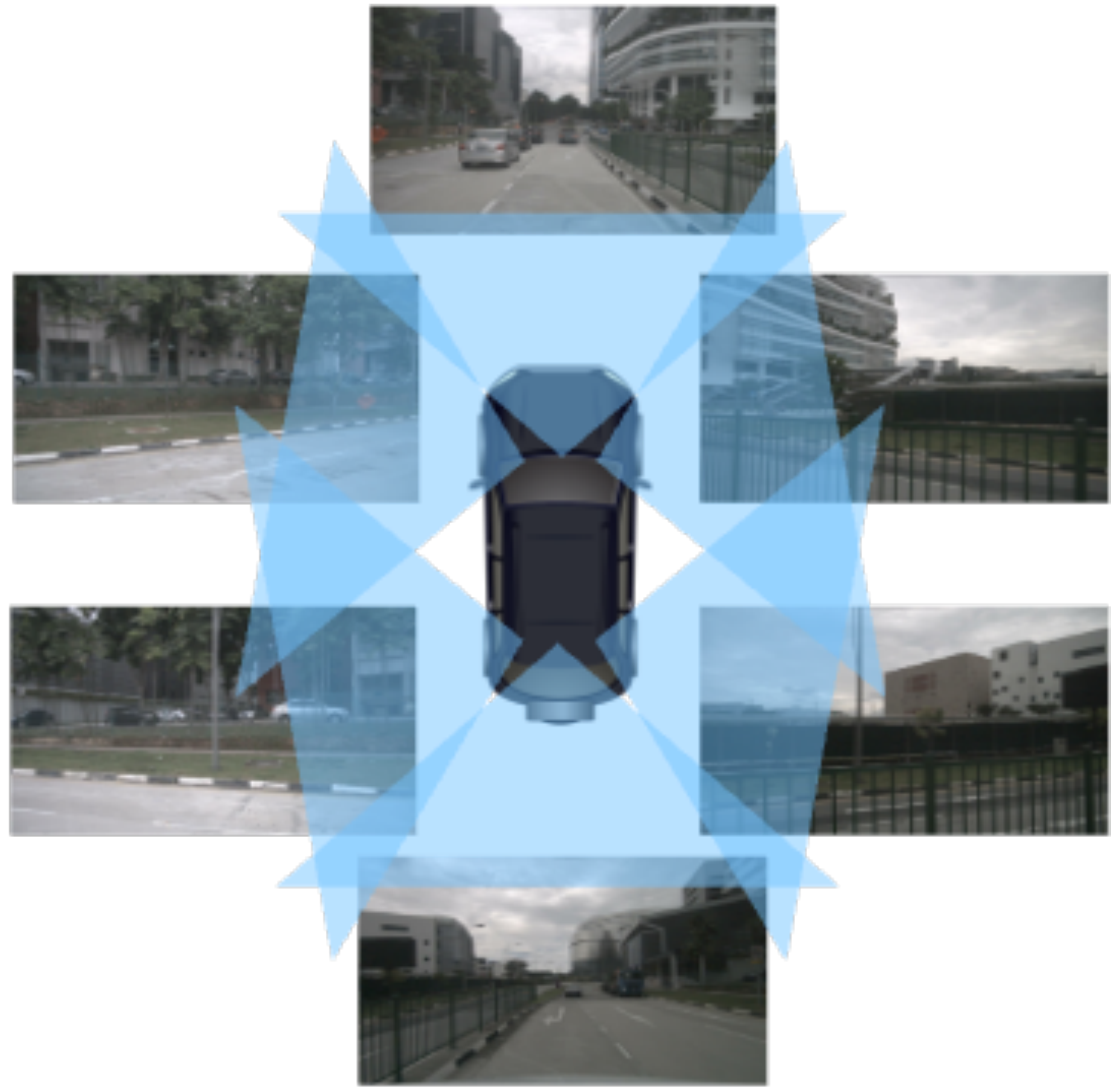
lack of depth information



Difficult to regress
3D bounding boxes

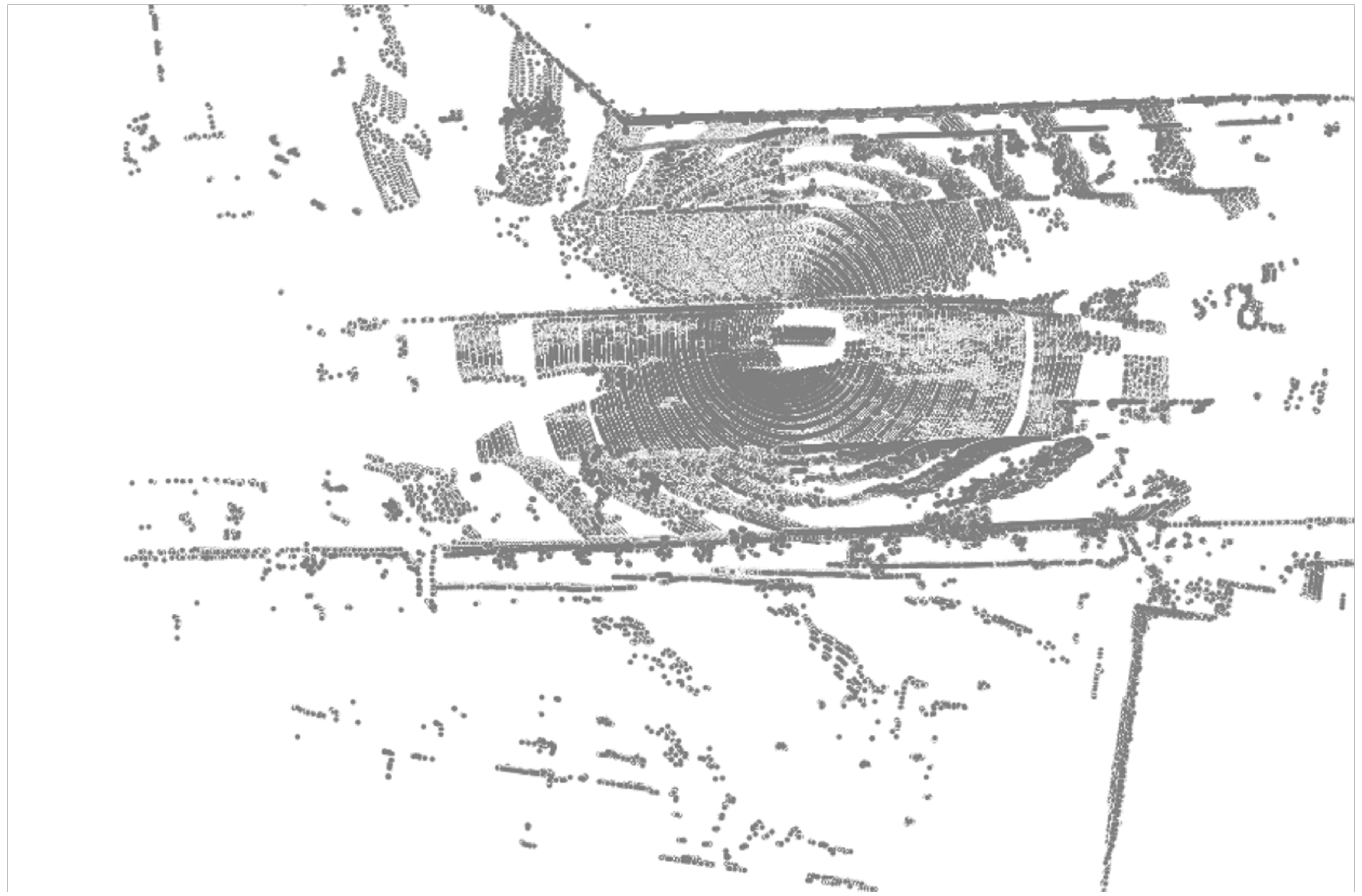
Necessary of Fusing Camera & LiDAR

Camera-only:
lack of depth information



Difficult to regress
3D bounding boxes

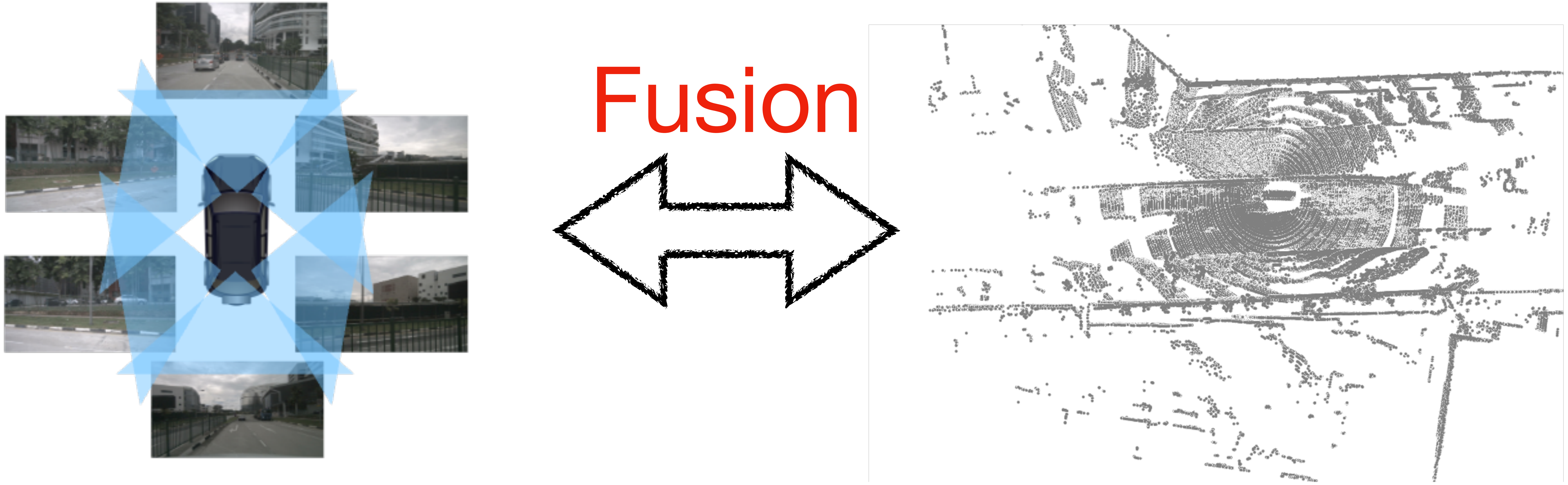
LiDAR-only:
lack of semantic information



