Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

Object Detection in Adverse Weather Conditions using Tightly-coupled Data-driven Multimodal Sensor Fusion

R&D Defense
February 15, 2024

Kevin Patel

Advisors
Prof. Dr.-Ing. Sebastian Houben
Santosh Thoduka, M.Sc.

Outline

- Introduction
- Challenges
- Related Works
- Methodology
- Evaluation and Results
- Conclusion
- Future Works

Object Detection in Adverse Weather Conditions using Tightly-coupled Data-driven Multimodal Sensor Fusion

2D object detection
Car, truck, pedestrian, cycle

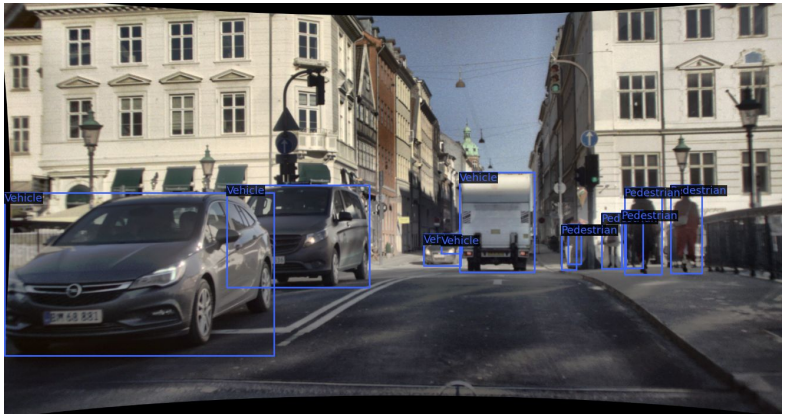


Figure 1: A sample from DENSE dataset [1]

[1] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, “Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather,” pp. 11 682–11 692, 2020.

Object Detection in **Adverse Weather Conditions** using Tightly-coupled Data-driven Multimodal Sensor Fusion

Such as fog, snow, rain, night

Figure 2: Dense fog, snow, and rain – from left to right (samples from DENSE dataset [1])

[1] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, “Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather,” pp. 11 682–11 692, 2020.

Object Detection in Adverse Weather Conditions using **Tightly-coupled** Data-driven Multimodal Sensor Fusion

How different modalities are combined at what level
Such as Early fusion, mid fusion/feature fusion, late fusion, etc.

Object Detection in Adverse Weather Conditions using Tightly-coupled **Data-driven** Multimodal Sensor Fusion

Supervised learning with publicly available datasets

Object Detection in Adverse Weather Conditions using Tightly-coupled Data-driven **Multimodal** Sensor Fusion

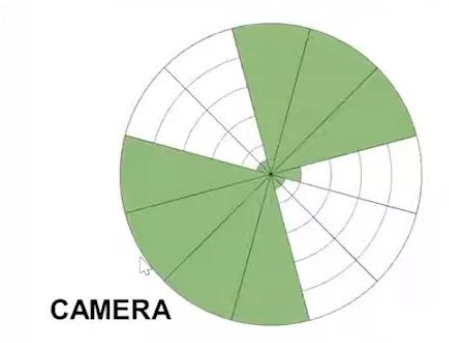
Different sensor data types
Such as point cloud, image, etc.

Object Detection in Adverse Weather Conditions using Tightly-coupled Data-driven Multimodal **Sensor Fusion**

Integration of multisensor data for enhanced situational understanding

What to Fuse?

What to Fuse? - Camera



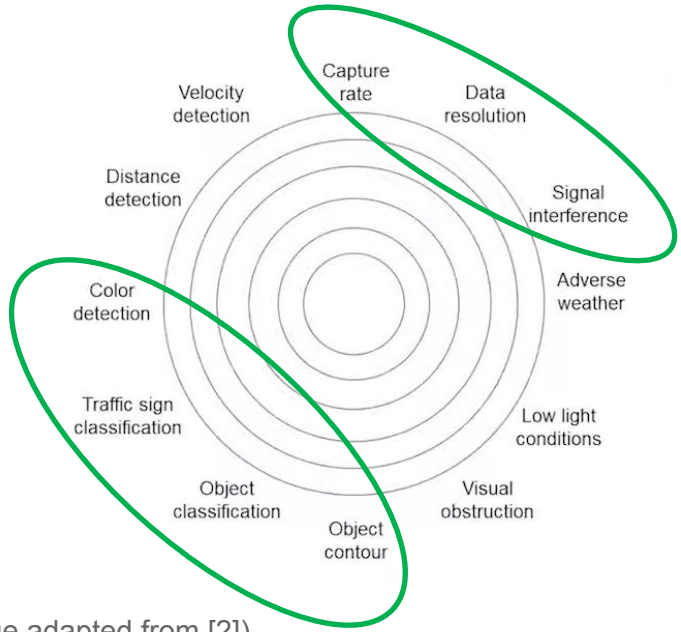
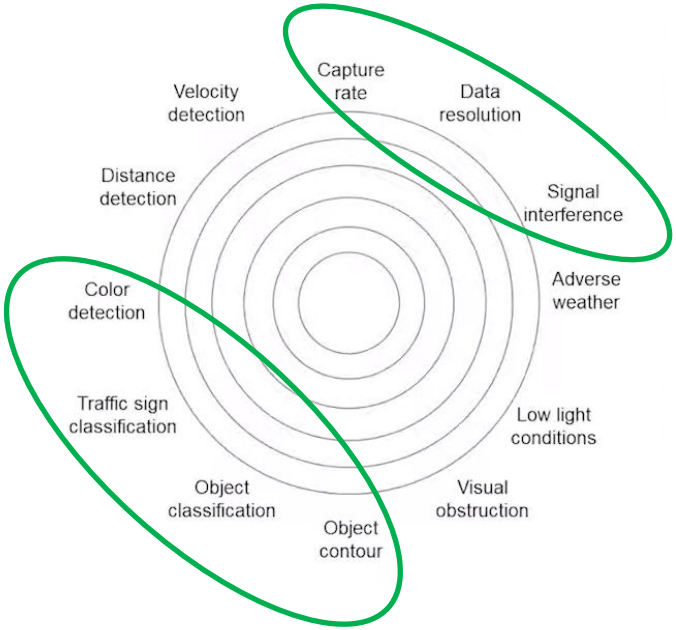


