Paper Review (ECCV 2024)

MLCF

SAMFusion: Sensor-Adaptive Multimodal Fusion for 3D Object Detection in Adverse Weather

Edoardo Palladin*[1], Roland Dietze*[2], Praveen Narayanan[1], Mario Bijelic[1,3], and Felix Heide[1,3]

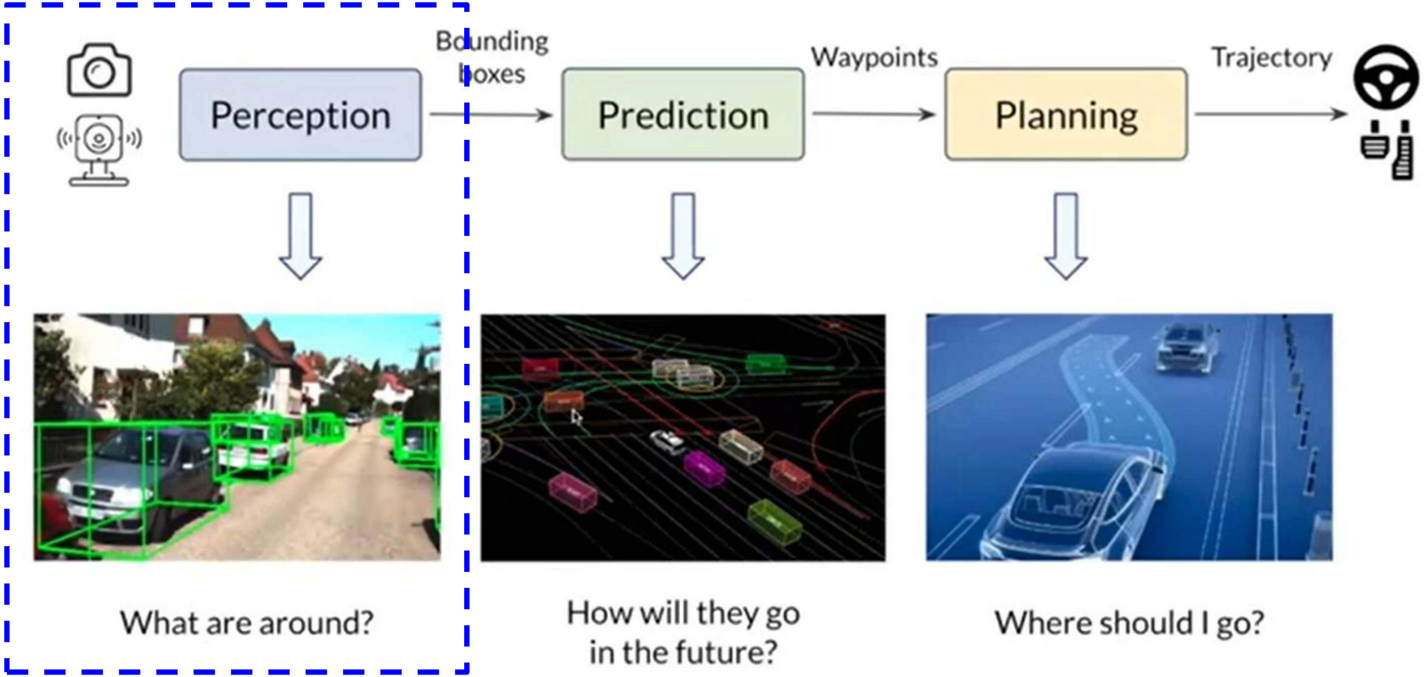
[1]Torc Robotics [2]University of Stuttgart [3]Princeton University

Susang Kim

Contents

1.Introduction - Perception in Autonomous Driving

Challenge | Various weathers, illuminations, and scenarios



They perform well under normal environmental conditions **but may fail in adverse weather**, such as heavy fog, snow, or obstructions caused by soiling.

1.Introduction - Various Types of Vision Sensors

(a) Camera (b) LiDAR (c) Radar

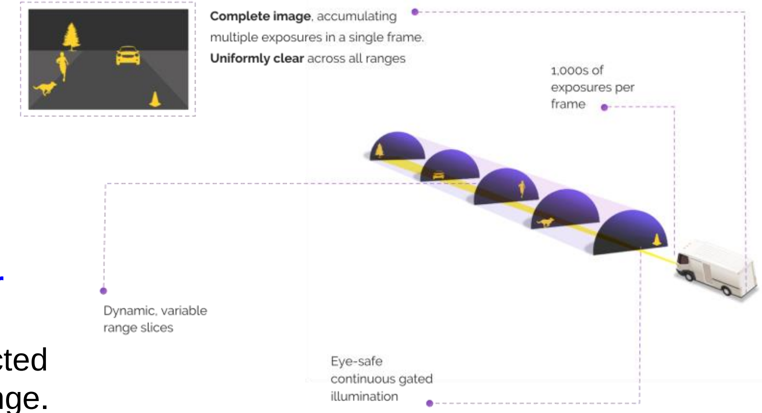
(d) Event camera (e) IMU (f) Thermal camera

(a) Camera image (b) LiDAR point cloud

(c) Radar point cloud (d) Event-based camera image (e) Thermal camera image

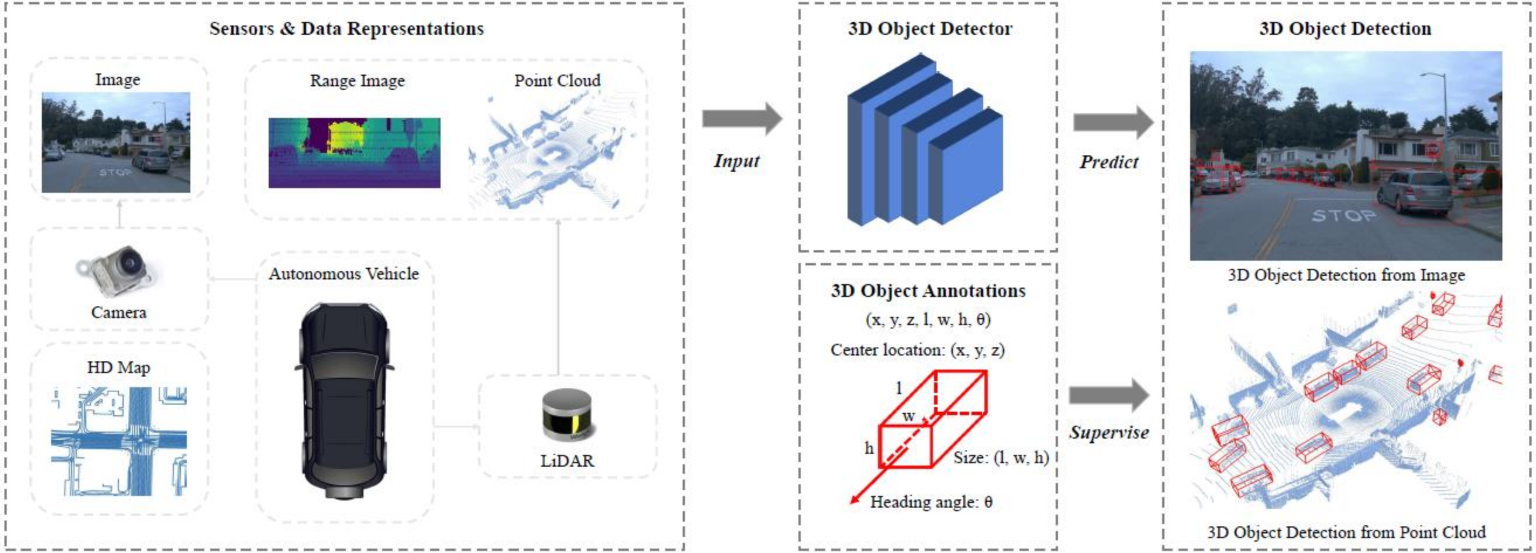
NIR(Near-infrared)
gated camera

A camera that opens its **shutter only during a specific time window (gate)** to capture reflected light from a desired distance range.

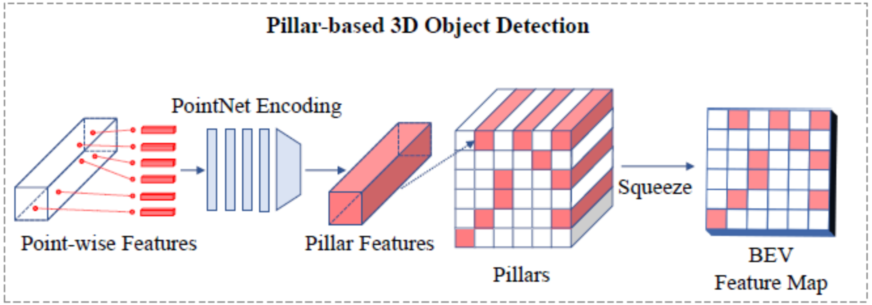


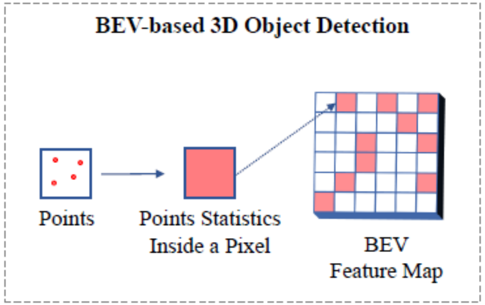
Gated camera : https://www.brightwayvision.com/technology
Liu, Mingyu, et al. "A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook." Transactions on Intelligent Vehicles 2024.