Difficult to classify
objects

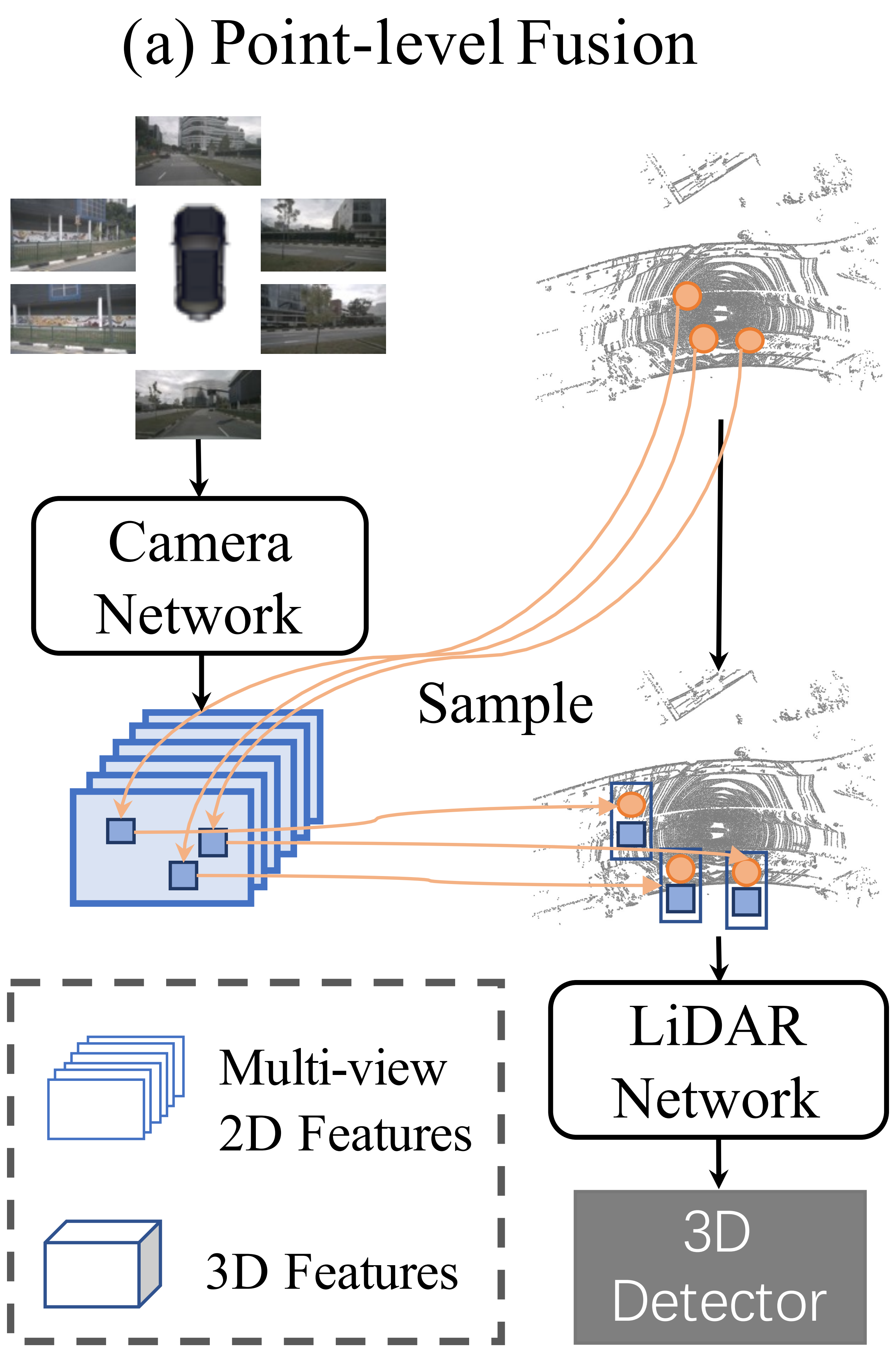
Necessary of Fusing Camera & LiDAR

LiDAR-Camera Fusion



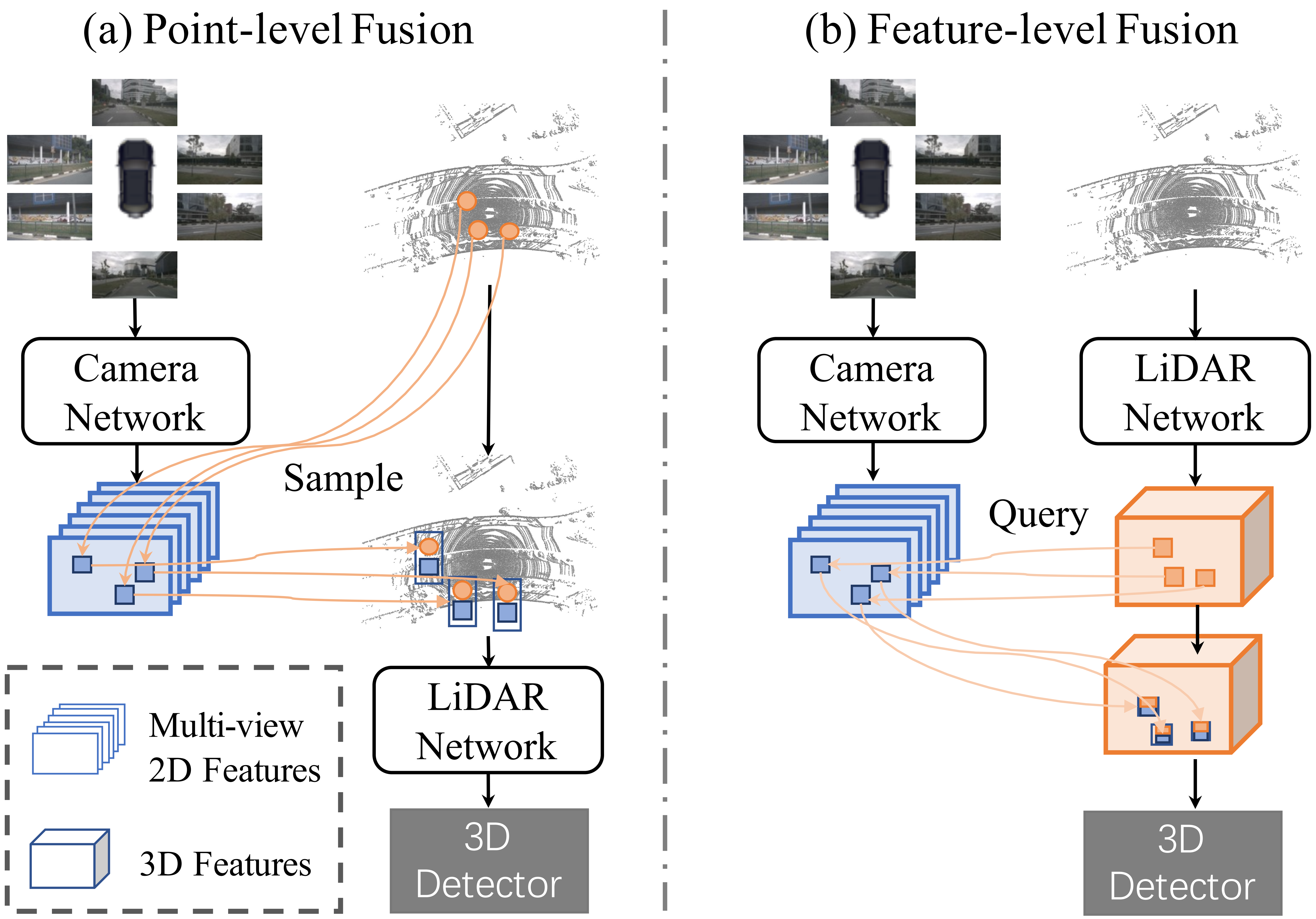
Challenges of Fusing Camera & LiDAR

Current LiDAR-Camera Fusion methods depend highly on the point cloud of the LiDAR sensor