Figure 3: Sensors' modality characteristics [Image adapted from [2])

[2] R. Laganière, "Sensor fusion for autonomous vehicles: Strategies, methods, and tradeoffs," 2022, accessed on 18.12.2022. [Online]. Available: https://youtu.be/2Fcmh7SLPBI

What to Fuse? - Camera

Figure 3: Van occluded by a water droplet on the lens (Image source [3])



[3] Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, “A Deep Learning-based Radar and Camera Sensor Fusion Architecture for Object Detection,” arXiv, May 2020. [Online]. Available: https://arxiv.org/abs/2005.07431v1

What to Fuse? - LiDAR

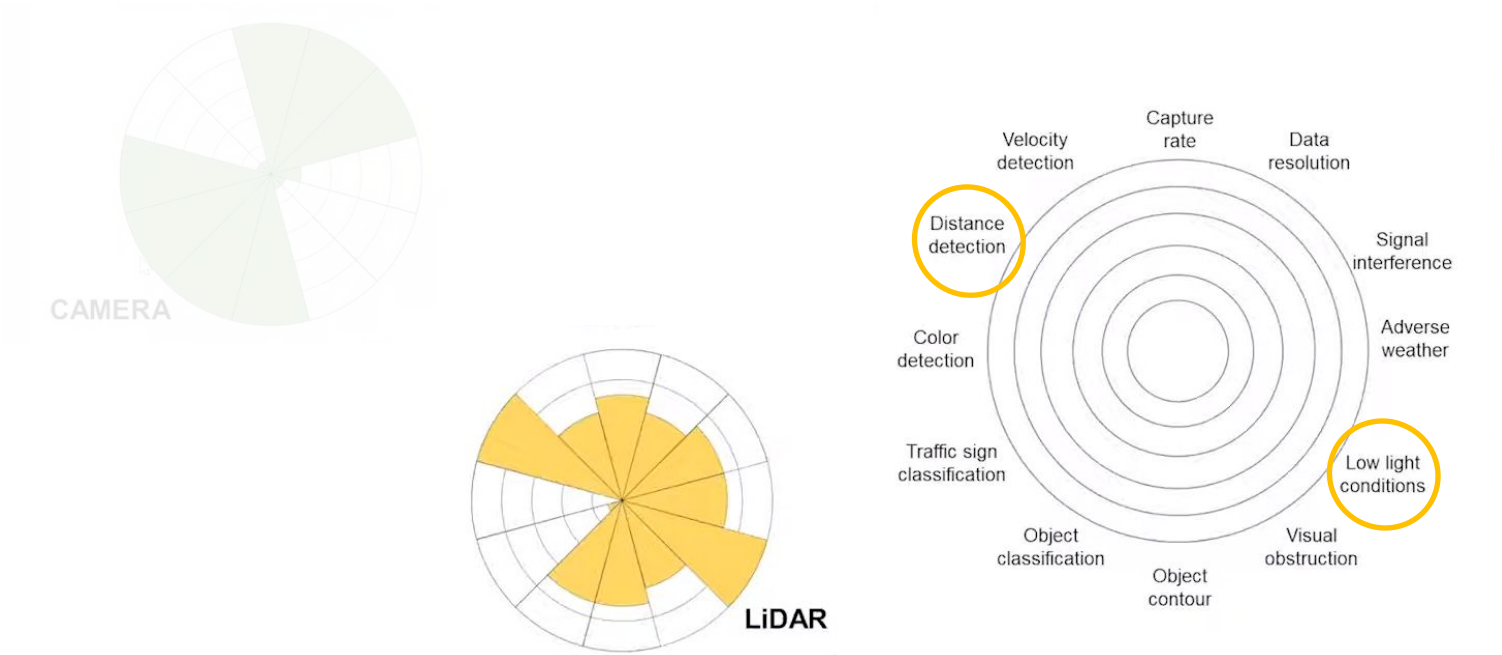
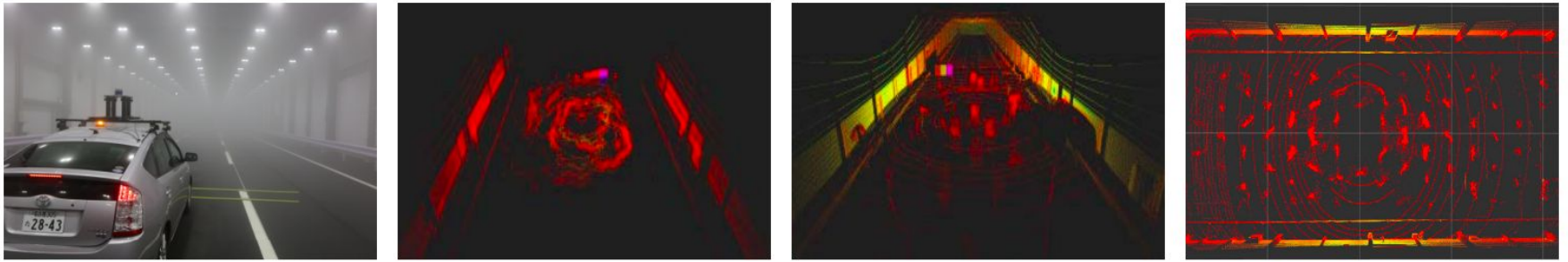