1.Introduction – 3D Object Detection



$$B = [x_c, y_c, z_c, l, w, h, \theta, class]$$





[Geometric consistency]

the extrinsics matrix $T \in SE(3)$

the camera intrinsics matrix K

3D point $[x, y, z]$

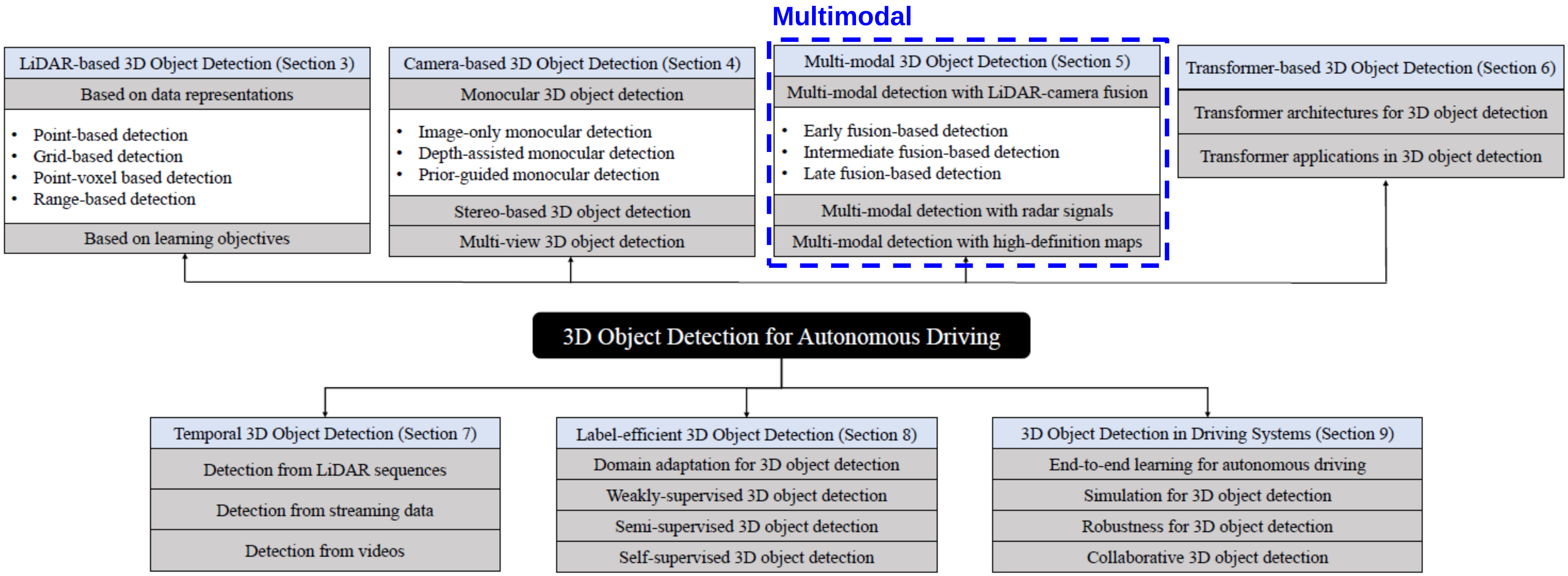
image pixel coordinate $[u, v]$

Depth d

$$d \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K\,T \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

Mao, Jiageng, et al. "3D object detection for autonomous driving: A comprehensive survey." IJCV 2023.

1.Introduction – Multi-modal 3D Object Detection



Mao, Jiageng, et al. "3D object detection for autonomous driving: A comprehensive survey." IJCV 2023.

1.Introduction - Multimodal sensor fusion

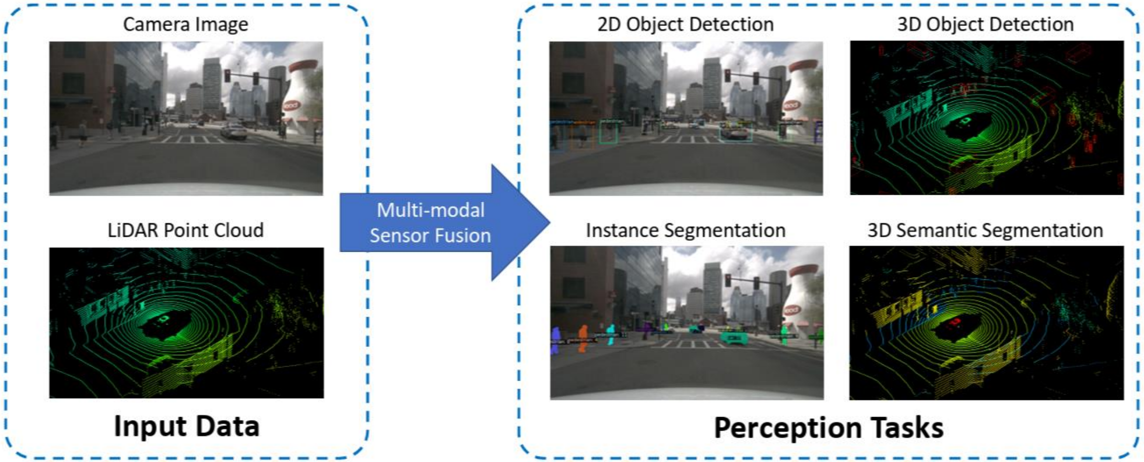


Figure 1. Perception Tasks of Autonomous Driving by Multi-modal Sensor Fusion Model.

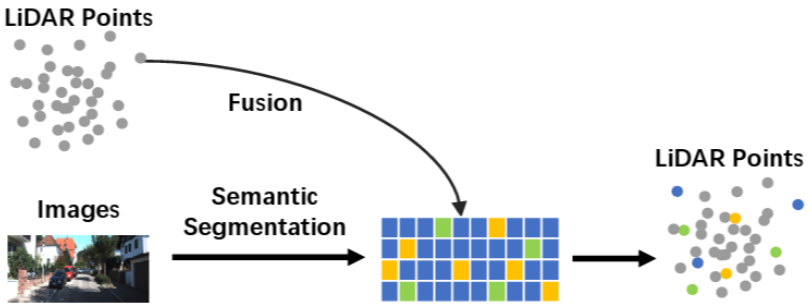


Figure 4. An Example of Early-Fusion

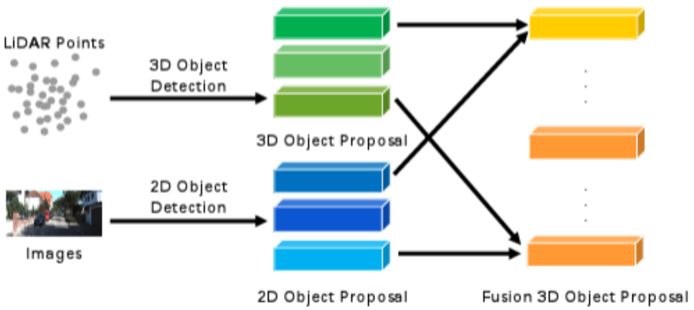
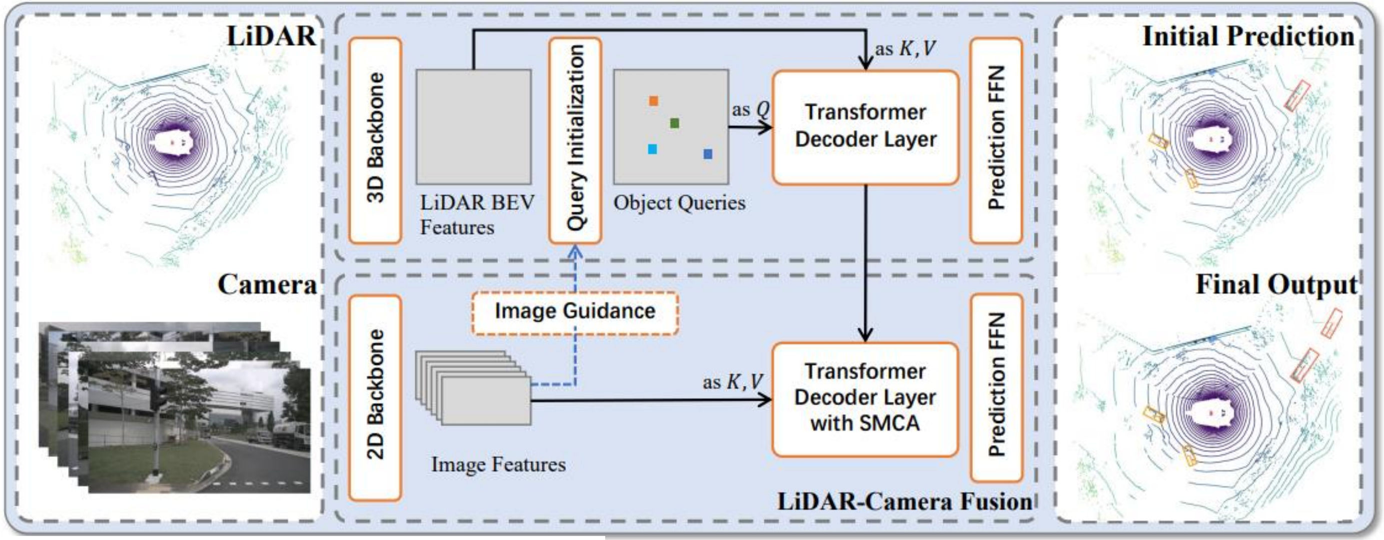


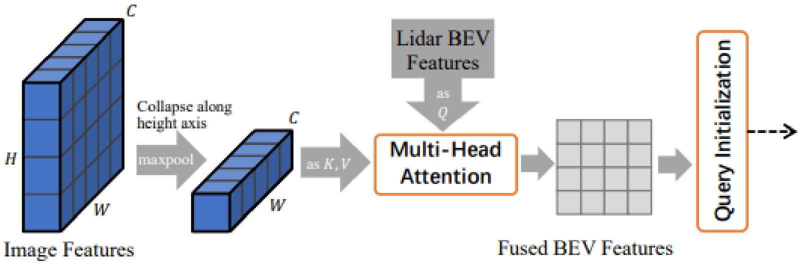
Figure 6. An Example of Late-Fusion

Huang, Keli, et al. "Multi-modal sensor fusion for auto driving perception: A survey." arXiv 2022.

2.Related Works – Transfusion (Lidar+Camera) (CVPR 2022)

TransFusion is a robust solution for LiDAR-camera fusion, employing a soft-association mechanism to handle challenging image conditions. Specifically, TransFusion consists of convolutional backbones and a detection head based on a transformer decoder.

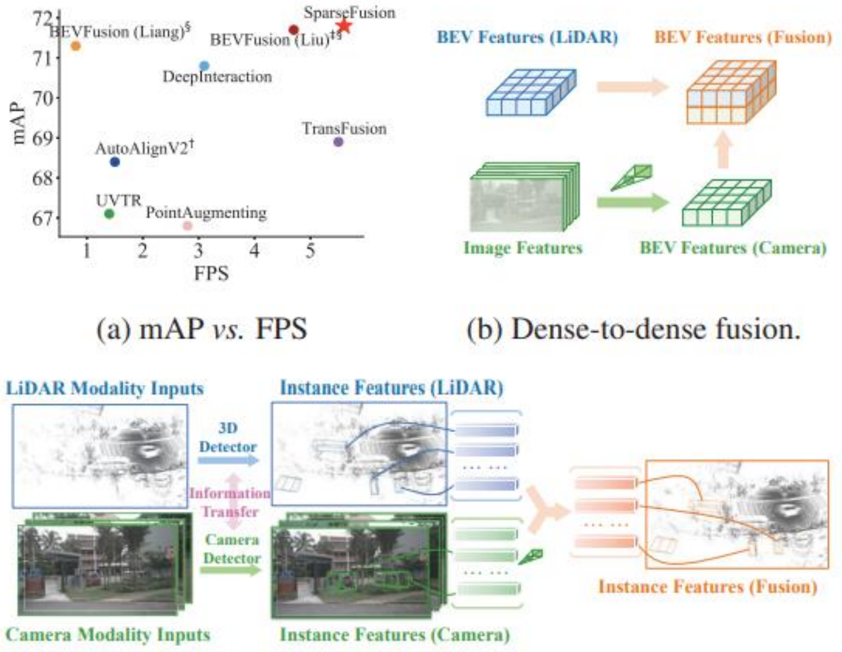




A novel transformer-based LiDAR-camera fusion model for 3D detection that performs fine-grained fusion in an attentive manner and demonstrates superior robustness against degraded image quality and sensor misalignment.

Bai, Xuyang, et al. "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers." CVPR 2022.