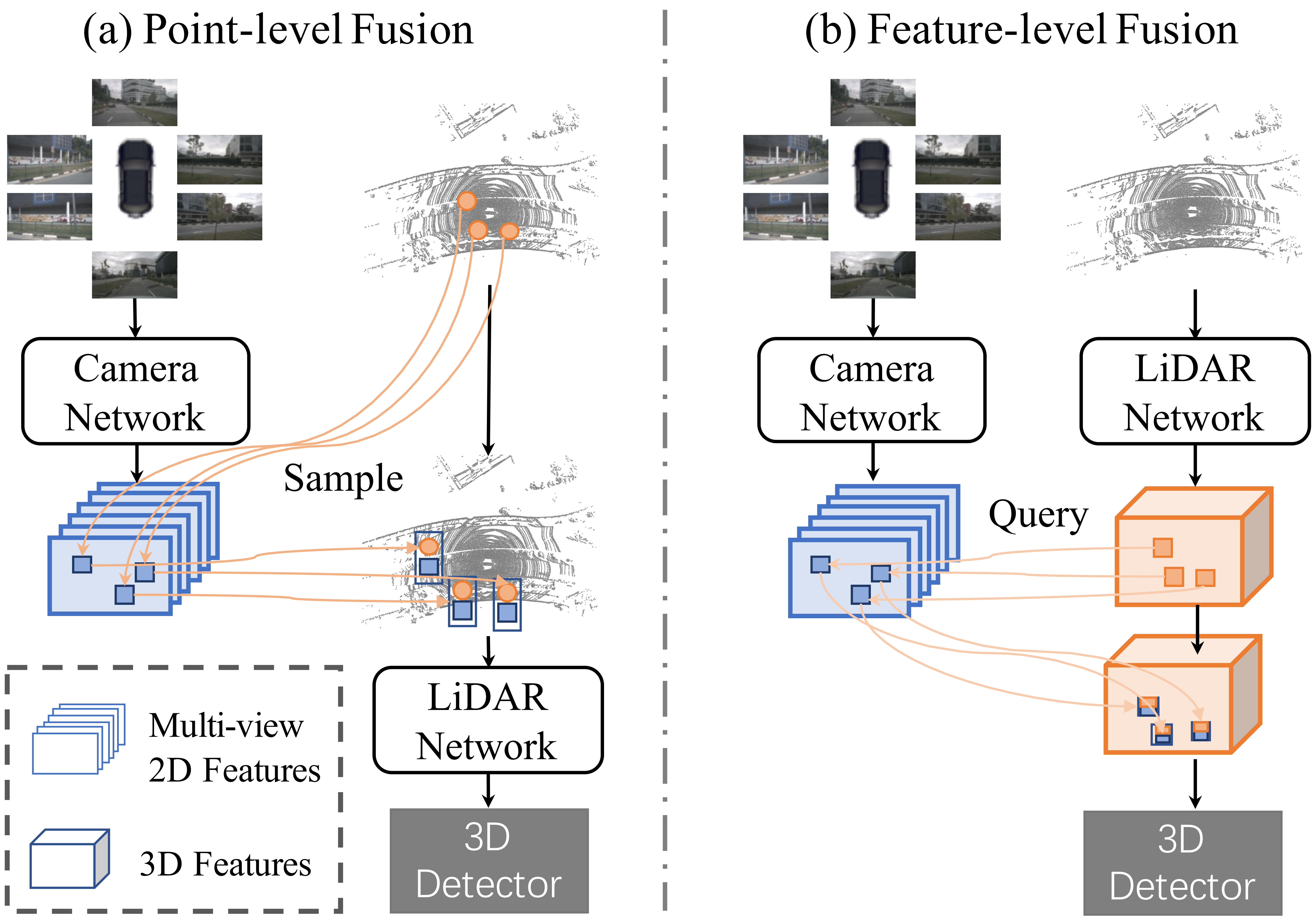
Challenges of Fusing Camera & LiDAR

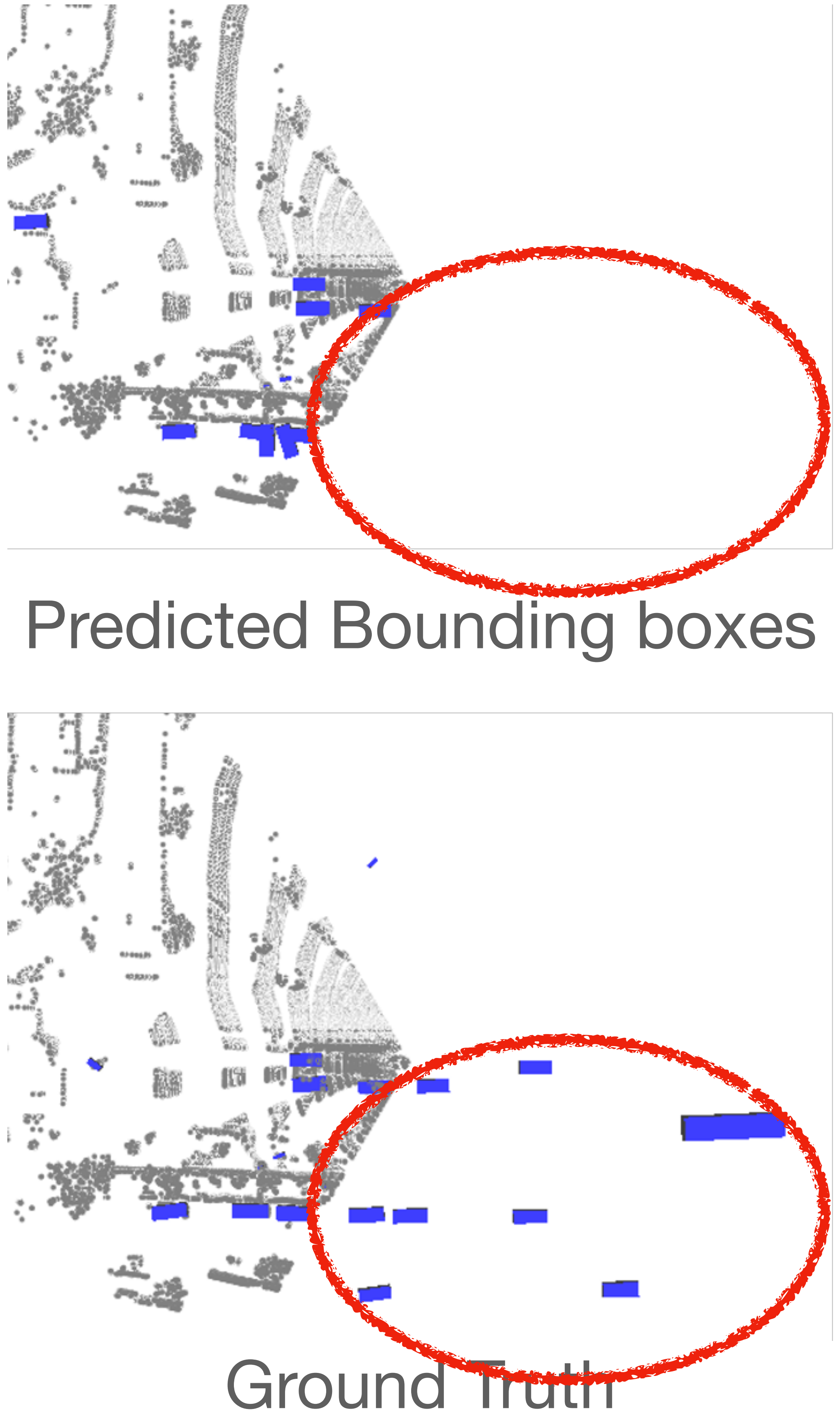
Current LiDAR-Camera Fusion methods depend highly on the point cloud of the LiDAR sensor



Challenges of Fusing Camera & LiDAR

Current LiDAR-Camera Fusion methods depend highly on the point cloud of the LiDAR sensor