Figure 3: Sensors' modality characteristics [Image adapted from [2])

[2] R. Laganière, "Sensor fusion for autonomous vehicles: Strategies, methods, and tradeoffs," 2022, accessed on 18.12.2022. [Online]. Available: https://youtu.be/2Fcmh7SLPBI

What to Fuse? - LiDAR



(a) Fog chamber (b) Point clouds in fog (c) Point clouds in rain (d) Rain pillars as detected by LiDAR

Figure 4: LiDAR performance test (a sample from LIBRE [3] dataset)

[3] A. Carballo, J. Lambert, A. Monrroy, D. Wong, P. Narksri, Y. Kitsukawa, E. Takeuchi, S. Kato, and K. Takeda, “Libre: The multiple 3d lidar dataset,” IEEE Intelligent Vehicles Symposium (IV), pp. 1094–1101, 2020.

What to Fuse? - LiDAR

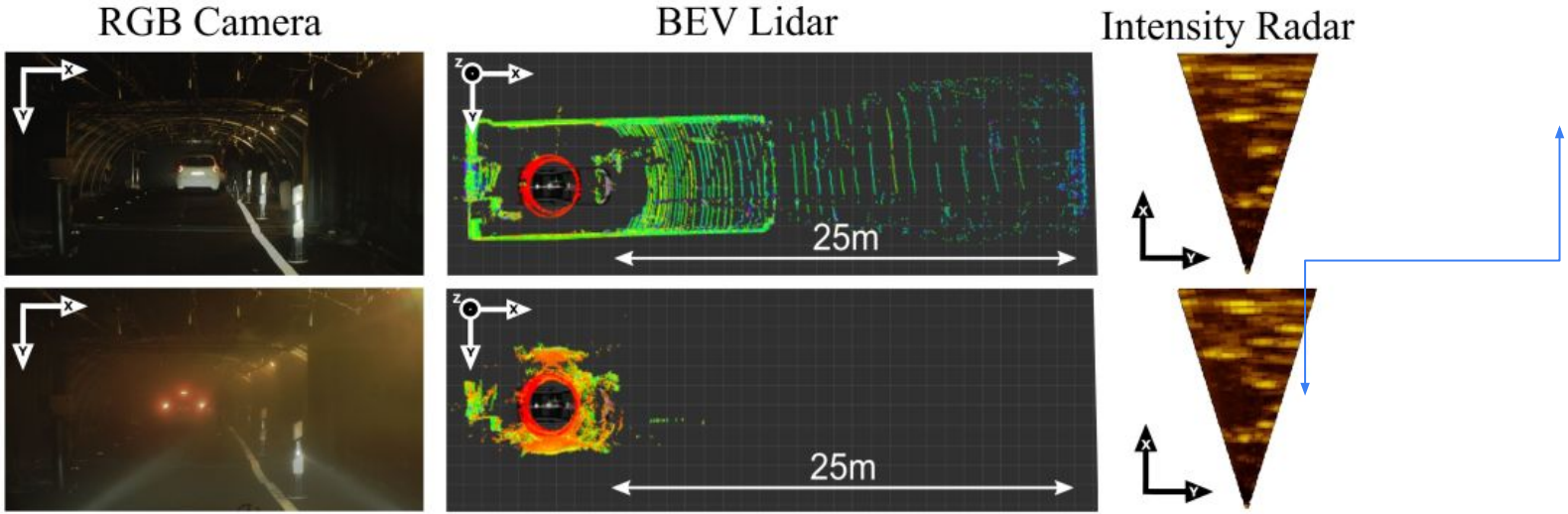


Figure 5: 1st row: clear weather condition, 2nd row: with fog.
Shows that LiDAR affects by the fog but Radar intensity remains the same (Image source [1])

[1] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, “Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather,” pp. 11 682–11 692, 2020.