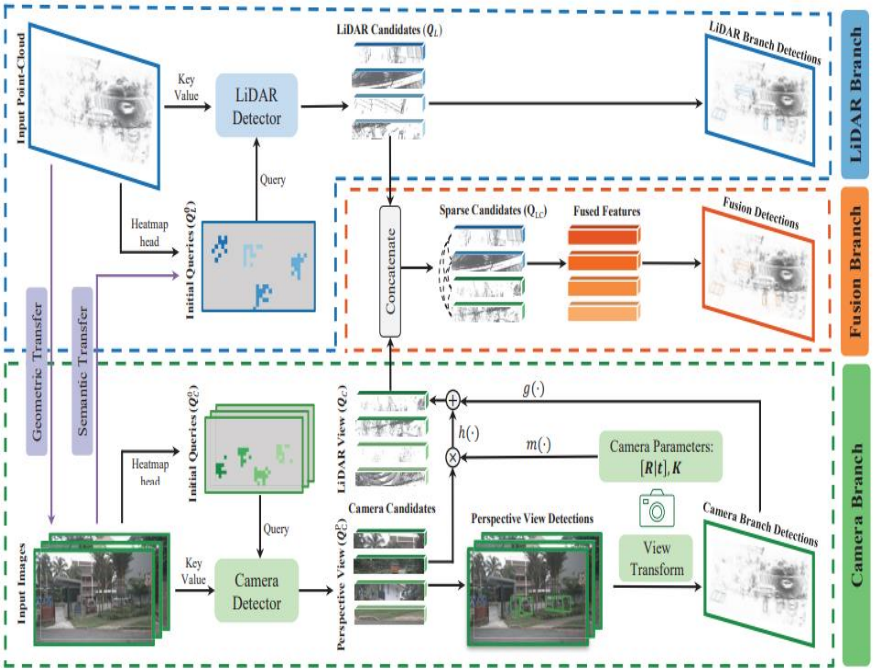
2.Related Works – SparseFusion (Lidar+Camera) (CVPR 2023)



(a) mAP *vs.* FPS (b) Dense-to-dense fusion.

(c) Overview of our sparse fusion strategy. We extract instance-level features from the LiDAR and camera modalities separately, and fuse them in a unified 3D space to perform detection.

SparseFusion utilizes the outputs of parallel detectors in the LiDAR and camera modalities as sparse candidates for fusion.

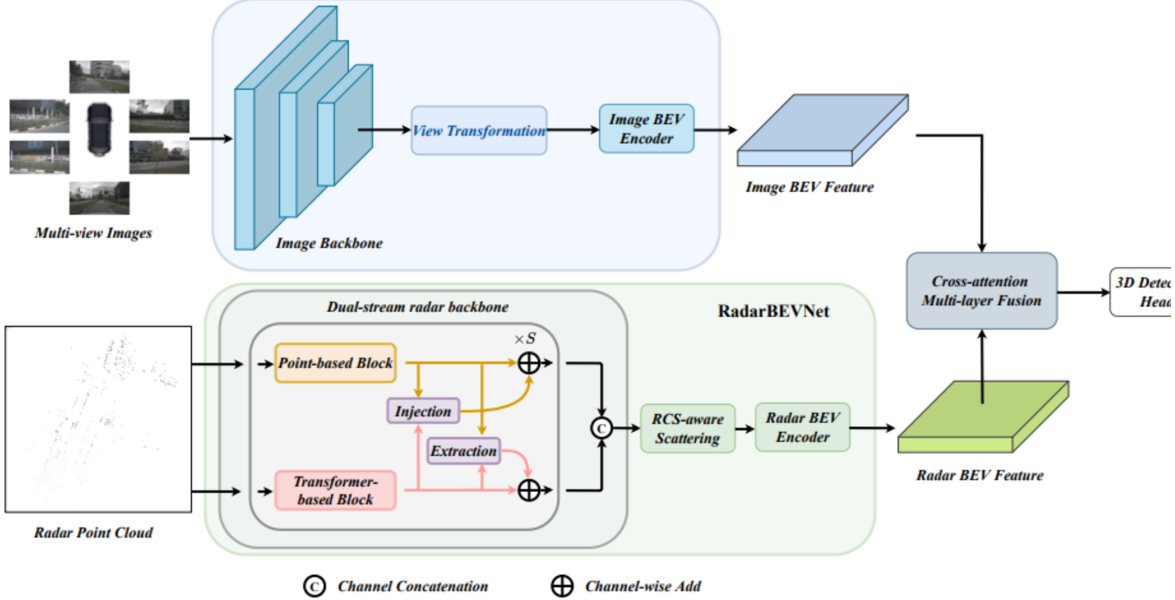


Sparse fuses sparse candidates from LiDAR and camera modalities to obtain a multi-modality instance-level representation in the unified LiDAR space

Xie, Yichen, et al. "Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection." CVPR 2023.

2.Related Works - RCBEVDet (Radar+Camera) (CVPR 2024)

RCBEVDet is a radar-camera fusion method for 3D object detection in BEV. It introduces RadarBEVNet, which uses a dual-stream radar backbone and an RCS-aware BEV encoder for radar feature extraction.



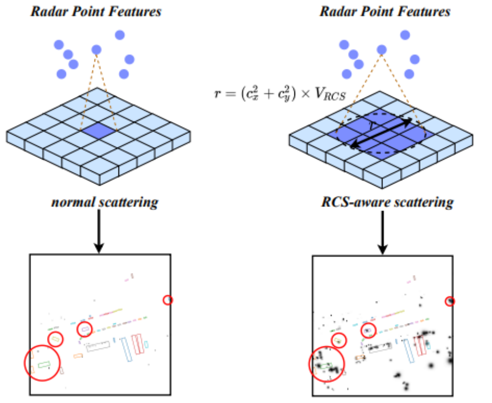


Figure 5. **Illustration of RCS-aware scattering.** RCS-aware scattering uses RCS as the object size prior to scatter the feature of one radar point to many BEV pixels.

Method	Input	Backbone	Image Size	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓	FPS↑
CenterFusion [30]	C+R	DLA34	448 × 800	45.3	33.2	0.649	**0.263**	0.535	0.540	**0.142**	-
CRAFT [12]	C+R	DLA34	448 × 800	51.7	41.1	0.494	0.276	0.454	0.486	0.176	4.1
RCBEVDet (Ours)	C+R	DLA34	448 × 800	**56.3**	**45.3**	**0.492**	0.269	**0.449**	**0.230**	0.188	**4.7**
RCBEV4d [50]	C+R	Swin-T	256 × 704	49.7	38.1	0.526	0.272	0.445	0.465	0.185	-
RCBEVDet (Ours)	C+R	Swin-T	256 × 704	**56.2**	**49.6**	**0.496**	**0.271**	**0.418**	**0.239**	**0.179**	18.2
CRN [13]	C+R	R18	256 × 704	54.3	**44.8**	0.518	**0.283**	0.552	0.279	0.180	27.9
RCBEVDet (Ours)	C+R	R18	256 × 704	**54.8**	42.9	**0.502**	0.291	**0.432**	**0.210**	**0.178**	**28.3**

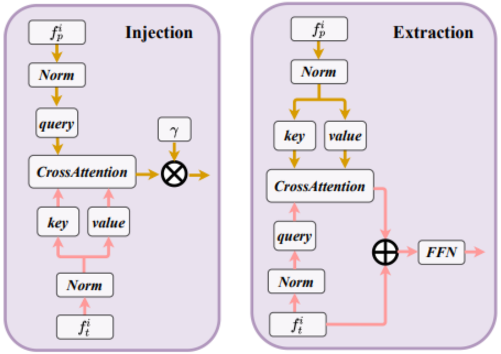


Figure 4. **Architecture of the Injection and Extraction module.** The left figure shows the details of the injection operation. The right figure displays the structure of the extraction operation.

Lin, Zhiwei, et al. "RCBEVDet: radar-camera fusion in bird's eye view for 3D object detection." CVPR 2024.

2.Related Works - Gated2Depth(Gated Camera) (CVPR 2019)

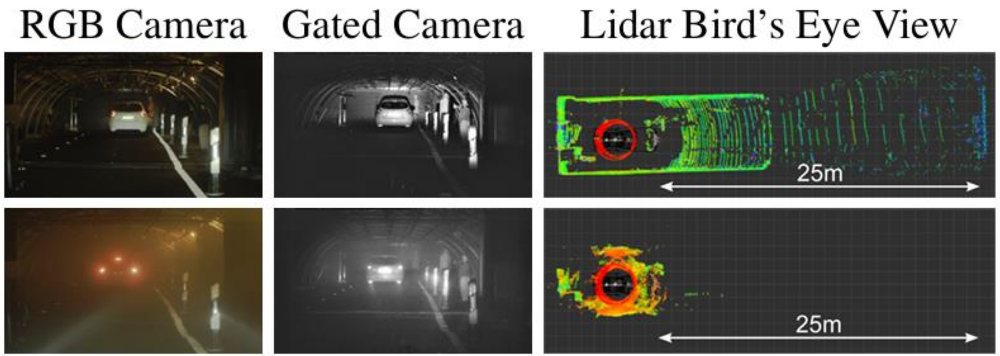
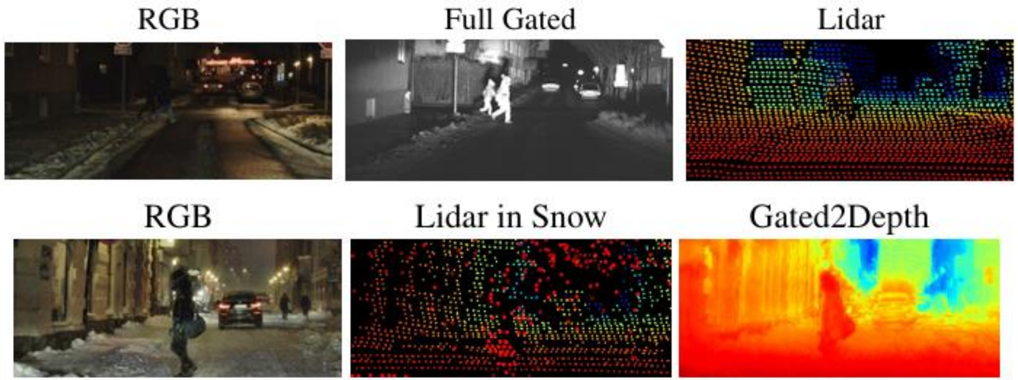


Figure 2: Sensor performance in a fog chamber with very dense fog. The first row shows recordings without fog while the second row shows the same scene in dense fog.