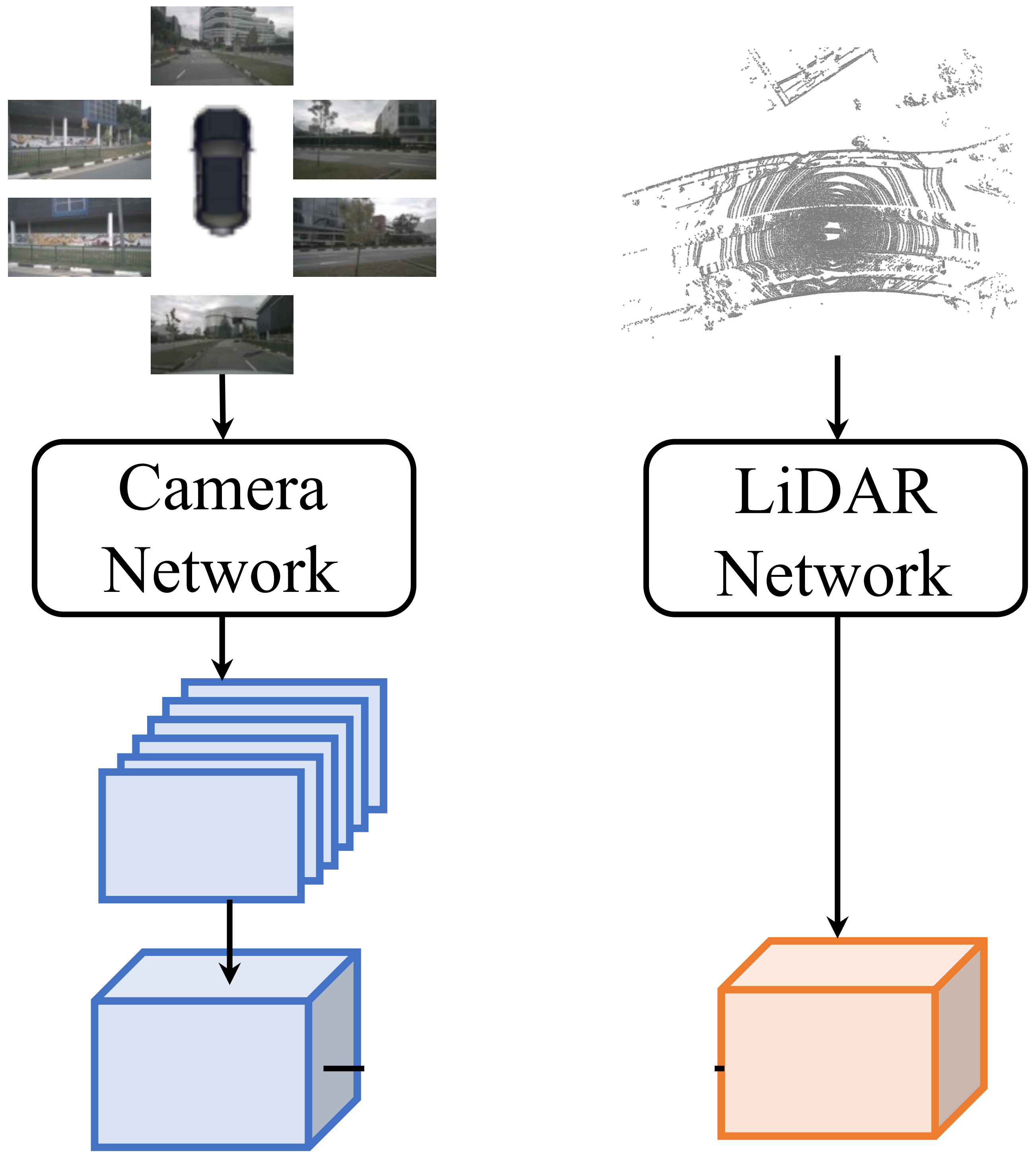
BEVFusion

A Simple and Robust LiDAR-Camera Fusion Framework

- Disentangle the two modalities during fusion

- Choose a suitable unified coordinate system

(c) Our BEVFusion

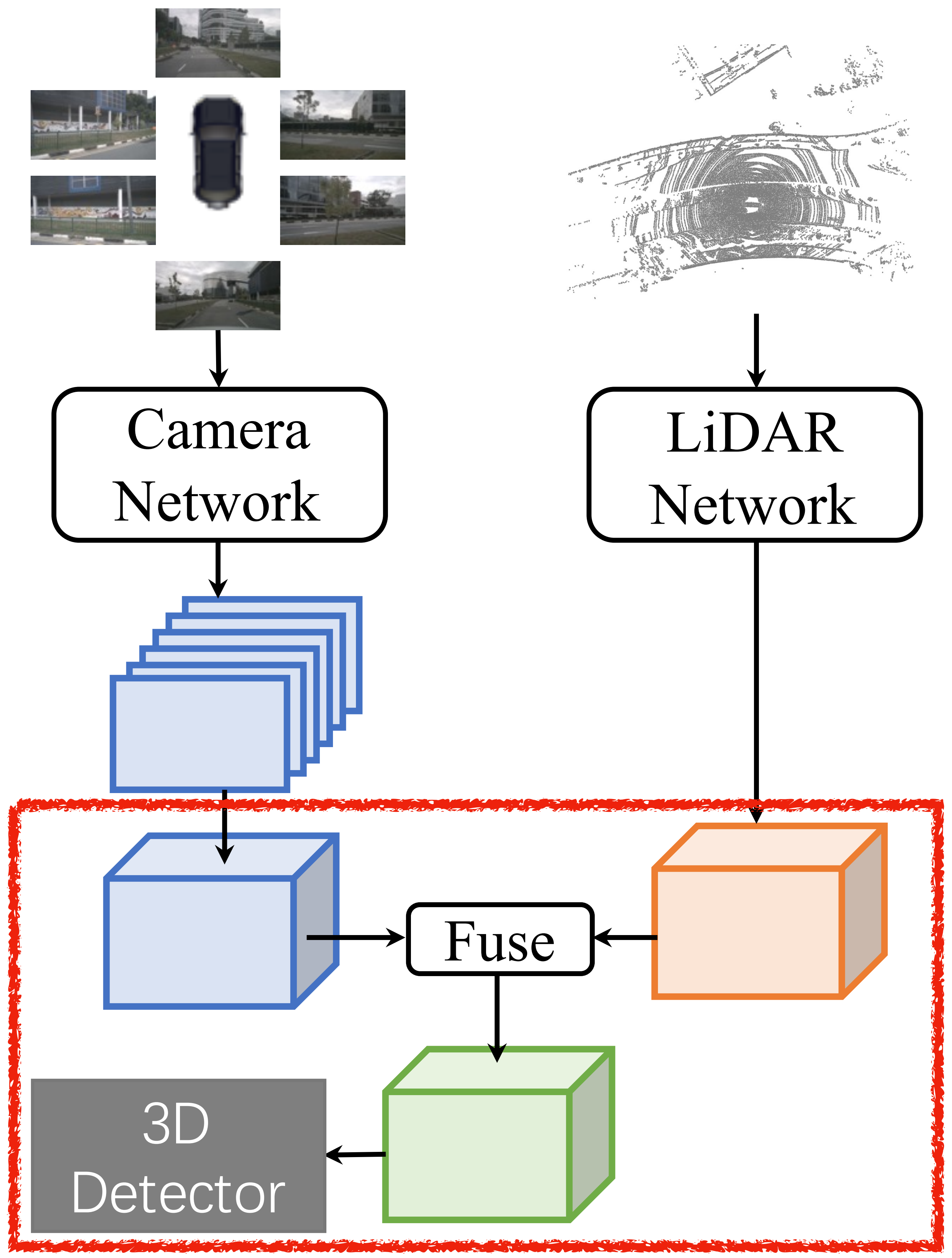


BEVFusion

A Simple and Robust LiDAR-Camera Fusion Framework

- Disentangle the two modalities during fusion
- Choose a suitable unified coordinate system
- Both modalities work complementary to each other