What to Fuse? - Radar

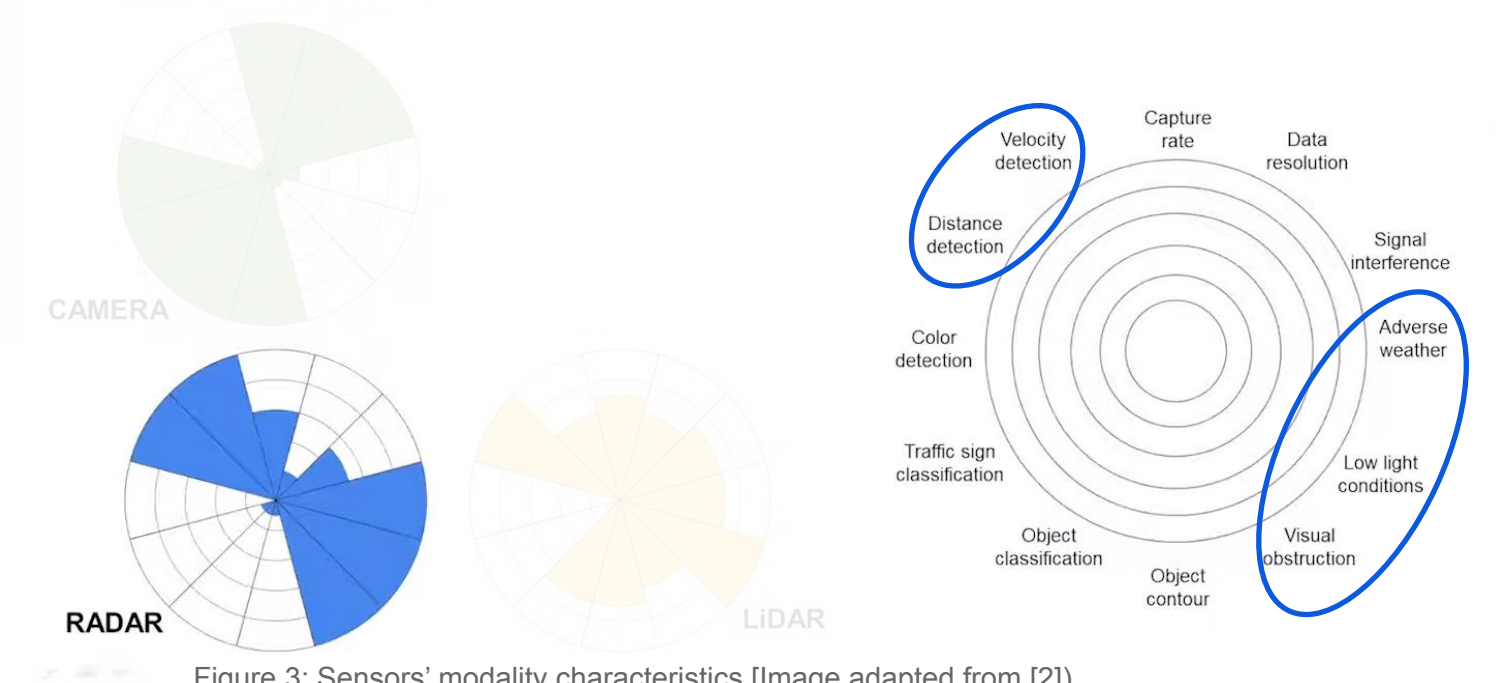
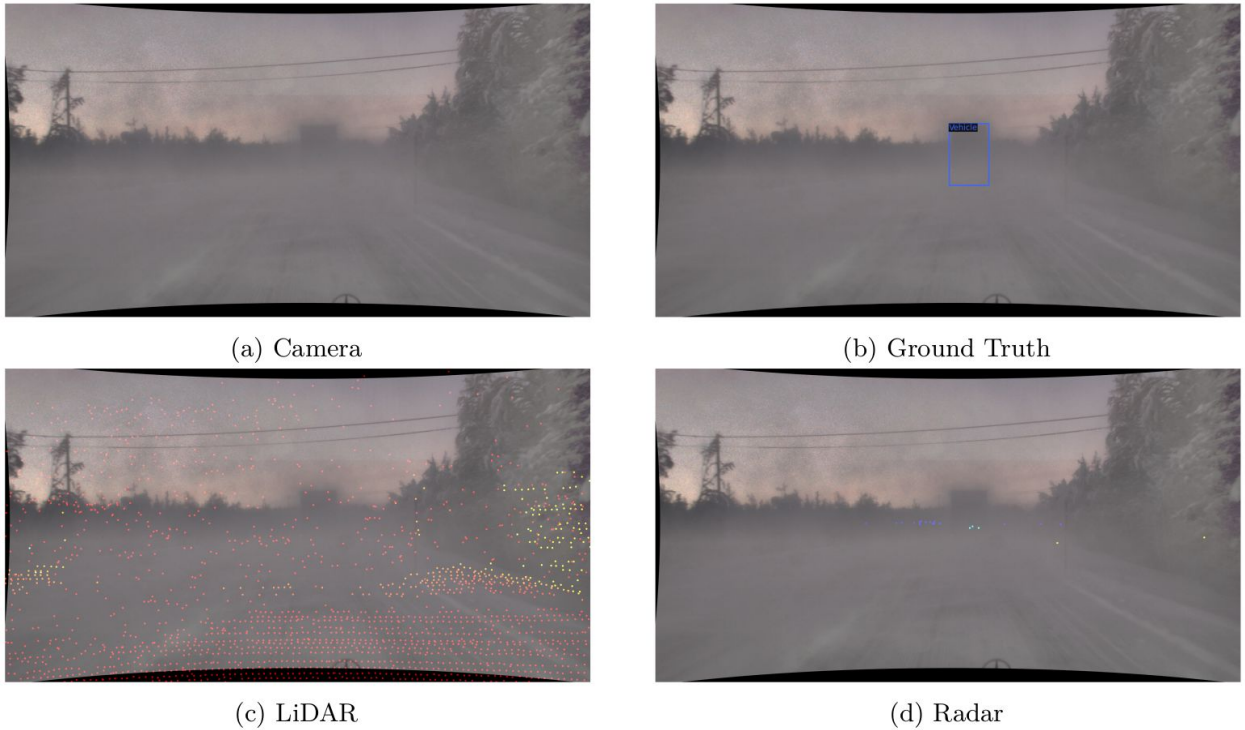