Standard RGB stereo camera (Aptina AR0230), lidar system (Velodyne HDL64-S3) and a **gated camera (BrightwayVision BrightEye)**



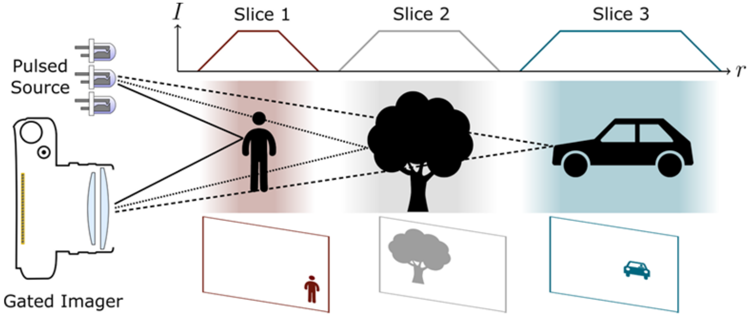
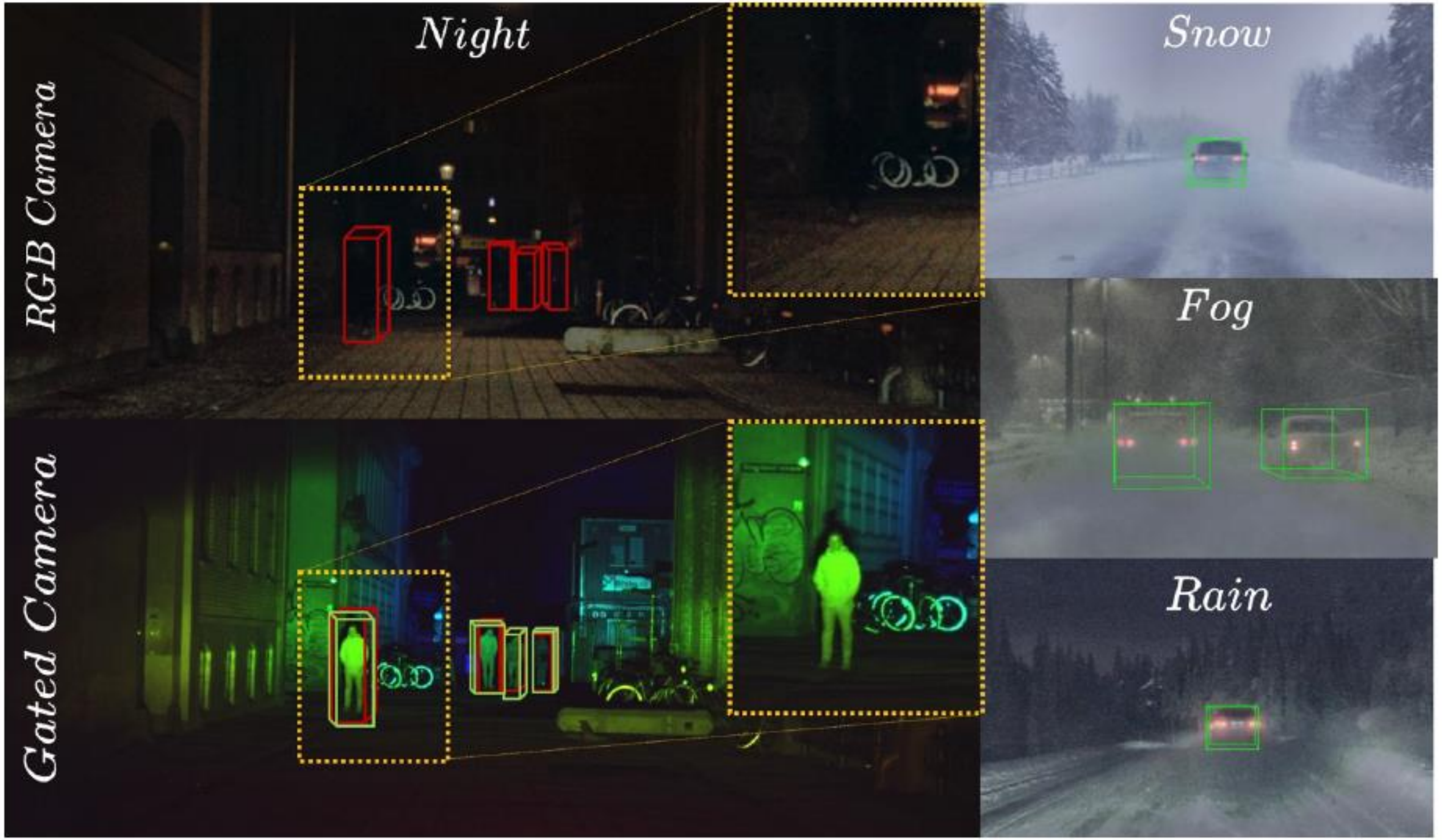


Figure 3: A gated system consists of a pulsed laser source and a gated imager that are time synchronized. By setting the delay between illumination and image acquisition, the environment can be sliced into single images that contain only a certain distance range.

Gruber, Tobias, et al. "Gated2depth: Real-time dense lidar from gated images." CVPR 2019.

3.Method - Challenging Adverse Weather Conditions



gated NIR, RGB color-imaging, LiDAR, and radar.

Ground truth bounding boxes in red
Predictions in green.

3.Method - The contributions of SAMFusion

- We propose a novel transformer-based multi-modal sensor fusion approach, **improving object detection in the presence of severe sensor degradation**.

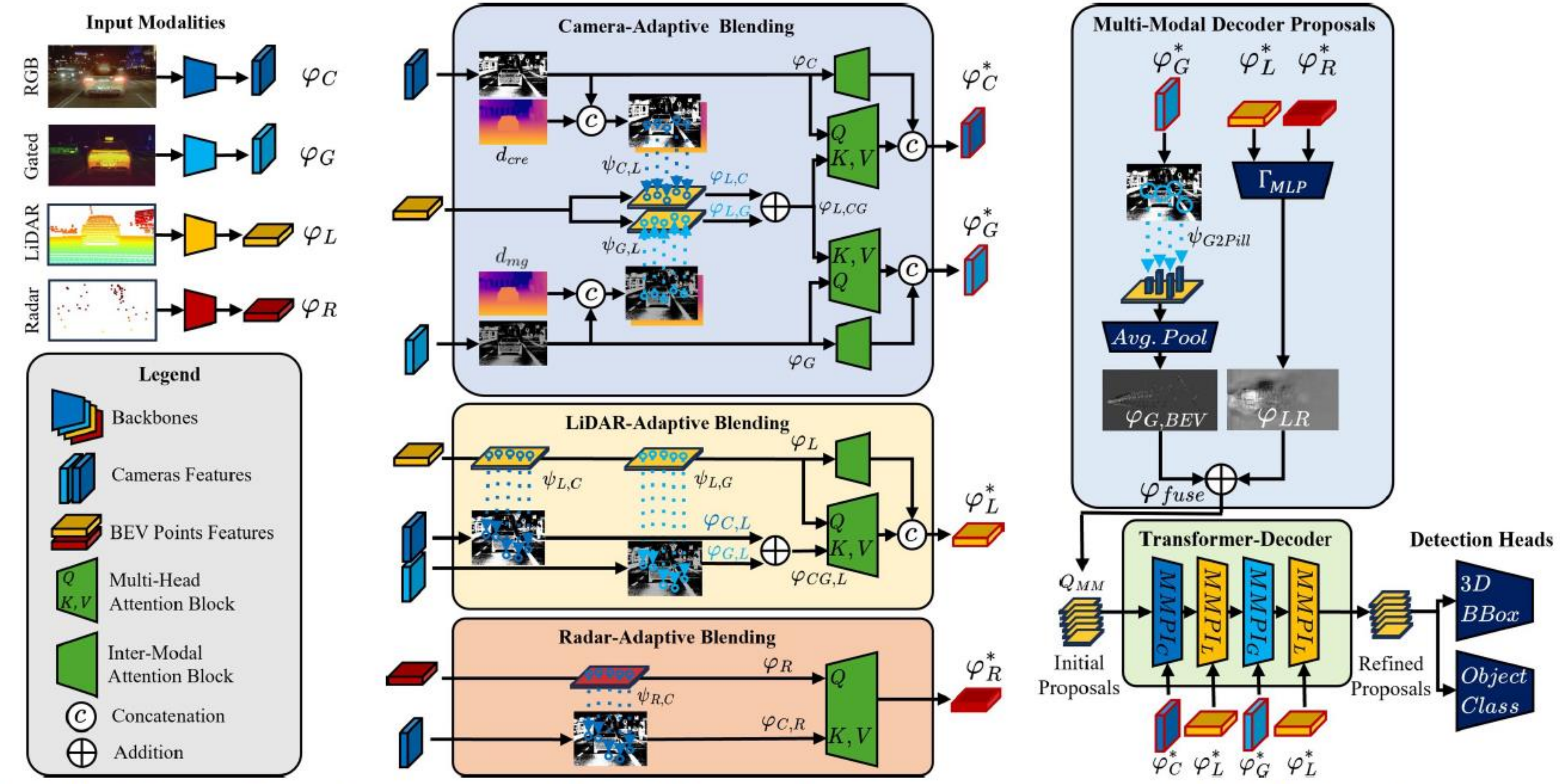
- We introduce an encoder architecture **combining early camera fusion, depthbased cross-modal transformation**, and **adaptive blending in conjunction with learned distance-weighted multimodal decoder proposals** to increase the reliability of object detection across lighting and weather conditions.

- We **design a transformer decoder that aggregates multimodal information in BEV** through multimodal proposal initialization.

- We validate the method on automotive **adverse weather scenes and improve 3D-AP**, especially for **the pedestrian class by more than 17.2 AP in dense fog and 15.62 AP in heavy snow on the most challenging distance category from 50 m-80** m relative to the state of the art

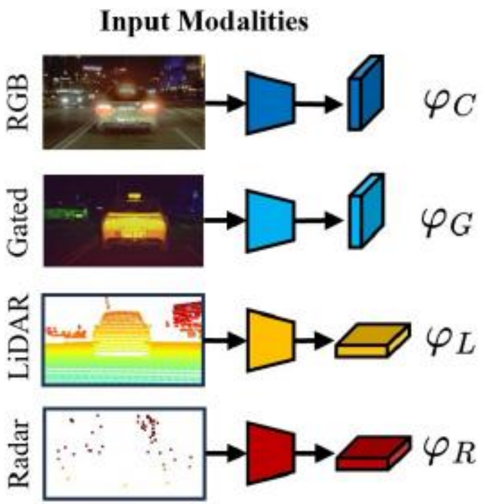

SAMFusion

3.Method - SAMFusion architecture for multimodal 3D object detection.