(c) Our BEVFusion

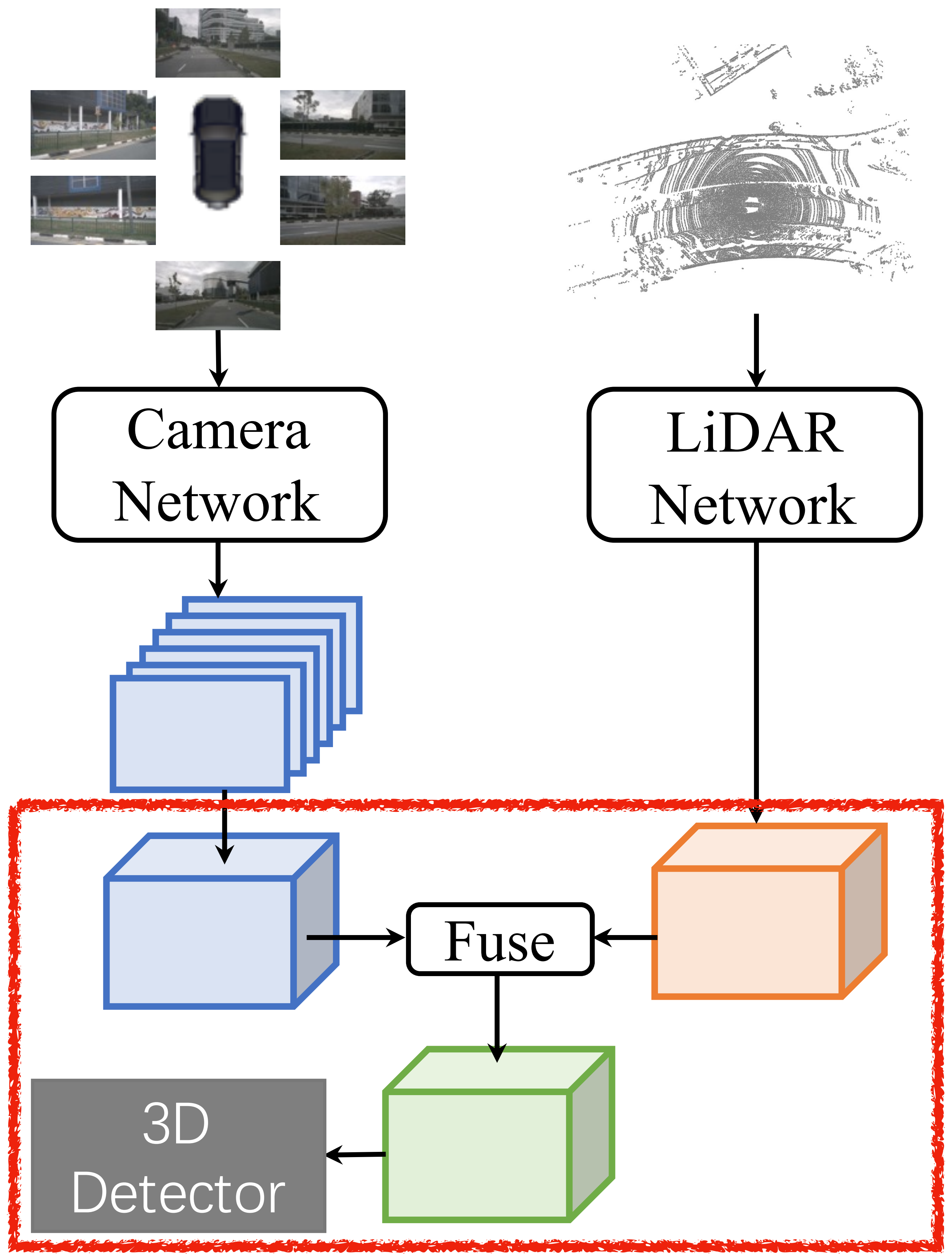


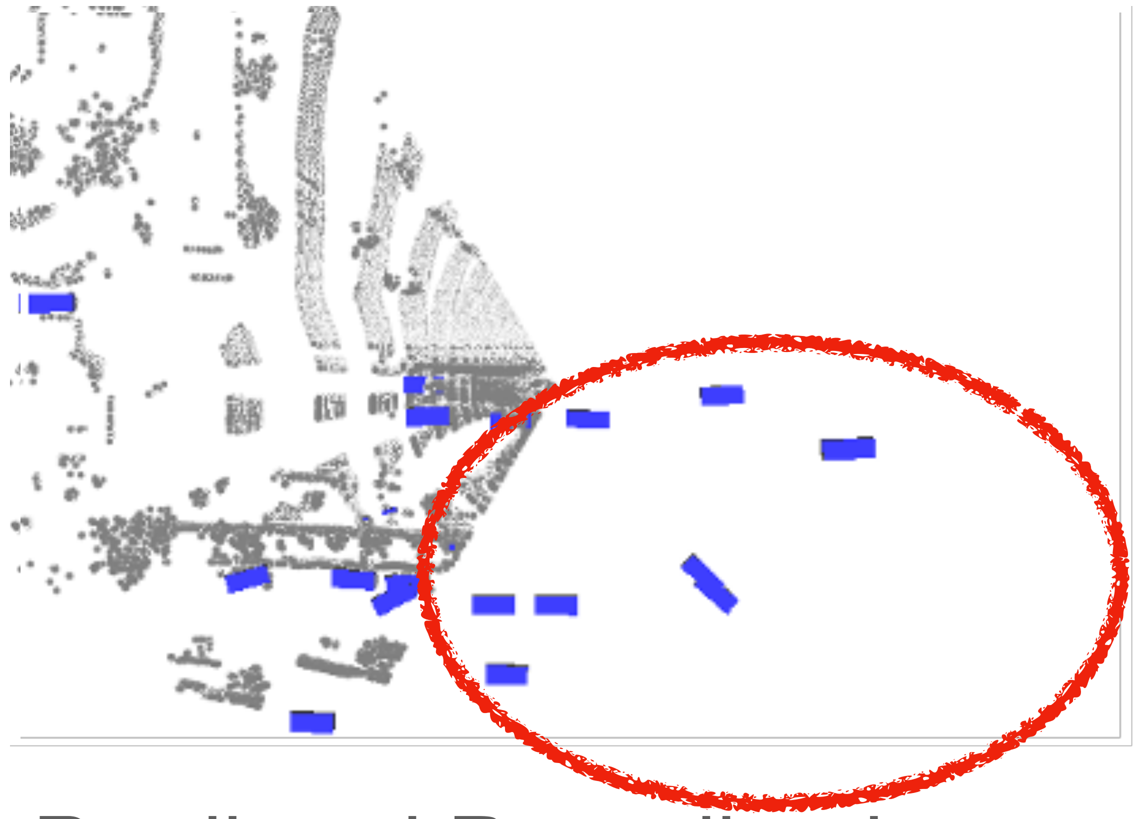
BEVFusion

A Simple and Robust LiDAR-Camera Fusion Framework

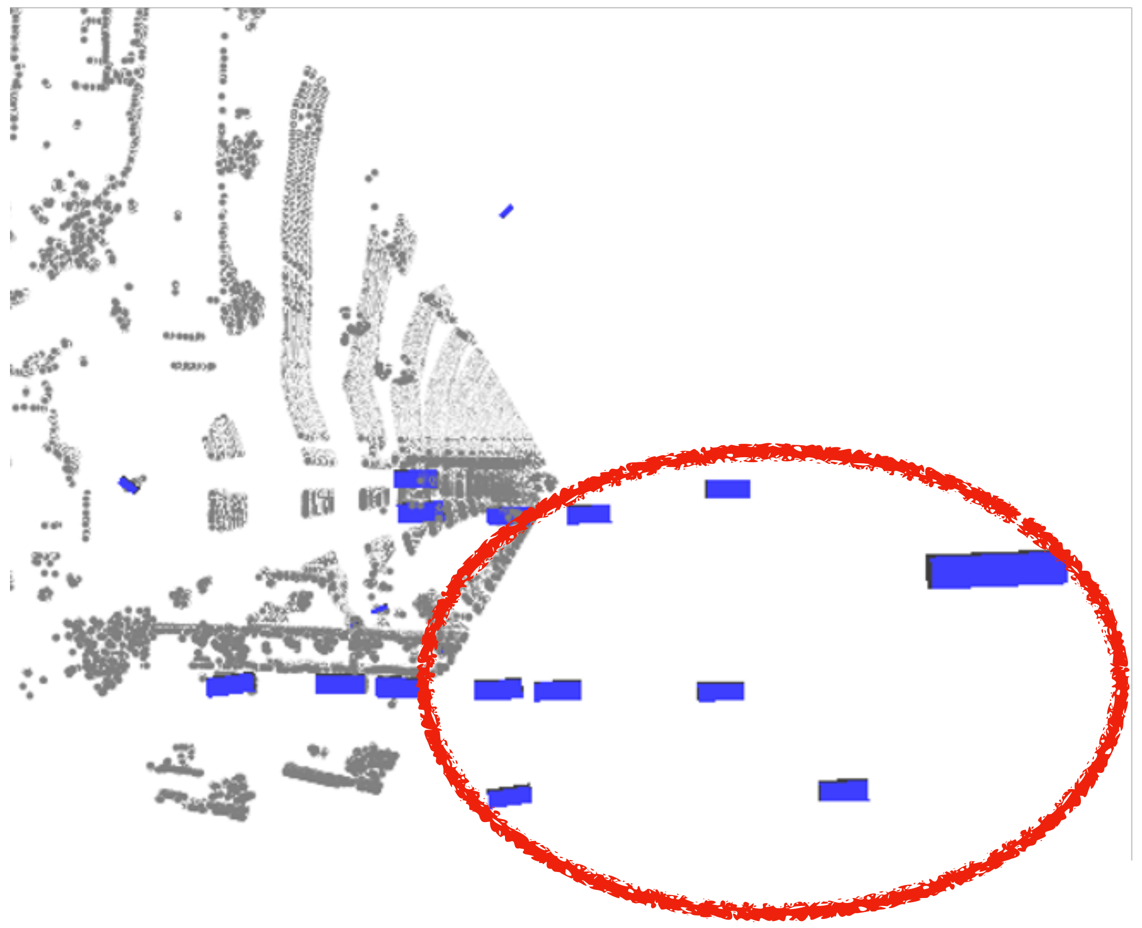
- Disentangle the two modalities during fusion
- Choose a suitable unified coordinate system
- Both modalities work complementary to each other