Figure 3: Sensors' modality characteristics [Image adapted from [2])

[2] R. Laganière, "Sensor fusion for autonomous vehicles: Strategies, methods, and tradeoffs," 2022, accessed on 18.12.2022. [Online]. Available: https://youtu.be/2Fcmh7SLPBI

What to Fuse? - Radar



(a) Camera (b) Ground Truth

(c) LiDAR (d) Radar

Figure 6: Dense fog influence on sensors (a sample from DENSE dataset [1])

[1] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, “Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather,” pp. 11 682–11 692, 2020.

What to Fuse? - Radar

(c) LiDAR

(d) Radar

Figure 6: Dense fog influence on sensors (a sample from DENSE dataset [1])

[1] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, “Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather,” pp. 11 682–11 692, 2020.

What to Fuse?- Radar

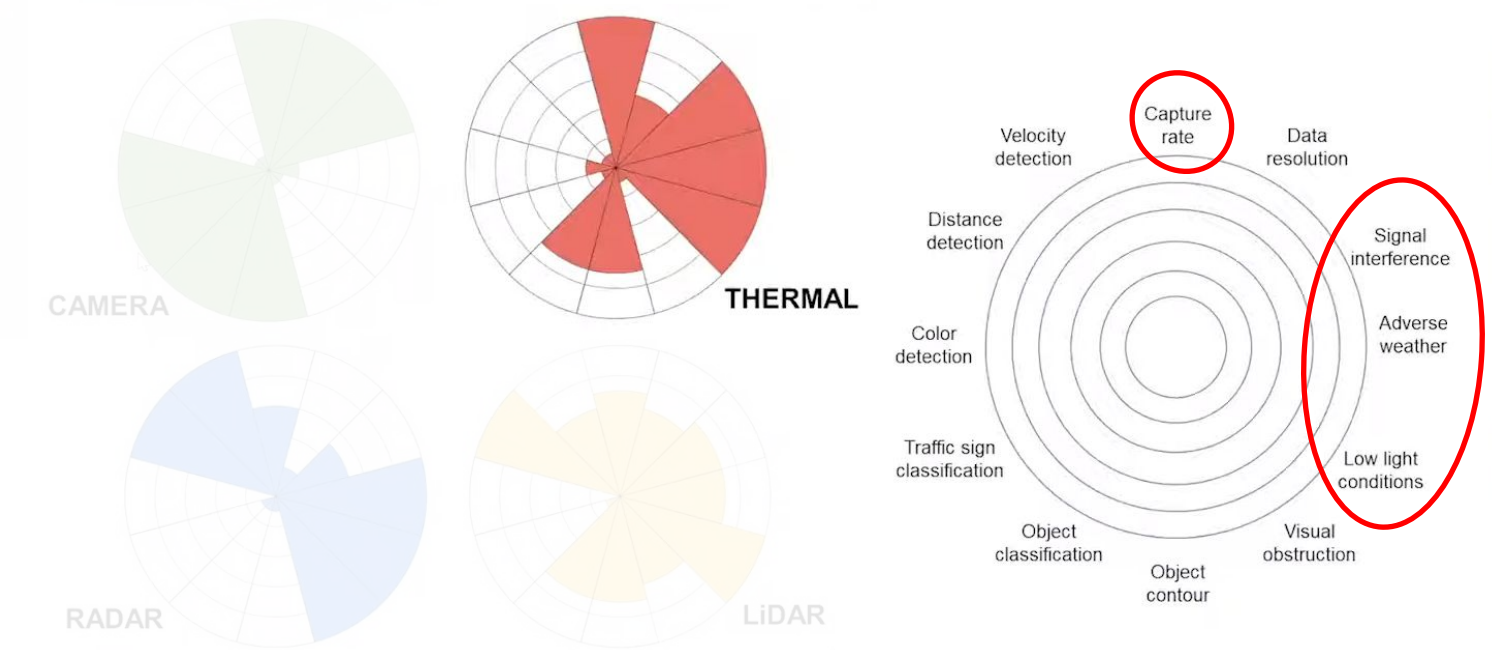


Figure 3: Sensors' modality characteristics [Image adapted from [2])

[2] R. Laganière, "Sensor fusion for autonomous vehicles: Strategies, methods, and tradeoffs," 2022, accessed on 18.12.2022. [Online]. Available: https://youtu.be/2Fcmh7SLPBI

What to Fuse? - All?

Answer: Sensor Fusion !