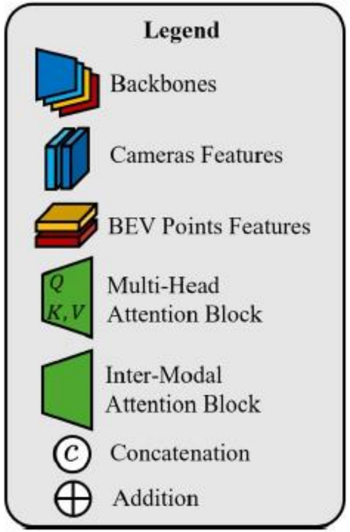


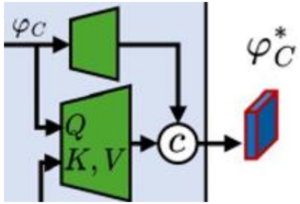
(a) Backbones, legend (b) SAM encoder (c) SAM decoder

3.Method - Backbones



$\varphi_L\ \varphi_R\ \in BEV\ coordinates$





cross-modal attention

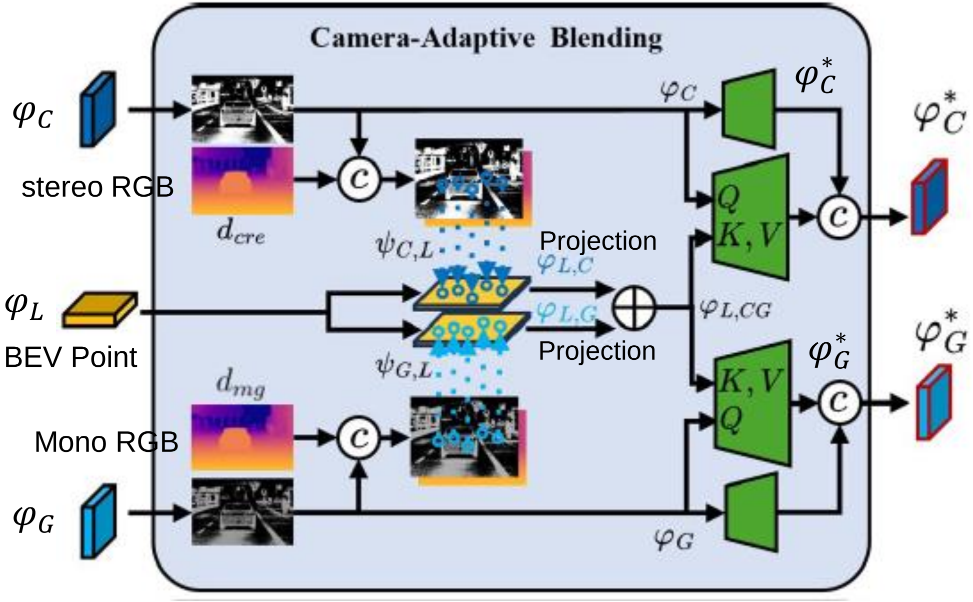
$$\varphi^*_{C;G} = \sum_{\varphi_{L,CG}\in J_s} softmax\left(\frac{\varphi_{C;G}\,\varphi^T_{L,CG}}{\sqrt{d}}\right)\varphi_{L,CG}.$$

Intra modal-attention

$$\varphi^*_{C;G} = \sum_{\varphi_{C;G}\in J_s} softmax\left(\frac{\varphi_{C;G}\,\varphi^T_{C;G}}{\sqrt{d}}\right)\varphi_{C;G}.$$

RGB/gated camera, LiDAR, radar are transformed into features through their respective feature extractors. By integrating these sensors into a **depth-based feature transformation**, a multi-modal query proposal and a decoder head, SAMFusion ensures robust and reliable 3D object detection across diverse scenarios.

3.Method – Camera-Adaptive Blending



$\psi_{C,L}$: The projection for RGB
$\psi_{G,L}$: The projection for Gated Camera

2D Image(depth + projection) → 3D Point
transform all the camera pixels (u, v) onto the LiDAR coordinate frame.

$$\begin{cases} z = \mathbf{d}(u, v), \\ x = (u - C_x) \times z / f_x, \\ y = (v - C_y) \times z / f_y, \end{cases}$$

(f_x, f_x) are the horizontal and vertical focal lengths of the camera and
(C_x, C_y) is the pixel location corresponding to the camera center

Each camera point is transformed into the LiDAR coordinate frame using the extrinsic matrix.
BEV Grid Projection + LiDAR feature sampling

LiDAR context fusion (based on both cameras)
$\psi_{C,L} \oplus \psi_{G,L} = \varphi_{L,CG}$

Queries from RGB and gated cameras are compared against weighted LiDAR context samples (RGB camera against Sampled LiDAR and gated camera against Sampled LiDAR).

3.Method – Camera-Adaptive Blending process

RGB/Gated Image + Depth concatenate

↓

3D projection $(\mathrm{u}, \mathrm{v}, \mathrm{d}) \rightarrow (x, y, z)$

↓

BEV coordinate projection (x, z)

↓

LiDAR $\varphi_L(x, z)$ Feature Sampling

↓

$\varphi_{L,C} \oplus \varphi_{L,G} \rightarrow \varphi_{L,CG}$

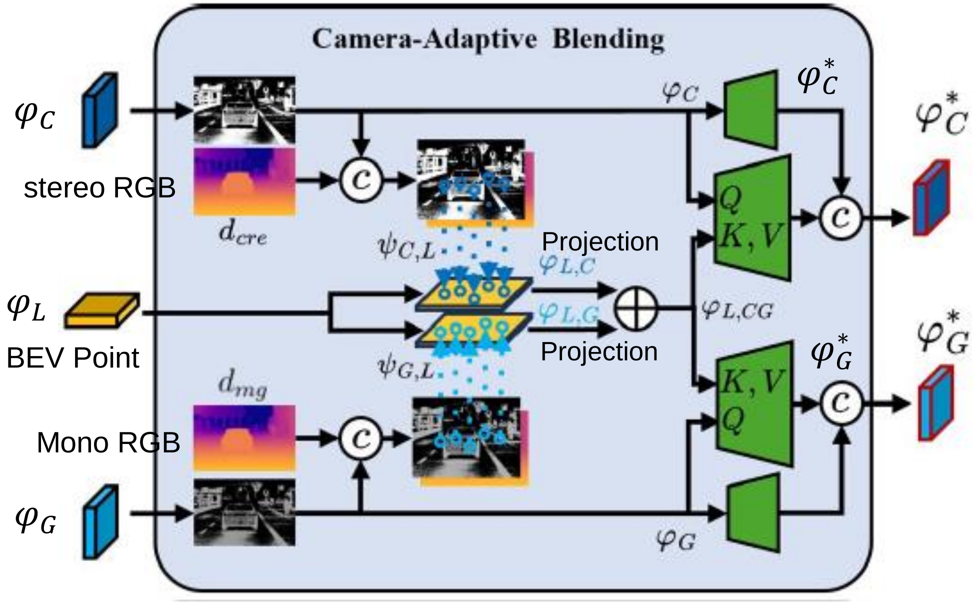
↓

Attention: $\varphi_C \longrightarrow \varphi_C^*$, $\varphi_g \longrightarrow \varphi_g^*$

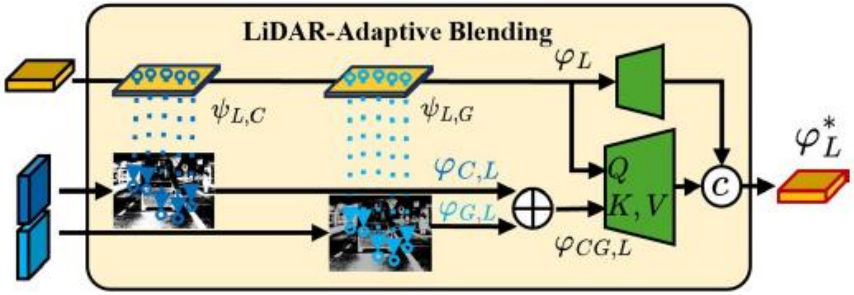
Cross-modal Enriched Features: φ_C $\varphi_{L,CG}$

↓

Concat Enriched Features: φ_C^*, φ_g^*



3.Method – LiDAR-Adaptive Blending & Radar-Adaptive Blending



In this module, we blend LiDAR features φ_L with a weighted context from RGB and gated camera features $\varphi_{CG,L}$ using attention, with **LiDAR features serving as queries and camera(+gated) features as keys and values**.

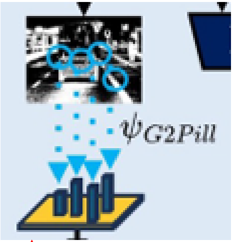
3D LiDAR features $\varphi_L(x_L, y_L, z_L)$ are mapped onto the corresponding 2D image points $(u_{C;G,L}, v_{C;G,L})$ by projection, through the $\psi_{L,C;G}$ LiDAR-to-camera (RGB; gated) projection matrix.

Blend the LiDAR-aware sampled image features from the two camera modalities

$$\varphi_{CG,L} = \varphi_{C,L} \oplus \varphi_{G,L}$$

3D Radar features $\varphi_R(x_R, y_R, z_R)$ are mapped onto the corresponding 2D image points $(u_{R,C}, v_{R,C})$ by projection, through the $\psi_{R,C}$ Radar-to-camera (RGB; gated) projection matrix.
Radar features serving as queries and camera features as keys and values.

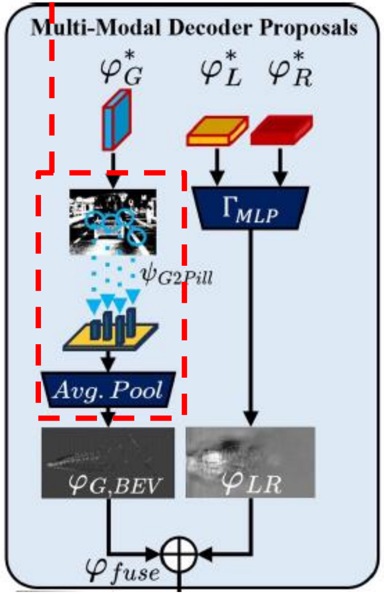
3.Method - Late Gated Camera Features Fusion.



Camera features are assigned to the corresponding LiDAR pillars

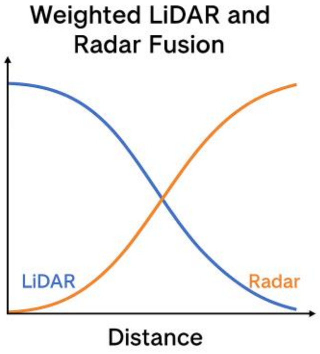
ψ_{G2Pill} 2D image feature → 3D BEV pillar feature

Gated Camera => BEV Grid Mapping => Avg Pool(BEV Grid < Image Feature)



Weighted Radar And LiDAR Feature Map Fusion.