(c) Our BEVFusion





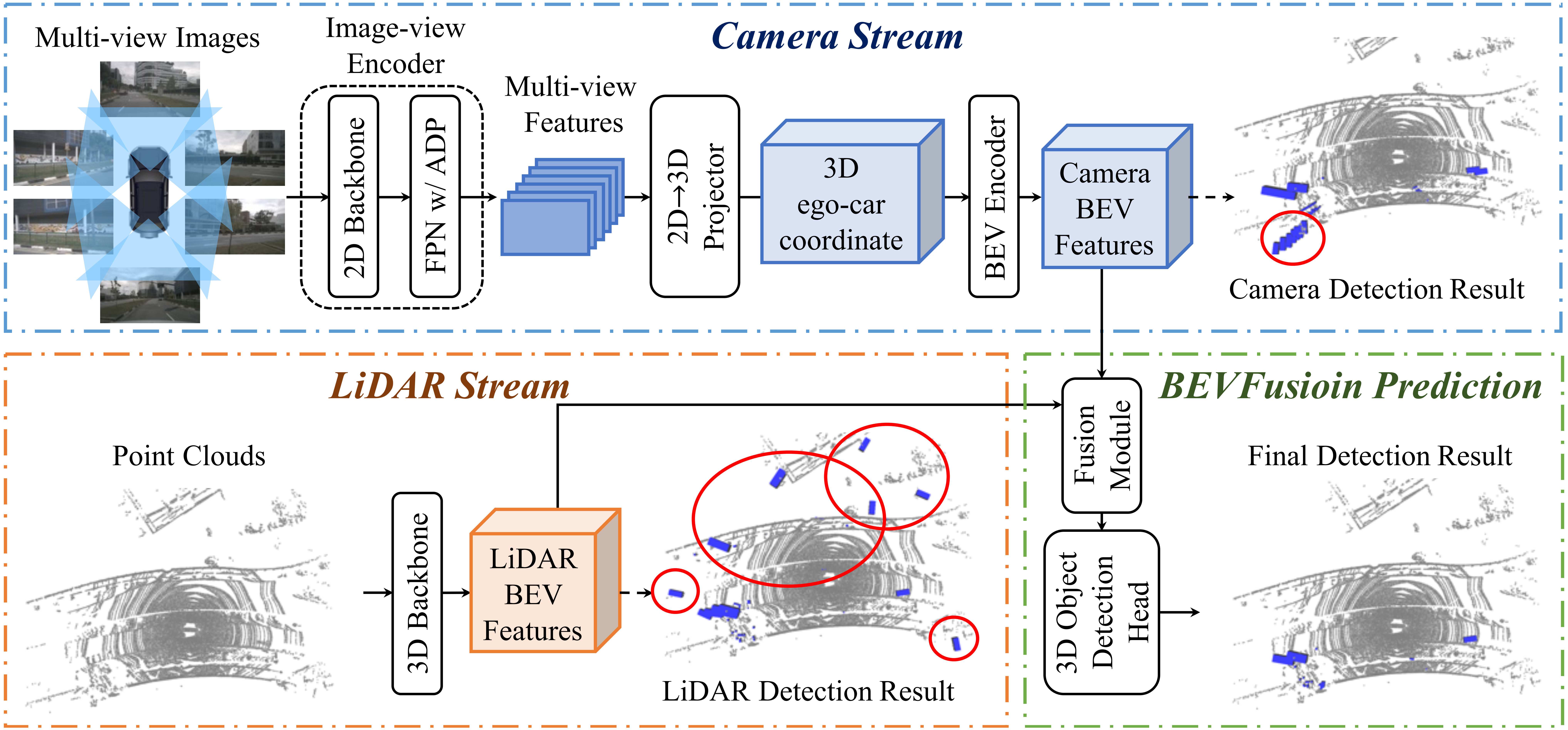
Predicted Bounding boxes



Ground Truth

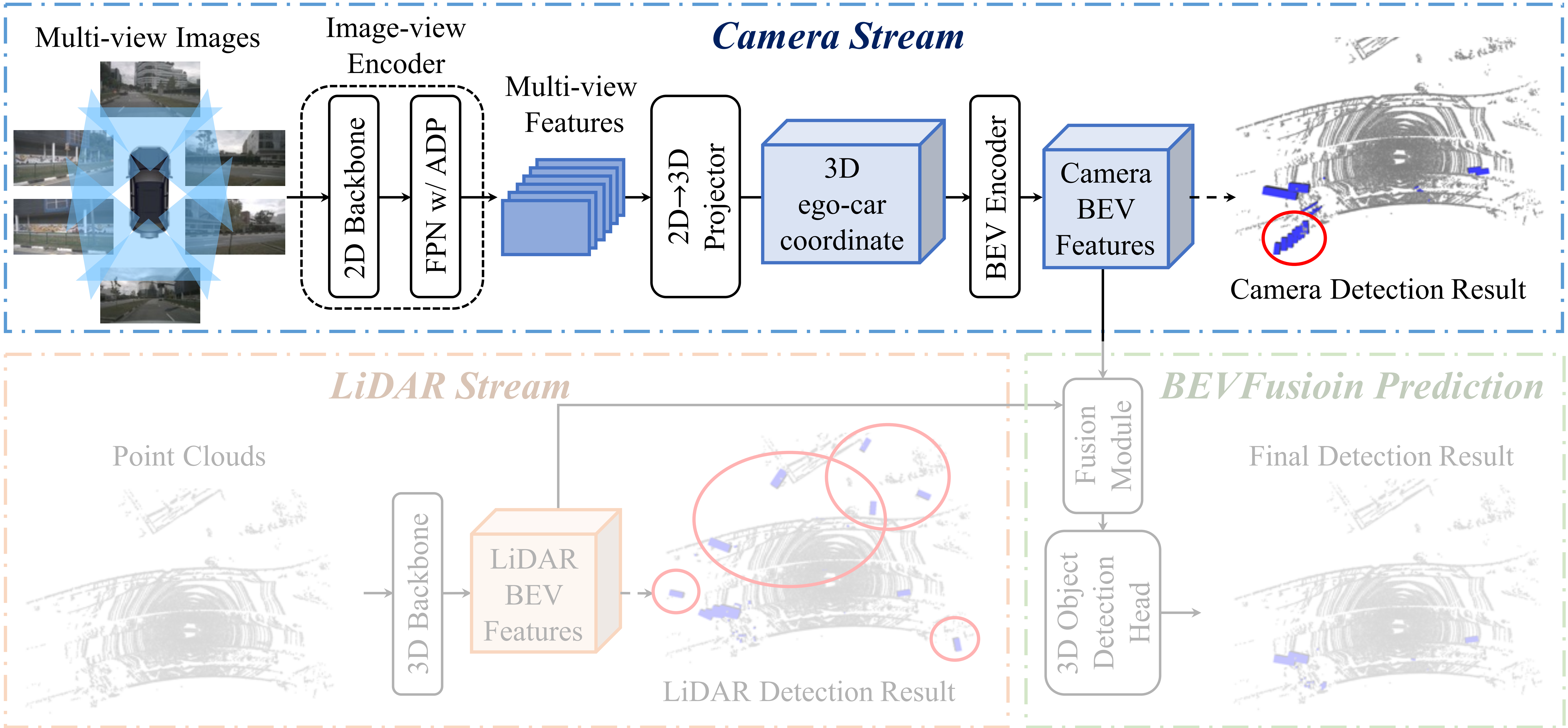
Methods

Overall Framework



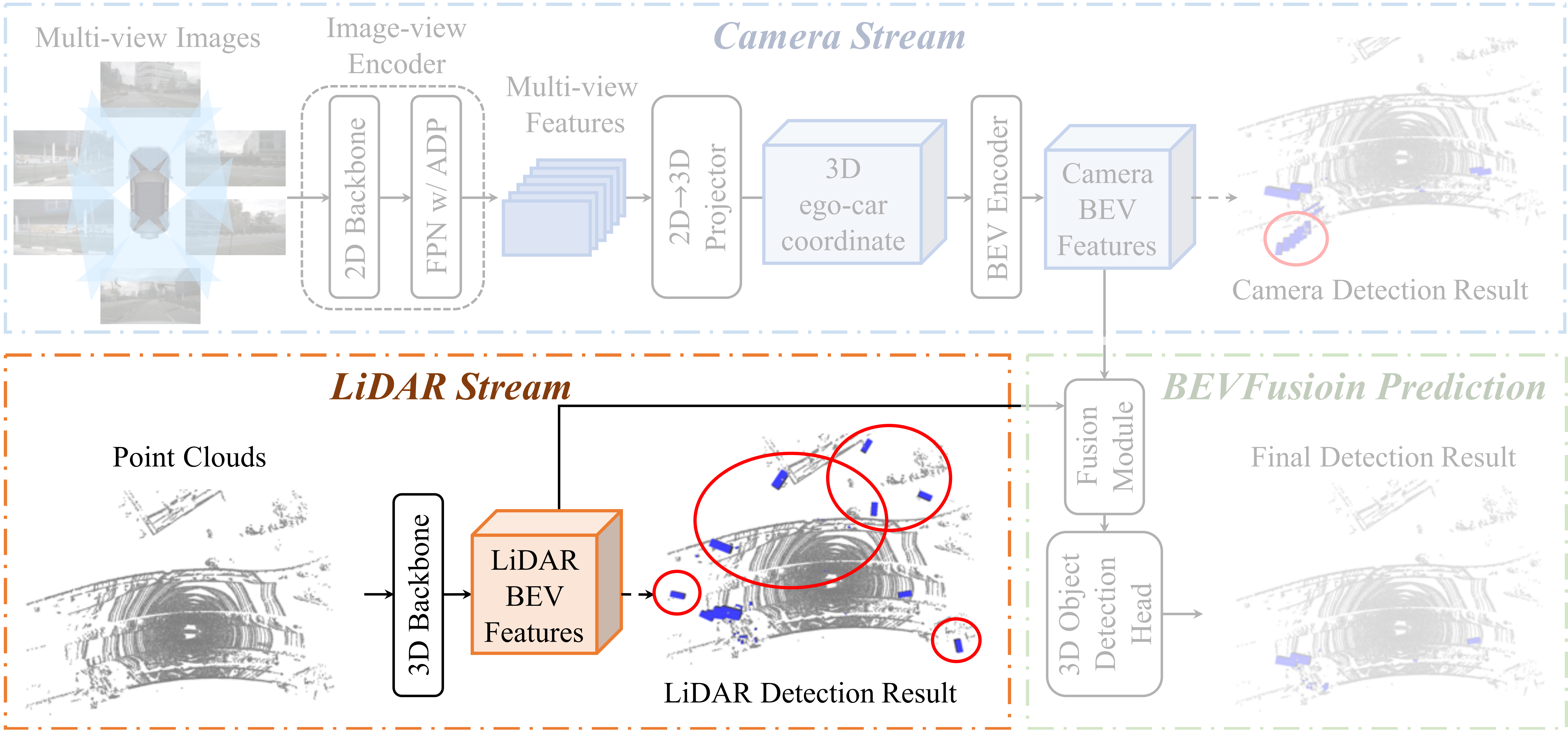
Methods

Camera Stream



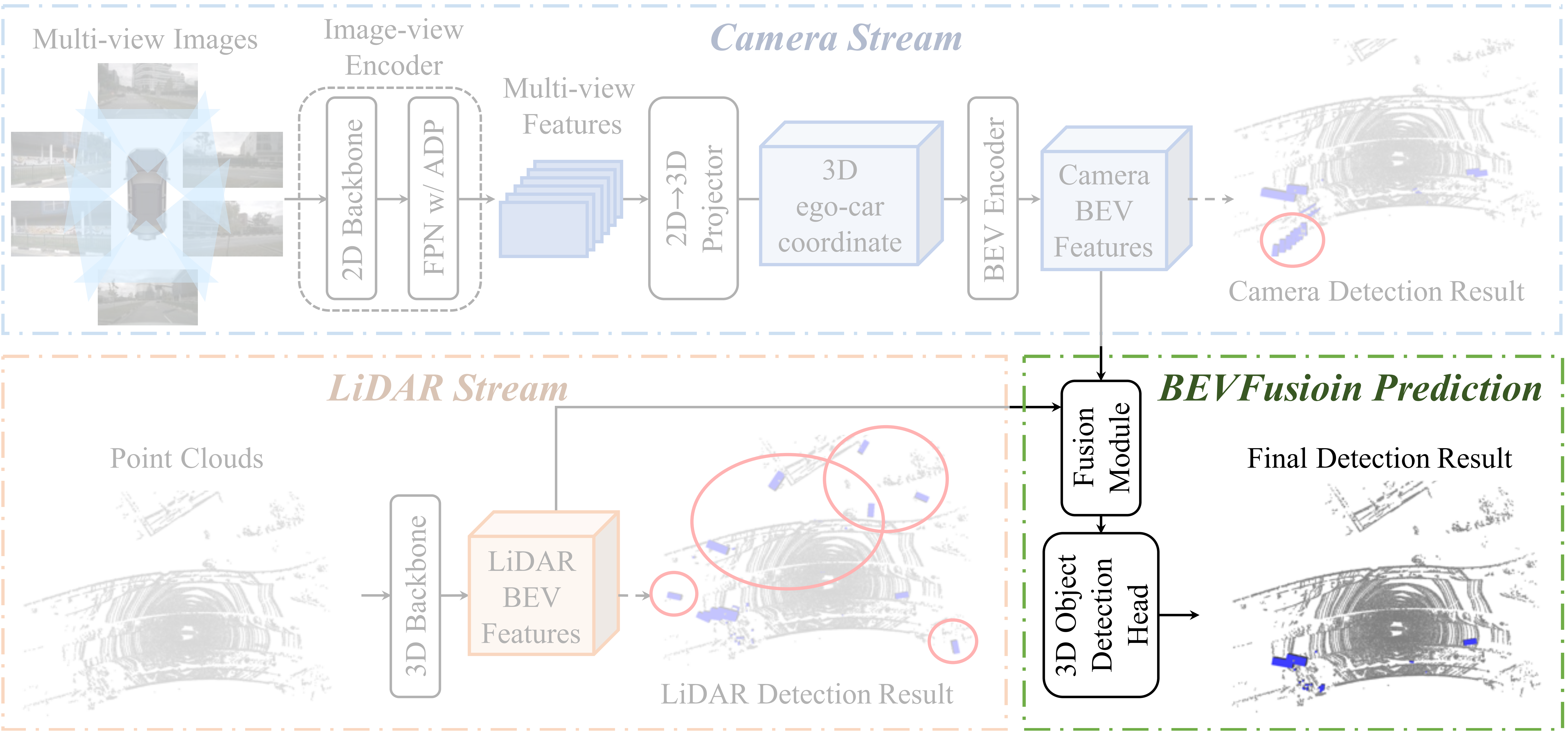
Methods

LiDAR Stream



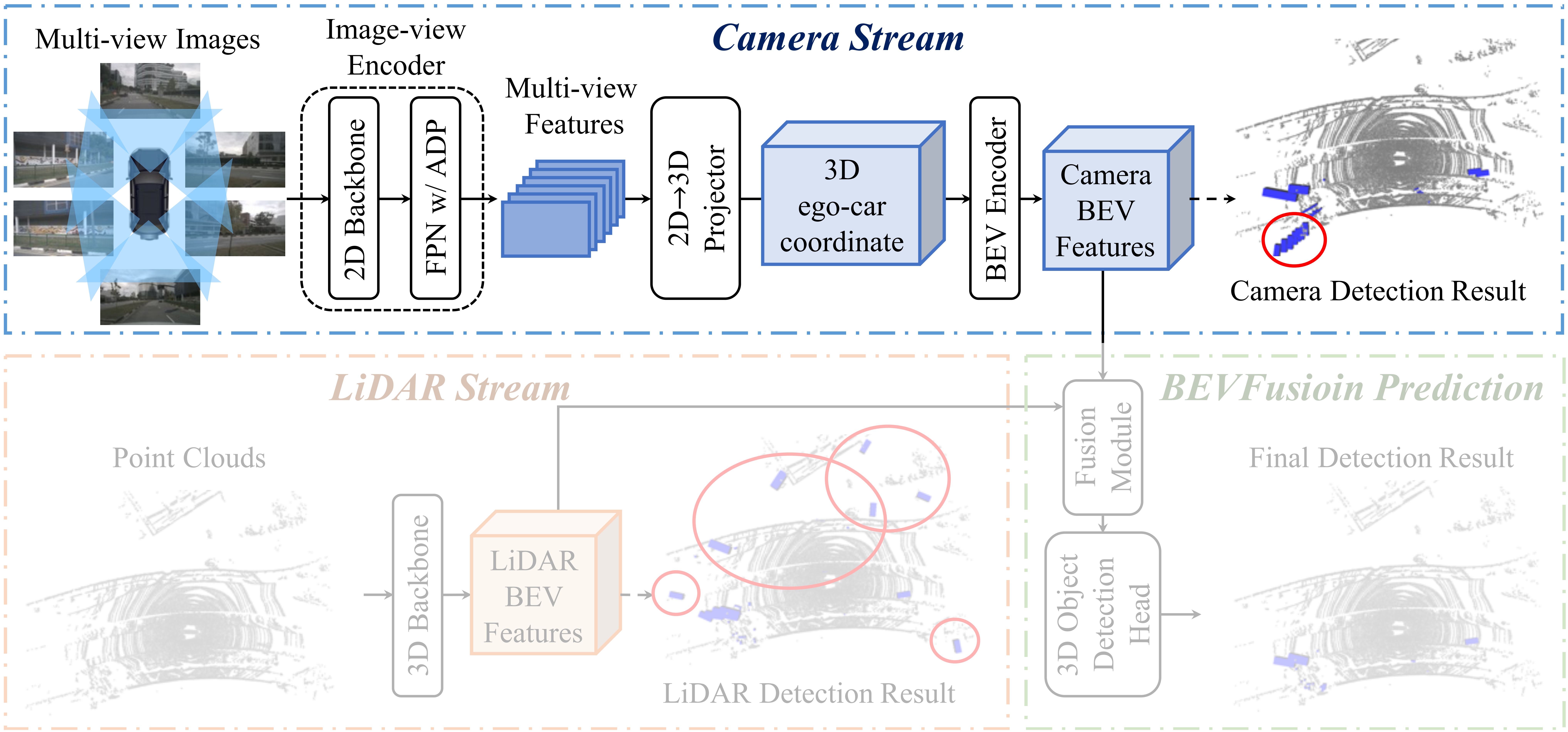
Methods

Fusion Prediction



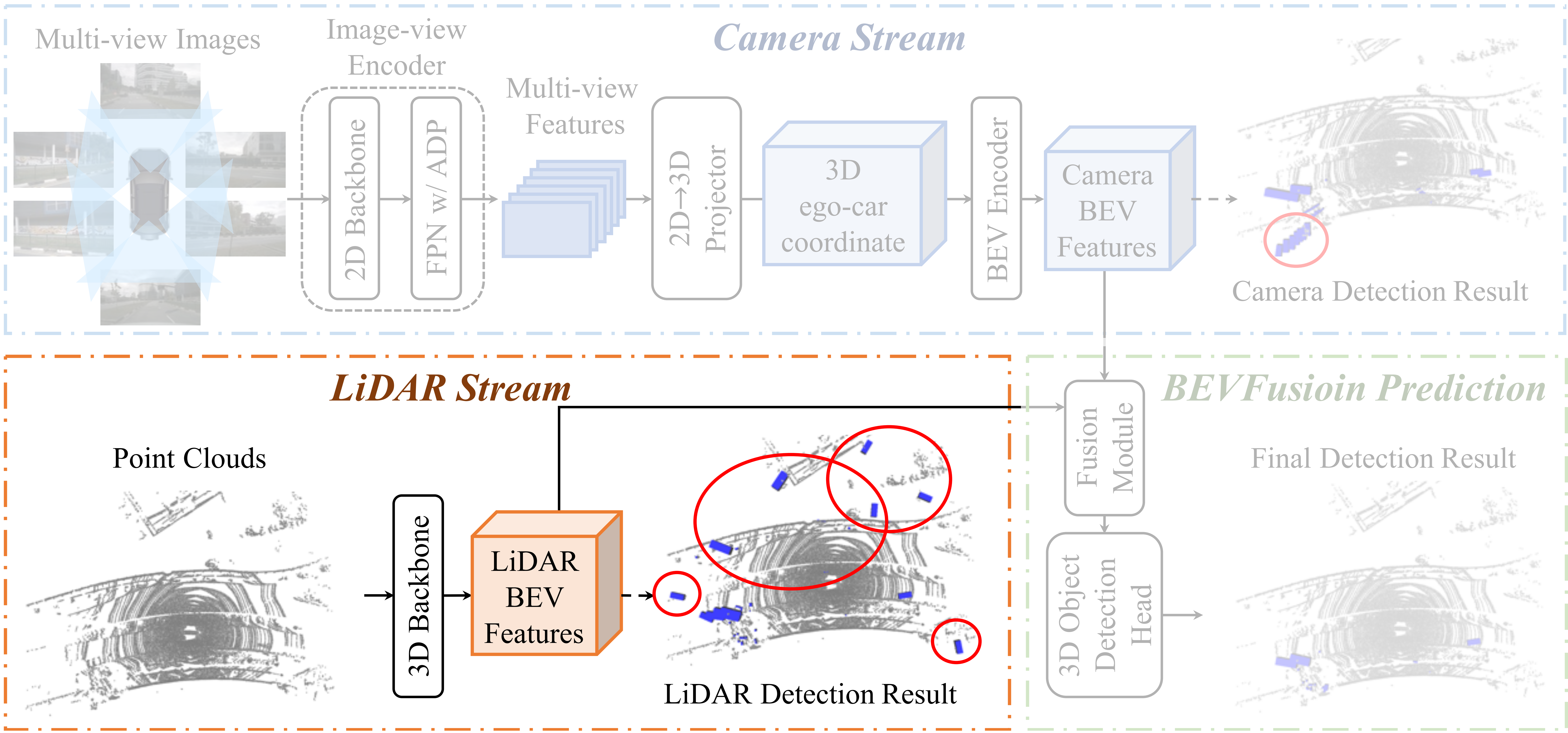
Methods

Camera Stream



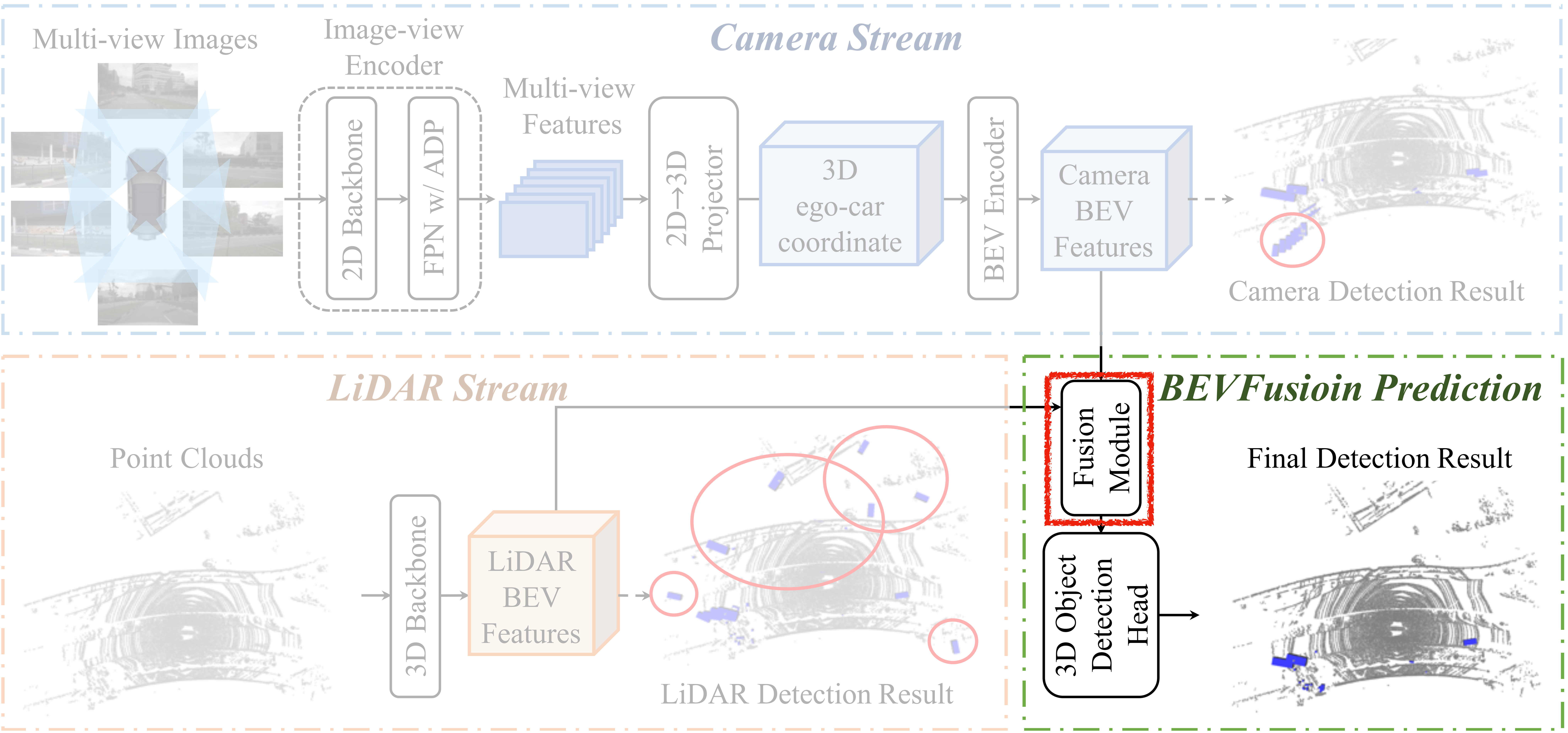
Methods

LiDAR Stream

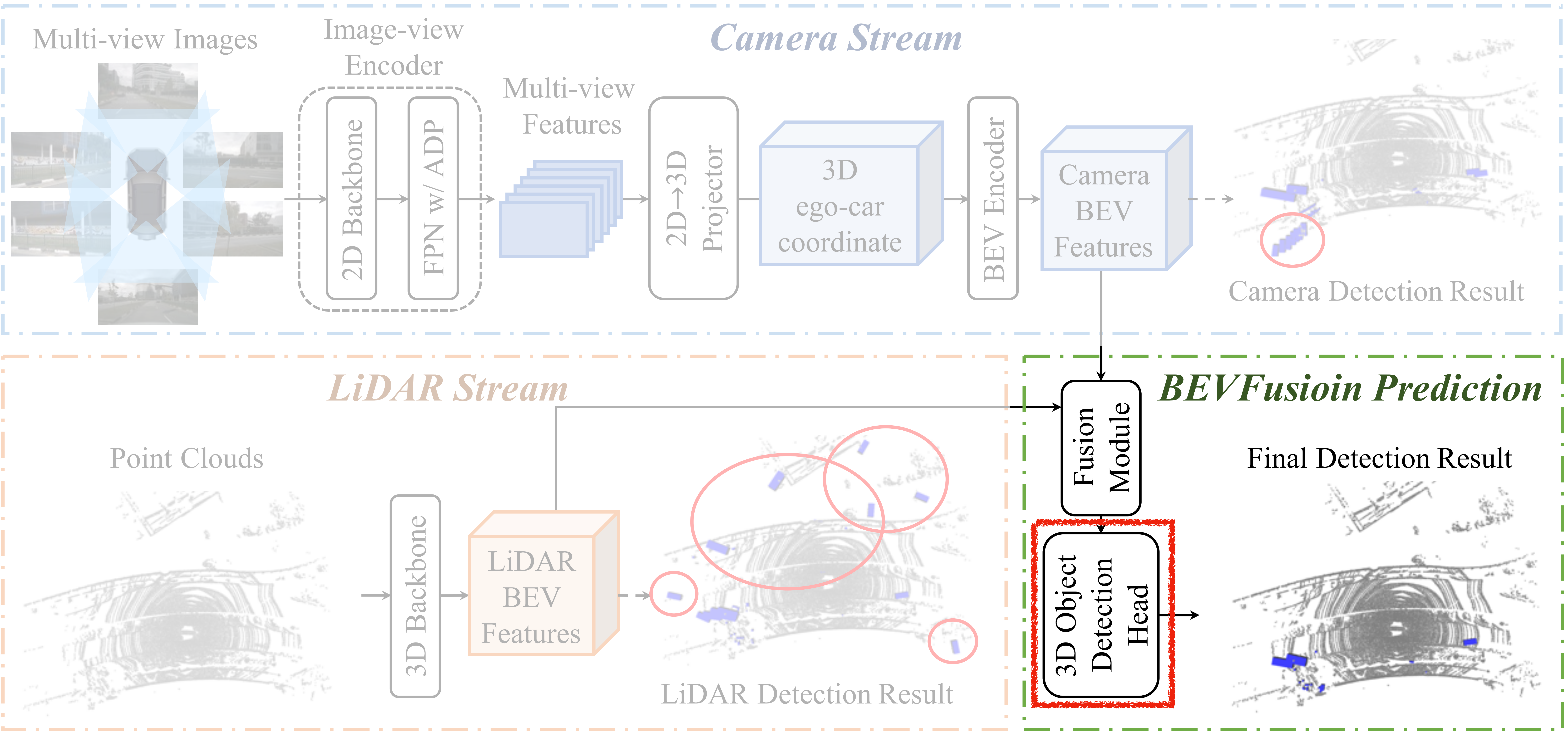


Methods

Dynamic Fusion Module



Methods
Prediction Head



Experiments

Generalization

On nuScenes validation set, BEVFusion boosts single modality streams by 3.0%-18.4%mAP over three popular methods.

Modality		PointPillars		CenterPoint		TransFusion-L	
Camera	LiDAR	mAP	NDS	mAP	NDS	mAP	NDS
✓		22.9	31.1	27.1	32.1	22.7	26.1
	✓	35.1	49.8	57.1	65.4	64.9	69.9
✓	✓	53.5	60.4	64.2	68.0	67.9	71.0

Experiments

Robustness

- Under limited Field-of-View (FOV) , BEVFusion improves its LiDAR stream by a large margin by 18.6-25.1% mAP.

FOV	Metrics	PointPillars LiDAR	PointPillars ↑BEVFusion	CenterPoint LiDAR	CenterPoint ↑BEVFusion	TransFusion LiDAR	TransFusion ↑BEVFusion	TransFusion LC
$(-\pi/2, \pi/2)$	mAP	12.4	36.8 (+24.4)	23.6	45.5 (+21.9)	27.8	46.4 (+18.6)	31.1