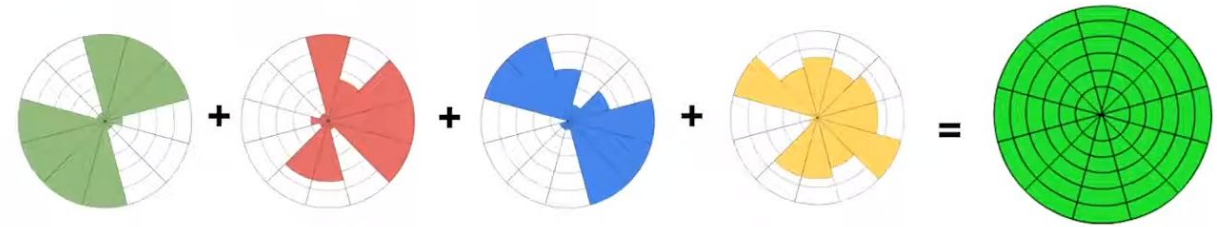


Figure 3: Sensors' modality characteristics [Image adapted from [2])

- This is not system redundancy
- Complementary sensors (multimodal)
- It's union not intersection

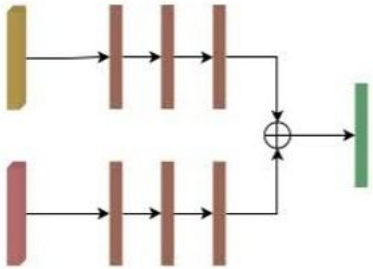
[2] R. Laganière, "Sensor fusion for autonomous vehicles: Strategies, methods, and tradeoffs," 2022, accessed on 18.12.2022. [Online]. Available: https://youtu.be/2Fcmh7SLPBI

What to Fuse? - Example

Figure 7: Importance of sensor fusion in adverse weather conditions, *Fusion of Camera, LiDAR, Radar (Image adapted from [1])

[1] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, “Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather,” pp. 11 682–11 692, 2020.

When to Fuse?



a) Late Fusion

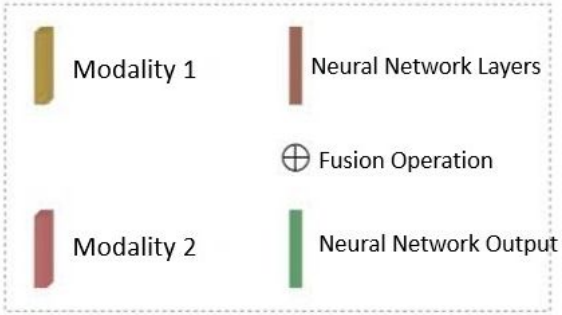


Figure 8: Taxonomy of fusion levels. (a) Late Fusion. (b) Early Fusion. (c) Middle Fusion. (d) Tightly-coupled Fusion (Image adapted from [4])

[4] Yao, Shanliang, et al. "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review." arXiv preprint arXiv:2304.10410 (2023).

When to Fuse?

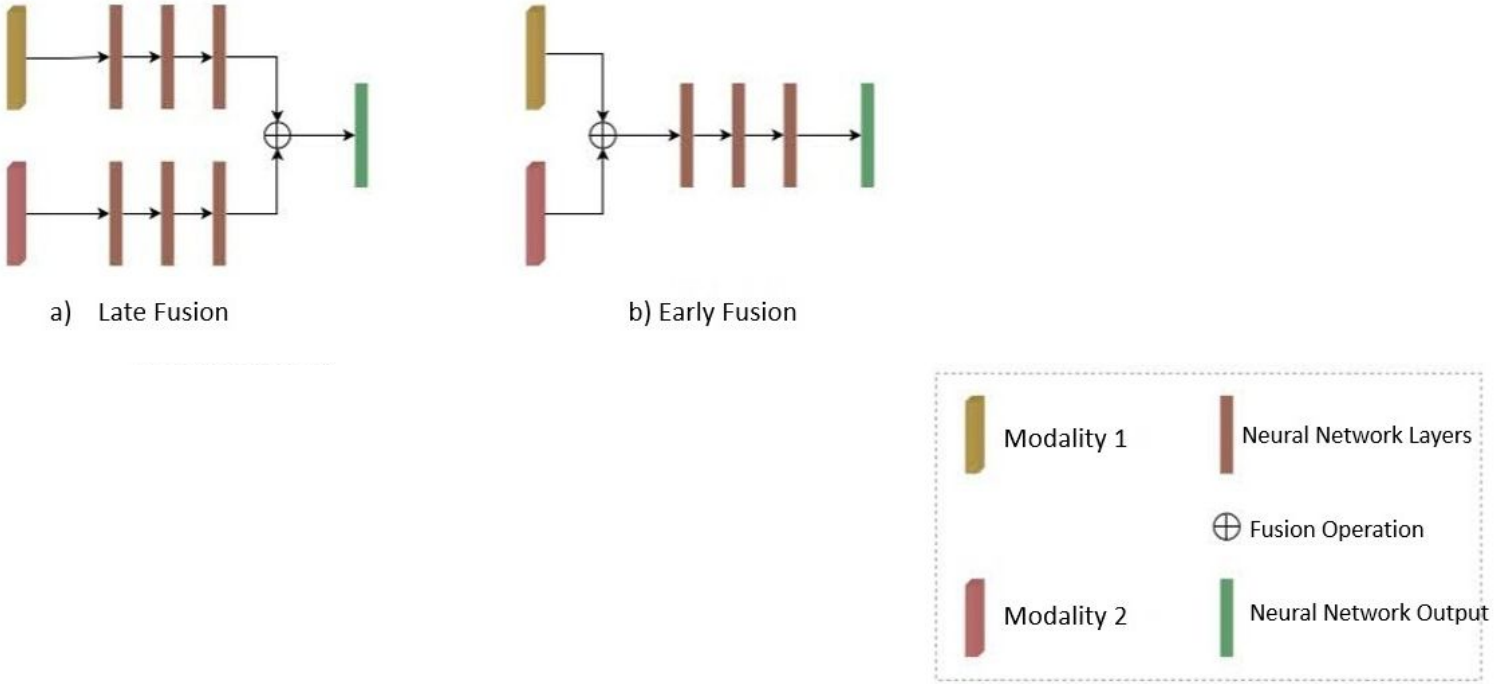


Figure 8: Taxonomy of fusion levels. (a) Late Fusion. (b) Early Fusion. (c) Middle Fusion. (d) Tightly-coupled Fusion (Image adapted from [4])