The features of two sensors in a variable way,
to **dynamically adjust the ratio** so that
LiDAR Is more reliable at close distances
Radar is more reliable at longer distances.



$$f = \exp((-\frac{d}{2\sigma^2})^2)$$

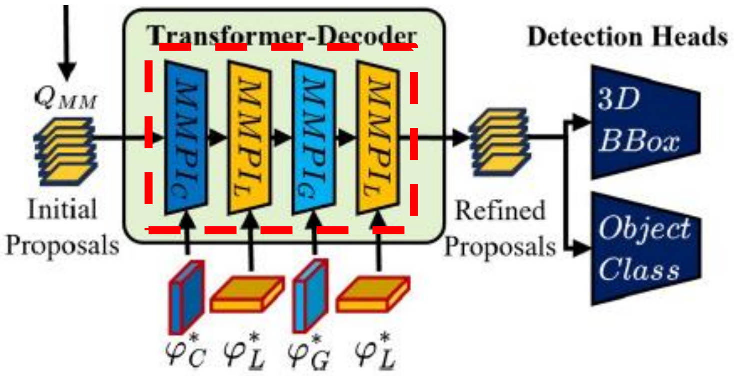
Distance-based weighting functions

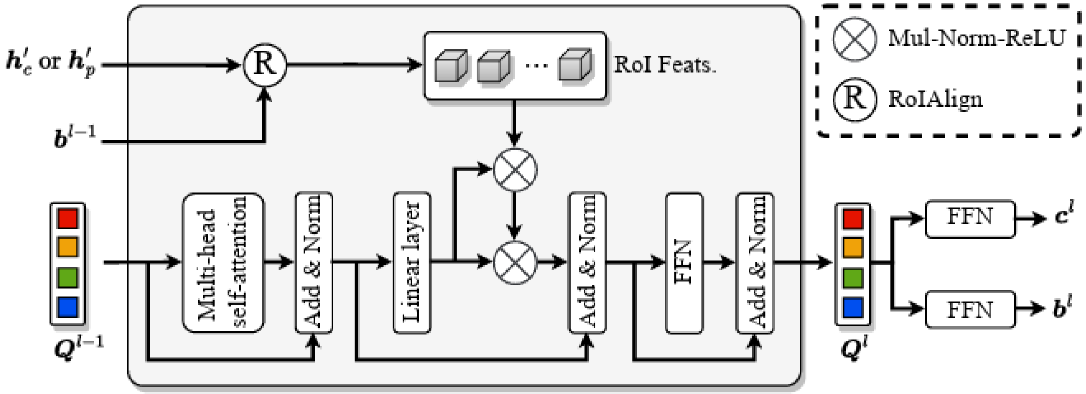
d : Distance of each feature point from the ego vehicle

σ : Variance (learned parameter)

$$\varphi_{LR} = \Gamma_{MLP}(f(d,\sigma)\varphi_L^* + (1 - f(d,\sigma))\varphi_R^*)$$

3.Method – MMPI module (Deepinteraction - NeurIPS 2022)





(b) Multi-modal predictive interaction layer (MMPI)

Multi-modal predictive interaction layer (MMPI) For the l-th decoding layer, the set prediction is computed by taking the object queries $\left\{Q_n^{(l-1)}\right\}_{n=1}^N$ and the bounding box predictions $\left\{b_n^{(l-1)}\right\}_{n=1}^N$ from previous layer as inputs and enabling interaction with the intensified image h'_p or LiDAR h'_c representations (h'_c if l is odd, h'_p if l is even). We formulate the multi-modal predictive interaction layer (Figure 3(b)) for specific modality as follows:

Yang, Zeyu, et al. "Deepinteraction: 3d object detection via modality interaction." NeurIPS 2022.

4.Experiments – Implementation Details

Framework : Pytorch, MMDetection3D
Camera branch backbone : Initialized ResNet-50
Pretrained weight : Cascade Mask R-CNN
Input image size : RGB, Gated Image [800,400] (cener-based cropping – reduce computational cost)
Voxel size : 0.075m deep, 0.075m wide and 0.2m high.
LiDAR point clouds : (0 m, 100 m) in range, (-40 m, 40 m) in width and the height range (-3 m, 1m)
Radar point clouds : (0 m, 100 m) in range, (-40 m, 40 m) in width and the height range (-0.2 m, 0.4 m)
Decoder layers : four stacked transformer, guided by RGB, gated camera, and LiDAR modalities with 200 initial multi-modal proposals.

We train all models for 12 epochs in an end-to-end manner with a batch size of 4 on NVIDIA V100 GPUs.

MULTI MODAL FEATURE MAP WEIGHTING

Layer #	Component	Sigmoid mask	Output Shape
0_a	Convfuser $(\varphi_L^*, \Gamma_{MLP})$	✓	$128 \times 180 \times 180$
0_b	Convfuser $(\varphi_R^*, \Gamma_{MLP})$	✓	$128 \times 180 \times 180$
1	Convfuser $(0_a, 0_b)$	✗	$128 \times 180 \times 180$
Combined feature map φ_{fuse}		**Shape:**	$128 \times 180 \times 180$

FEATURE MAP BLENDING MODULE

Layer #	Layer Description	Output Shape
Convfuser	Conv2d (3x3)	$128 \times 180 \times 180$
	GroupNorm (num_groups=16)	
	ReLU	
	Conv2d (3x3)	
	GroupNorm (num_groups=16)	
	ReLU	
	Conv2d (3x3)	
	GroupNorm (num_groups=16)	
	ReLU	

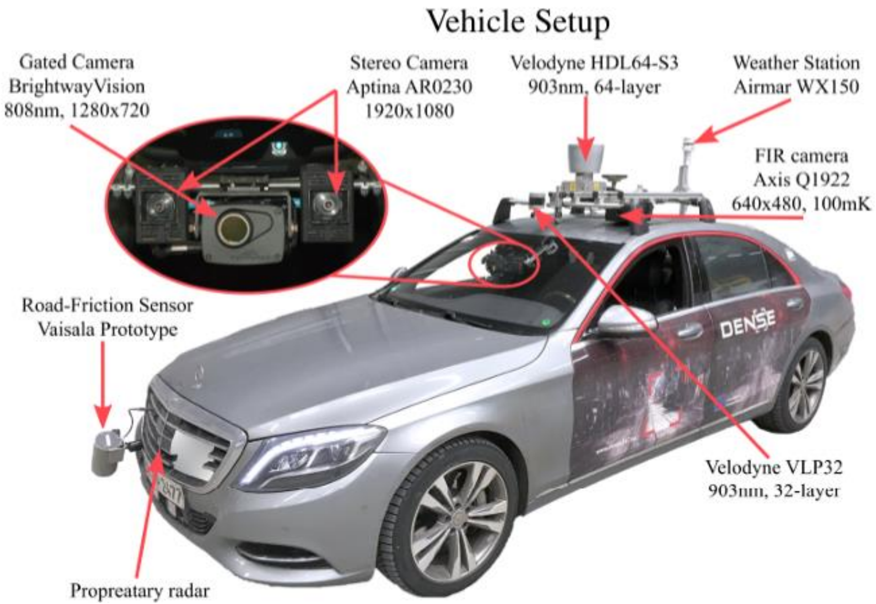
4.Experiments – Dataset and Evaluation Metrics

The **SeeingThroughFog Dataset**
2,997 annotated samples in adverse weather conditions,
covering night, fog, and snowy scenarios.

Following prior research(Gated3D),
we divide the dataset into **10,046 samples for training,**
1,000 for validation, and 1,941 for testing.
The test split is further divided into **1,046 daytime**
and 895 nighttime samples, with respective weather splits.

Evaluation Metrics.
Object detection performance is evaluated according to the metrics specified in the **KITTI evaluation framework, including 3D-AP and BEV-AP for the passenger car and pedestrian class**.
We incorporate 40 recall positions for the AP calculation. To match the predictions and ground truth we apply intersection over union (IoU) with an IoU of 0.2 for passenger cars and 0.1 for pedestrians.
Further, we follow and report results according to respective distance bins.

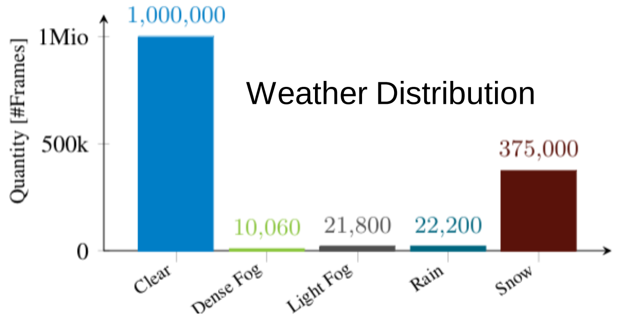
4.Experiments – Seeing Through Fog (CVPR 2020)

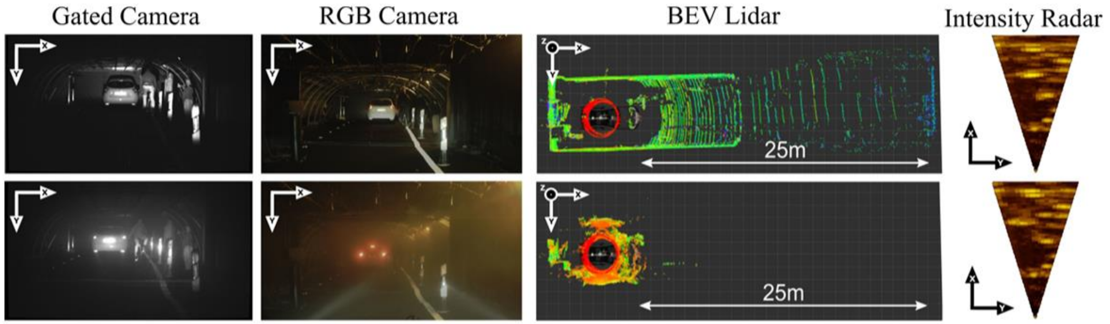


Geographical coverage of the data collection campaign covering two months and 10,000km in Germany, Sweden, Denmark, and Finland.

Dataset / Sensor Setup	KITTI [19]	BDD [69]	Waymo [59]	NuScenes [6]	Ours
RGB Cameras	2	1	5	6	2
RGB resolution	1242×372	1280×720	1920×1080	1600x900	1920x1024
Lidar Sensors	1	✗	5	1	2
Lidar resolution	64	0	64	32	64
Radar Sensor	✗	✗	✗	4	1
Gated Camera	✗	✗	✗	✗	1
FIR Camera	✗	✗	✗	✗	1
Frame rate	10 Hz	30 Hz	10 Hz	1 Hz/10 Hz	10 Hz
Dataset Statistics					
Labeled Frames	15K	100k	198k	40K	13.5K
Labels	80k	1.47M	7.87M	1.4M	100K
Scene Tags	✗	✓	✗	✓	✓
Night Time	✗	✓	✓	✓	✓
Light Weather	✗	✓	✗	✓	✓
Heavy Weather	✗	✗	✗	✗	✓
Fog Chamber	✗	✗	✗	✗	✓

Table 1: Comparison of the proposed multimodal adverse weather dataset to existing automotive detection datasets.





Bijelic, Mario, et al. "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather." CVPR 2020.

4.Experiments – nuScenes (CVPR 2020)

Sensor	Details
6x Camera	RGB, 12Hz capture frequency, 1/1.8” CMOS sensor, 1600×900 resolution, auto exposure, JPEG compressed
1x Lidar	Spinning, 32 beams, 20Hz capture frequency, 360° horizontal FOV, -30° to 10° vertical FOV, $\leq 70m$ range, $\pm$2cm accuracy, up to $1.4M$ points per second.
5x Radar	$\leq 250m$ range, 77GHz, FMCW, 13Hz capture frequency, $\pm$0.1km/h vel. accuracy
GPS & IMU	GPS, IMU, AHRS. 0.2° heading, 0.1° roll/pitch, 20mm RTK positioning, 1000Hz update rate

Table 2. Sensor data in nuScenes.