	NDS	37.1	45.8 (+8.7)	48.0	54.9 (+6.9)	50.5	55.8 (+5.3)	49.2
$(-\pi/3, \pi/3)$	mAP	8.4	33.5 (+25.1)	15.9	40.9 (+25.0)	19.0	41.5 (+22.5)	21.0
	NDS	34.3	42.1 (+7.8)	43.5	49.9 (+6.4)	45.3	50.8 (+5.5)	41.2

Experiments

Robustness

- Under camera malfunctions, BEVFusion outperforms camera-only and other LiDAR-camera fusion methods.

Approach	Clean		Missing F		Preserve F		Stuck	
	mAP	NDS	mAP	NDS	mAP	NDS	mAP	NDS
DETR3D[53]	34.9	43.4	25.8	39.2	3.3	20.5	17.3	32.3
PointAugmenting[47]	46.9	55.6	42.4	53.0	31.6	46.5	42.1	52.8
MVX-Net[43]	61.0	66.1	47.8	59.4	17.5	41.7	48.3	58.8
TransFusion[2]	66.9	70.9	65.3	70.1	64.4	69.3	65.9	70.2
BEVFusion	67.9	71.0	65.9	70.7	65.1	69.9	66.2	70.3

Experiments

Comparison with SOTA

Method	Modality	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bike	Ped.	T.C.
FUTR3D [5]	LC	64.2	68.0	86.3	61.5	26.0	71.9	42.1	64.4	73.6	63.3	82.6	70.1
BEVFusion	LC	67.9	71.0	88.6	65.0	28.1	75.4	41.4	72.2	76.7	65.8	88.7	76.9
BEVFusion*	LC	69.6	72.1	89.1	66.7	30.9	77.7	42.6	73.5	79.0	67.5	89.4	79.3
PointPillars[20]	L	30.5	45.3	68.4	23.0	4.1	28.2	23.4	38.9	27.4	1.1	59.7	30.8
CBGS[67]	L	52.8	63.3	81.1	48.5	10.5	54.9	42.9	65.7	51.5	22.3	80.1	70.9
CenterPoint[59]†	L	60.3	67.3	85.2	53.5	20.0	63.6	56.0	71.1	59.5	30.7	84.6	78.4
TransFusion-L [1]	L	65.5	70.2	86.2	56.7	28.2	66.3	58.8	78.2	68.3	44.2	86.1	82.0