[4] Yao, Shanliang, et al. "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review." arXiv preprint arXiv:2304.10410 (2023).

When to Fuse?

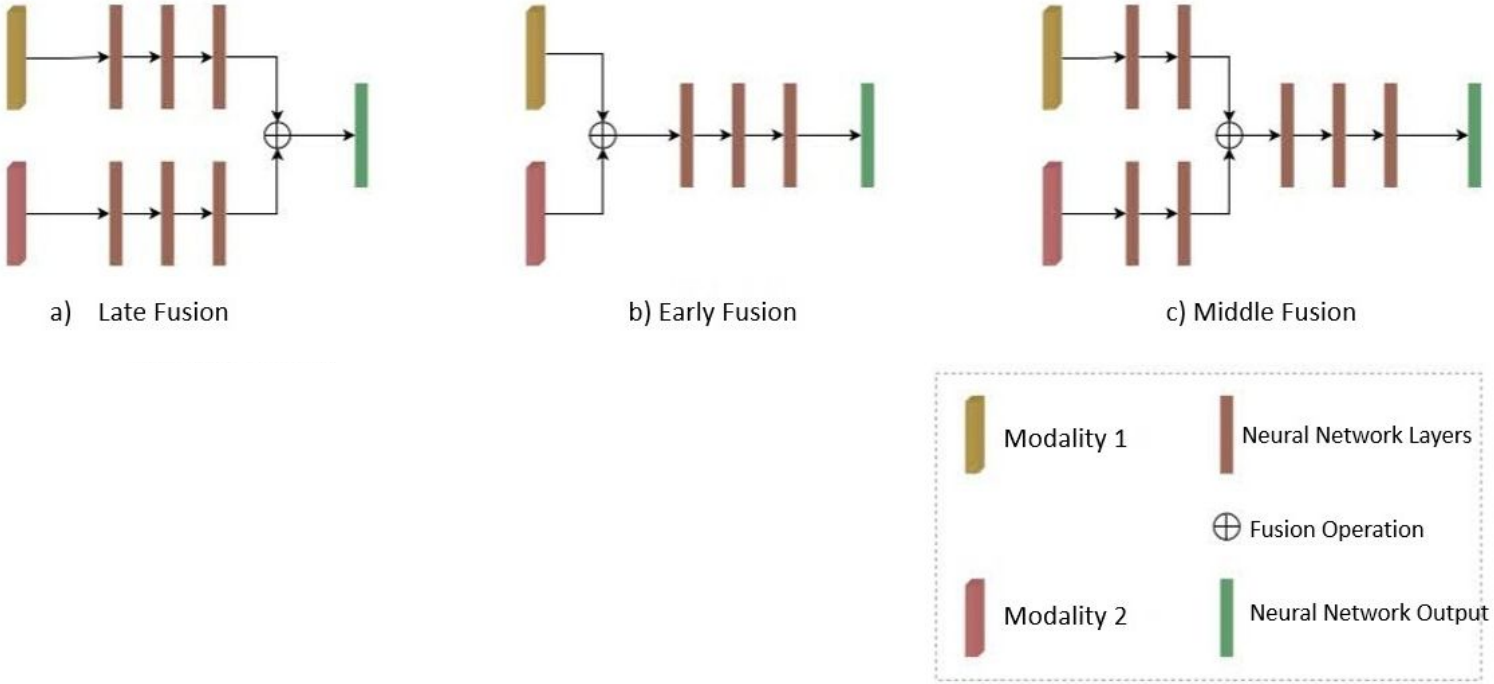


Figure 8: Taxonomy of fusion levels. (a) Late Fusion. (b) Early Fusion. (c) Middle Fusion. (d) Tightly-coupled Fusion (Image adapted from [4])

[4] Yao, Shanliang, et al. "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review." arXiv preprint arXiv:2304.10410 (2023).

When to Fuse?