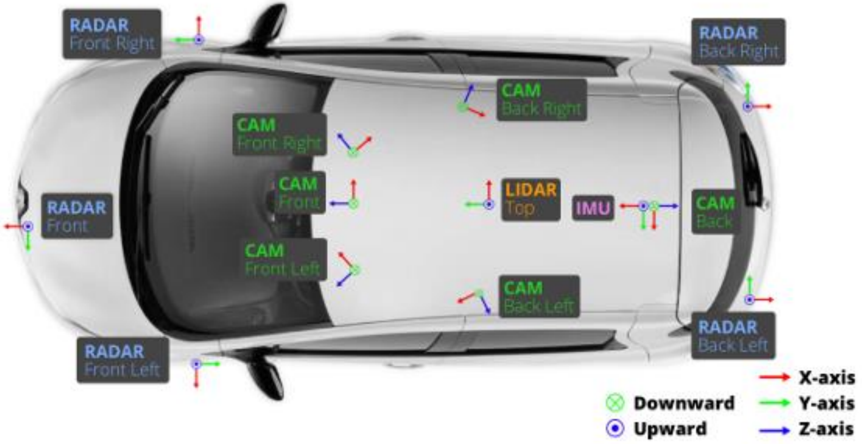


Figure 4. Sensor setup for our data collection platform.

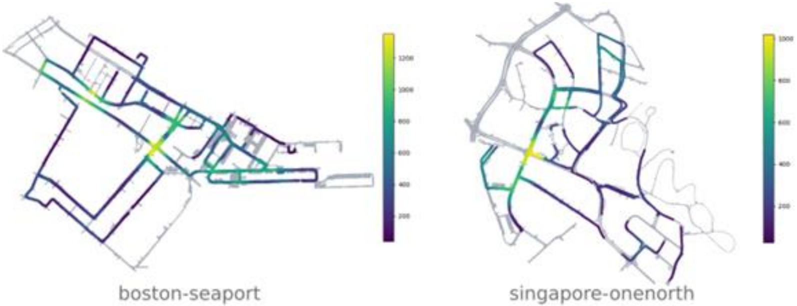
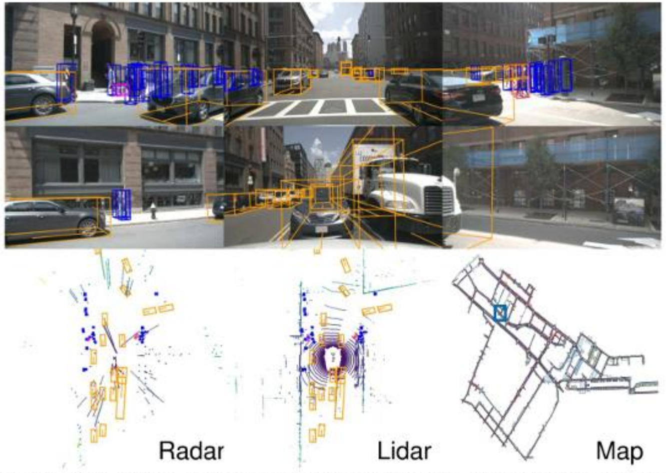


Figure 5. Spatial data coverage for two nuScenes locations. Colors indicate the number of keyframes with ego vehicle poses within a 100m radius across all scenes.



"Ped with pet, bicycle, car makes a u-turn, lane change, peds crossing crosswalk"

Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." CVPR. 2020.

4.Experiments – Comparison of benchmark datasets

Category	KITTI	nuScenes	Seeing Through Fog (STF)
Object Classes	Car, Van, Pedestrian, Cyclist, etc.	Car, Truck, Bus, Pedestrian, Bicycle, Motorcycle, etc.	Focus on Car, Pedestrian
Evaluation Unit	Per-frame 3D bounding box	Includes object tracking unit (detection + tracking)	3D bounding box (per-frame), includes annotations under different weather conditions
GT Labeling Criteria	Valid only if LiDAR point count ≥ 5, others treated as "don't care"	All objects labeled, includes metadata such as visibility, score	Pedestrian labeled even with 1–2 points (focus on completeness)
"Don't Care" Region	Clearly defined. No GT box=ignore surroundings	None. All included in evaluation	Vehicles with **insufficient points treated as "don't care"**
Evaluation Metrics	AP@IoU 0.7 (Car), 0.5 (Pedestrian)	mAP, mATE, mASE, mAAE, mAVE, NDS and other diverse metrics	AP@IoU 0.5, evaluated by distance range (0–30m, 30–50m, 50–80m)
Weather/Lighting Tags	None (all clear weather)	Some night/rain included, but **mostly clear conditions**	Includes weather condition tags **(Clear, Light Fog, Dense Fog, Snow)**
Occluded Object Handling	Not labeled	Includes occlusion level, visibility score	Pedestrians labeled even with poor visibility
Number of Cameras	2 (Stereo)	6-camera surround view	2 (Stereo) + Gated camera + FIR camera
LiDAR Resolution	Velodyne HDL-64E (64 channels)	Velodyne HDL-32E (32 channels)	Mix of HDL-64E + VLP-32C
Sensor Configuration	RGB + LiDAR	RGB (6) + LiDAR + RADAR	RGB + LiDAR + Radar + Gated NIR + FIR

4.Experiments – Evaluation of SAMFusion detection performance

Average Precision for *Pedestrian* class

Method	Modality	Day 3D object detection 0-30m	Day 3D object detection 30-50m	Day 3D object detection 50-80m	Day BEV detection 0-30m	Day BEV detection 30-50m	Day BEV detection 50-80m	Night 3D object detection 0-30m	Night 3D object detection 30-50m	Night 3D object detection 50-80m	Night BEV detection 0-30m	Night BEV detection 30-50m	Night BEV detection 50-80m
M3D-RPN [6]	C	26.20	14.50	9.84	30.68	17.47	10.07	25.09	6.43	2.07	26.42	7.69	2.74
PatchNet [48]	G	32.88	18.05	5.62	39.45	20.27	9.77	15.37	13.37	6.75	21.60	18.15	8.46
Gated3D [31]	G	50.94	20.59	14.14	53.26	22.15	16.51	48.53	23.99	14.98	49.82	25.57	15.46
Stereo-RCNN [36]	S	48.58	23.26	7.77	50.11	25.10	8.38	46.09	21.63	11.57	47.58	25.47	11.84
SECOND [80]	L	70.75	51.81	19.34	71.05	52.51	20.28	69.04	48.09	14.56	70.51	49.23	15.32
MVXNet [62]	CL	74.51	61.69	29.78	74.88	62.63	30.54	74.15	55.66	23.19	74.42	55.90	23.58
BEVFusion [42]	CL	64.25	57.91	8.86	64.76	59.41	8.86	65.78	52.91	7.25	66.25	54.40	7.27
DeepInteraction [83]	CL	78.01	66.59	28.55	77.98	66.67	28.54	71.98	61.10	20.53	71.96	61.29	20.72
SparseFusion [77]	CL	68.27	60.18	16.89	68.18	60.32	16.92	61.11	57.09	12.67	61.21	57.24	12.66
SAMFusion	**CGLR**	**80.09**	**70.97**	**40.16**	**79.97**	**70.99**	**40.35**	**75.49**	**67.59**	**27.14**	**75.49**	**67.56**	**27.16**

Average Precision for *Car* class

Method	Modality	Day 3D object detection 0-30m	Day 3D object detection 30-50m	Day 3D object detection 50-80m	Day BEV detection 0-30m	Day BEV detection 30-50m	Day BEV detection 50-80m	Night 3D object detection 0-30m	Night 3D object detection 30-50m	Night 3D object detection 50-80m	Night BEV detection 0-30m	Night BEV detection 30-50m	Night BEV detection 50-80m
M3D-RPN [6]	C	53.21	13.26	10.52	60.80	16.16	10.52	51.18	20.76	2.73	52.53	21.39	2.74
PatchNet [48]	G	23.91	10.86	7.34	24.87	11.33	7.84	23.74	16.79	7.16	25.15	17.76	8.29
Gated3D [31]	G	52.15	28.31	14.85	52.31	29.26	15.02	51.42	25.73	12.97	53.37	29.13	13.12
Stereo-RCNN [36]	S	54.17	17.16	6.17	57.92	17.69	6.26	47.36	17.21	13.02	53.81	18.34	13.08
SECOND [80]	L	95.68	81.90	46.81	95.70	82.18	47.55	98.01	84.10	48.53	98.03	84.23	50.39
MVXNet [62]	CL	96.29	84.09	50.35	96.30	84.09	51.83	96.36	85.99	**49.79**	96.36	86.06	**51.17**
BEVFusion [42]	CL	95.30	86.86	11.43	95.43	87.38	11.24	93.89	84.84	12.17	93.95	85.31	12.48
DeepInteraction [83]	CL	97.12	87.95	**51.84**	97.13	88.47	**51.99**	98.31	88.09	46.83	98.31	88.11	46.87
SparseFusion [77]	CL	**97.47**	88.10	31.02	**97.49**	88.26	31.11	96.12	86.49	27.99	96.13	86.51	28.01
SAMFusion	**CGLR**	97.25	**89.50**	50.68	97.26	**89.69**	50.80	**98.77**	**88.91**	44.40	**98.82**	**89.16**	45.46

SoTA mono- and multi-modal methods based on the car and pedestrian classes on the SeeingThoughFog test set.

Objects with fewer than five LiDAR points are excluded from evaluation, so correct detections in challenging conditions (e.g., fog, long distance) may be underestimated.

In contrast, the pedestrian class prioritizes completeness by labeling as many objects as possible, even with few LiDAR points.

Only clear objects are labeled, so detection performance may be underestimated.

4.Experiments – Ablation study

(a) Ablation of Input Modality configurations.