PointPainting[46]	LC	46.4	58.1	77.9	35.8	15.8	36.2	37.3	60.2	41.5	24.1	73.3	62.4
3D-CVF[61]	LC	52.7	62.3	83.0	45.0	15.9	48.8	49.6	65.9	51.2	30.4	74.2	62.9
PointAugmenting[47]†	LC	66.8	71.0	87.5	57.3	28.0	65.2	60.7	72.6	74.3	50.9	87.9	83.6
MVP[60]	LC	66.4	70.5	86.8	58.5	26.1	67.4	57.3	74.8	70.0	49.3	89.1	85.0
FusionPainting[55]	LC	68.1	71.6	87.1	60.8	30.0	68.5	61.7	71.8	**74.7**	**53.5**	88.3	85.0
TransFusion[1]	LC	68.9	71.7	87.1	60.0	33.1	68.3	60.8	78.1	73.6	52.9	88.4	**86.7**
BEVFusion (Ours)	LC	**69.2**	**71.8**	**88.1**	**60.9**	**34.4**	**69.3**	**62.1**	**78.2**	72.2	52.2	**89.2**	85.2
BEVFusion (Ours)*	LC	**71.3**	**73.3**	**88.5**	**63.1**	**38.1**	**72.0**	**64.7**	**78.3**	**75.2**	**56.5**	**90.0**	**86.5**

† These methods exploit double-flip during the test time. The best and second best results are marked in **red** and **blue**.

Notion of class: Construction vehicle (C.V.), pedestrian (Ped.), traffic cone (T.C.). Notion of modality: Camera (C), LiDAR (L).

* These methods exploit BEV-space data augmentation during training.

Conclusion: LiDAR-Camera Fusion

- Limitation of previous methods
  - Dependency of LiDAR inputs
- Our BEVFusion
  - Disentangle LiDAR / camera modality into two independent streams
  - Good generalization ability
  - Effective and robust

Thanks!

Code

Scan ME