	Input Modality	Proposal Modality	Day 3D object detection 30-50m	Day 3D object detection 50-80m	Night 3D object detection 30-50m	Night 3D object detection 50-80m
Ablation	CL	L	66.59	28.55	61.10	20.80
	GL	L	65.59	26.89	63.25	22.11
	CGL	L	66.88	28.94	64.17	22.34
	CLR	LR	69.06	35.02	65.97	20.95
	GLR	LR	69.52	32.17	67.05	24.40
	CGLR	LR	69.98	35.60	67.22	26.85
	CGLR	GLR	**70.99**	**40.16**	**67.56**	**27.14**

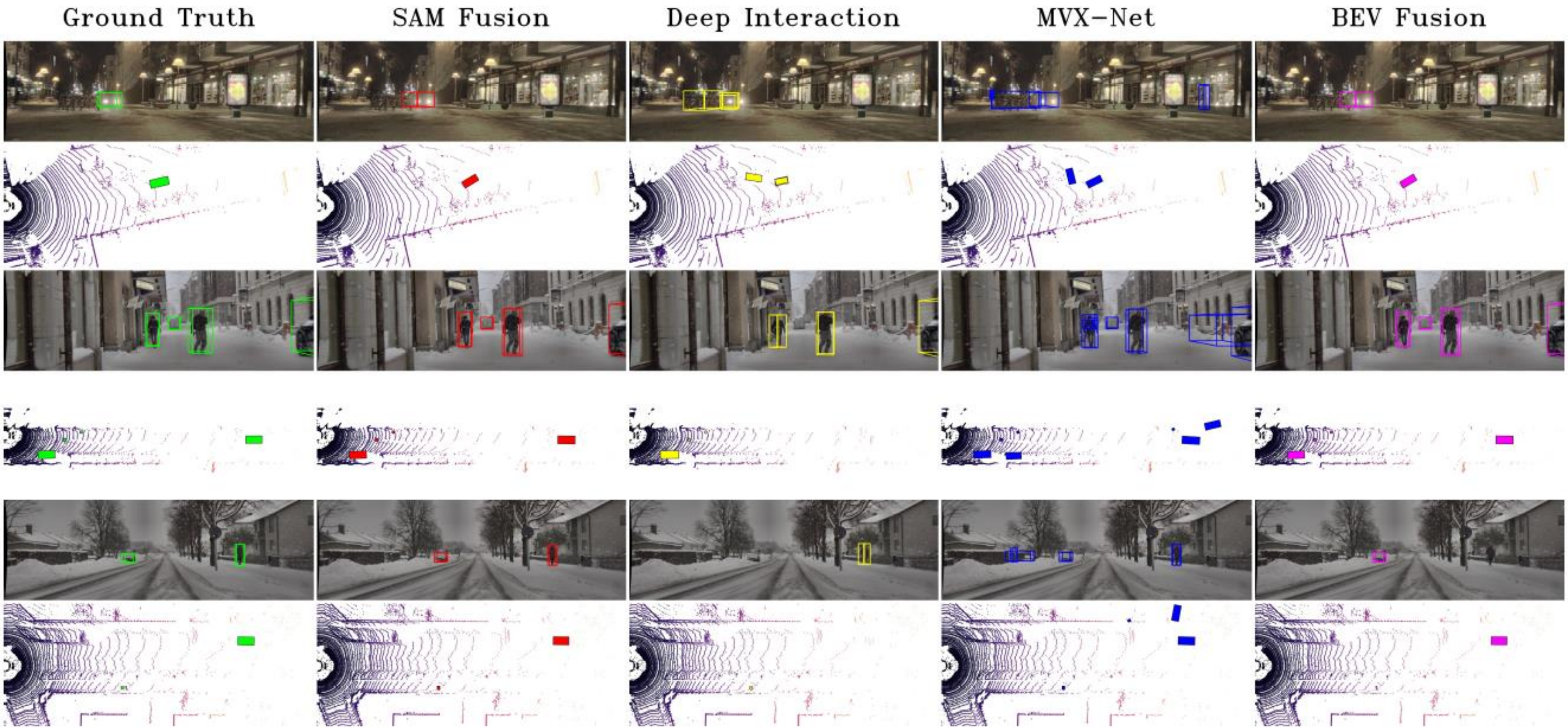
(b) Ablation of SAMFusion components.

	Input Modality	Depth-based Transformation	Proposal Modality C	G	R	L	Γ_{MLP}	Day 50-80m	Night 50-80m
Ablation	CGLR	✗	✗	✗	✗	✓	✗	28.94	22.34
	CGLR	✗	✗	✗	✓	✓	✗	29.48	23.02
	CGLR	✓	✗	✗	✓	✓	✗	29.49	24.01
	CGLR	✓	✗	✗	✓	✓	✓	35.60	26.85
	CGLR	✓	✓	✗	✓	✓	✓	36.19	22.79
	CGLR	✓	✗	✓	✓	✓	✓	**40.16**	**27.14**

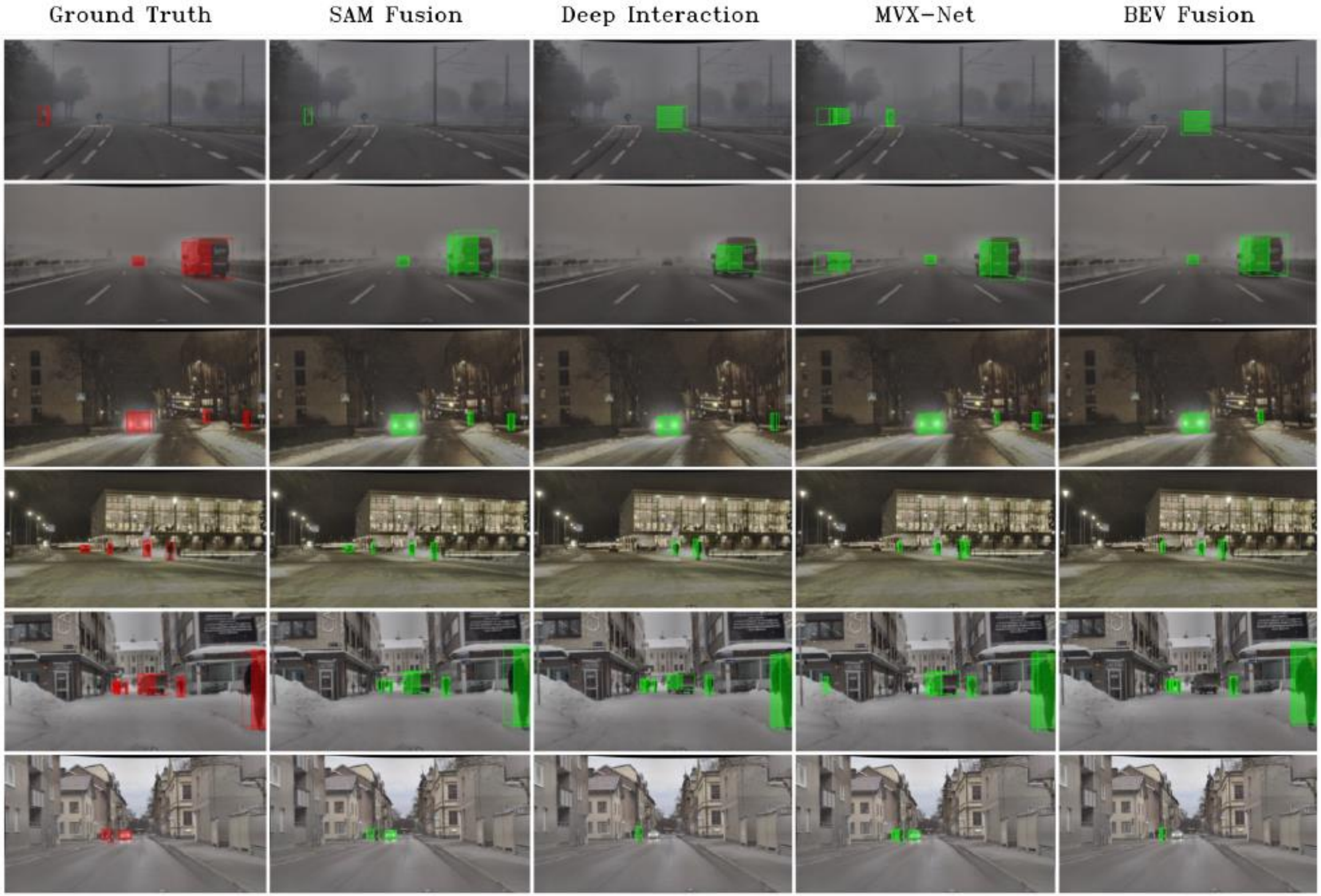
Table validates the proposed method in adverse weather, like snow and fog.(reduced number of road users in these weather

Method	Modality	Pedestrian Snow 0-30m	Pedestrian Snow 30-50m	Pedestrian Snow 50-80m	Pedestrian Fog 0-30m	Pedestrian Fog 30-50m	Pedestrian Fog 50-80m	Car Snow 0-30m	Car Snow 30-50m	Car Snow 50-80m	Car Fog 0-30m	Car Fog 30-50m	Car Fog 50-80m
		Average Precision for *Pedestrian* class: Snow 3D Object Detection			Fog 3D Object Detection			Average Precision for *Car* class: Snow 3D Object Detection			Fog 3D Object Detection		
MVXNet [62]	CL	76.23	59.73	25.83	73.89	50.98	16.73	95.82	86.02	50.28	92.81	84.62	52.30
BEVFusion [42]	CL	71.12	62.61	10.01	76.24	58.04	8.61	92.55	89.74	10.79	92.20	84.04	13.97
DeepInteraction [83]	CL	72.91	57.56	18.38	66.62	50.32	10.64	95.36	82.05	56.21	95.44	83.55	49.30
SparseFusion [77]	CL	73.33	66.84	19.87	79.25	58.39	17.05	96.79	91.35	32.11	95.81	87.71	25.16
SAMFusion	**CGLR**	**87.44**	**80.51**	**41.45**	**83.18**	**66.96**	**34.31**	**97.36**	**93.06**	**56.22**	**96.50**	**92.41**	**52.99**
Improvement in AP		+11.2	+13.6	+15.62	+3.9	+8.5	+17.2	+0.5	+1.7	+0.01	+0.7	+4.6	+0.7

4.Experiments – Qualitative results (adverse weather)



4.Experiments – Qualitative results (different sequences)



4.Experiments – Additional Results

The model enhances performance in adverse weather while maintaining accuracy in normal conditions.

Method	Modality	mAP ↑	NDS ↑
FUTR3D [7]	CL	64.5	68.3
AMVP [24]	CL	67.1	70.8
AutoAlignV2 [8]	CL	67.1	71.2
TransFusion [1]	CL	67.5	71.3
BEVFusion [16]	CL	67.9	71.0
BEVFusion [18]	CL	68.5	71.4
DeepInteraction [23]	CL	**69.9**	**72.7**
SAMFusion	CLR	<u>68.6</u>	<u>71.7</u>

Table 2: Results on nuScenes dataset validation split.

Model	Inference time [ms] ↓	Frames per Second ↑
MVXNet [21]	74.0	13.5
BEVFusion [18]	<u>57.4</u>	<u>17.5</u>
DeepInteraction [23]	**48.3**	**20.7**
SAMFusion	70.7	14.3

Table 4: Inference time comparison to existing multi-modal detection methods.

Method	Modality	mAP ↑	NDS ↑
DeepInteraction [23]	CL	<u>56.6</u>	<u>64.6</u>
SAMFusion	CLR	**58.8**	**65.6**

Table 3: Results on nuScenes dataset validation split for detections on the 20-50 meters range.

5.Conclusion & Limitation

(+) The first research on **four-sensor fusion integrating** camera, gated camera, radar, and LiDAR.

(+) Sensor Reliability Estimation Module : **Learns an adaptive weighting for each sensor modality** based on quality indicators. (Sensor-Specific Encoders)

(+) Uses a late-fusion approach with a **shared feature space in Bird's-Eye View (BEV)**, allowing flexible integration of feature maps from multiple sensors.

(+) Introduces a **robust sensor-level architecture** for fog, snow, and heavy rain, significantly enhancing detection of narrow-profile and vulnerable road users in low-light and **adverse weather conditions**.

(-) As the number of sensor types increases, **computational overhead also increases**.

(-) Solving the problem with **a single sensor is more practical** than using multimodal sensors.

(-) The performance naturally **improves as the number of input modalities increases**.

※ How about applying event cameras from a multimodal perspective?

Thanks
Any Questions?

You can send mail to
Susang Kim(healess1@gmail.com)