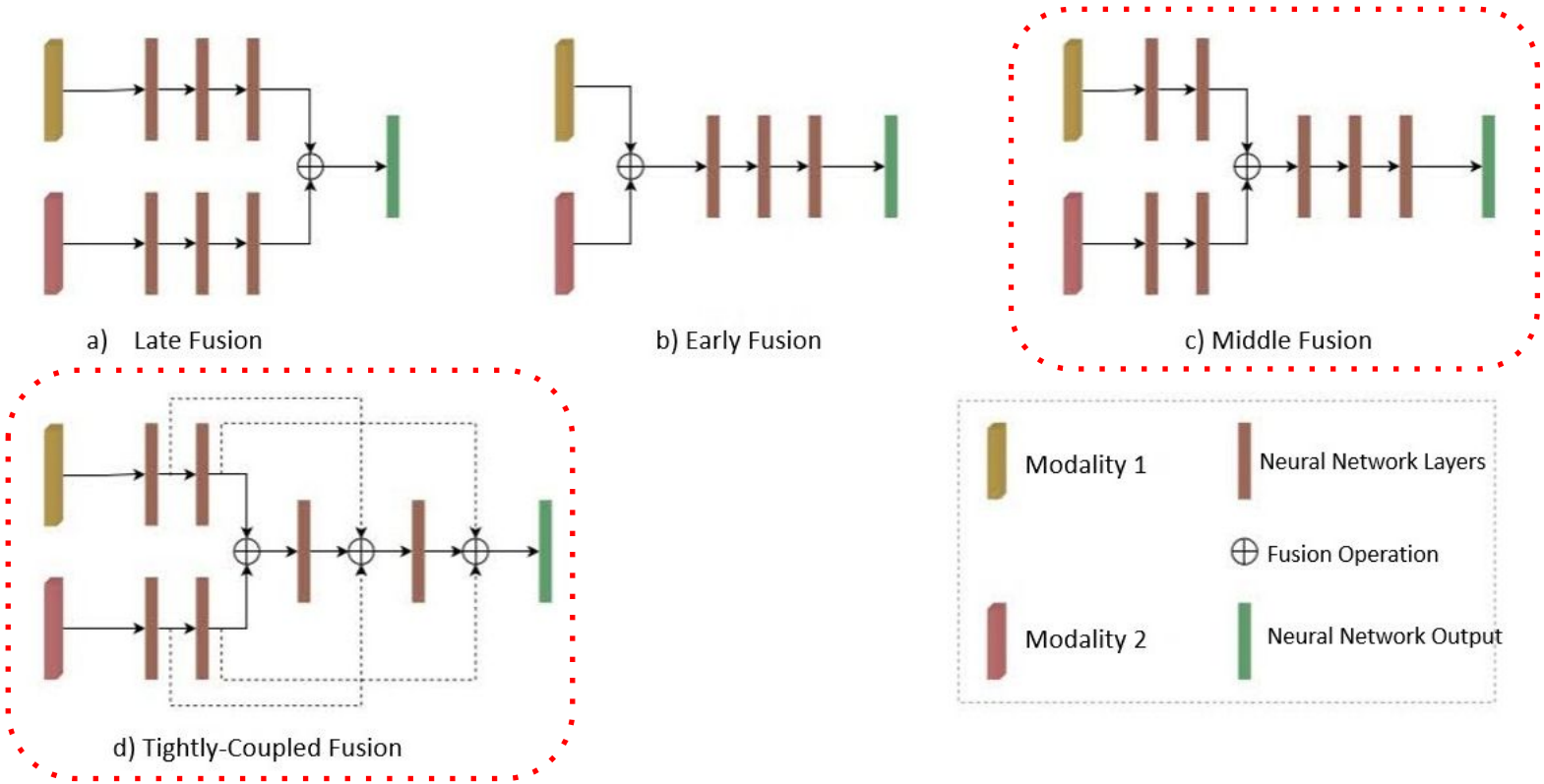


Figure 8: Taxonomy of fusion levels. (a) Late Fusion. (b) Early Fusion. (c) Middle Fusion. (d) Tightly-coupled Fusion (Image adapted from [4])

[4] Yao, Shanliang, et al. "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review." arXiv preprint arXiv:2304.10410 (2023).

Challenges

- Scarcity of adverse weather-related datasets
  - ✅ Camera, Lidar
  - ✅ Clear weather
  - ❌ Radar
  - ❌ Adverse weather conditions

KITTI[5], Waymo[6], ApolloScape[7]

[5] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2012, pp. 3354–3361.
[6] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine et al., “Scalability in perception for autonomous driving: Waymo open dataset,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2446–2454.
[7] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, “The apolloscape open dataset for autonomous driving and its application,” IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 10, pp. 2702–2719, 2019.

Challenges

- Scarcity of adverse weather-related datasets
  - Camera, Lidar, and clear weather but no radar, adverse weather
- Limitations in fusion architecture
  - ✅ Late fusion or middle/feature fusion (mostly camera and lidar)
  - ❌ Tightly-coupled fusion

Challenges

- Scarcity of adverse weather-related datasets
  - Camera, Lidar, and clear weather but no radar, adverse weather
- Limitations in fusion architecture
  - Late fusion or middle/feature fusion but no tightly-coupled fusion
- General guidelines for network architecture
  - What to fuse – camera, lidar, radar, thermal camera, ultrasonic

Figure 9: Automotive sensors (Image source [5])

[5] McLellan, Paul. "Automotive Sensors: Cameras, Lidar, Radar, Thermal - Breakfast Bytes - Cadence Blogs." Cadence Community, 14 Dec. 2018, https://community.cadence.com/cadence_blogs_8/b/breakfast-bytes/posts/autos182.

Challenges

- Scarcity of adverse weather-related datasets
  - Camera, Lidar, and clear weather but no radar, adverse weather
- Limitations in fusion architecture
  - Late fusion or middle/feature fusion but no tightly-coupled fusion
- General guidelines for network architecture
  - What to fuse – camera, lidar, radar, thermal camera, ultrasonic
  - How to fuse – addition, averaging, multiplication, concatenation

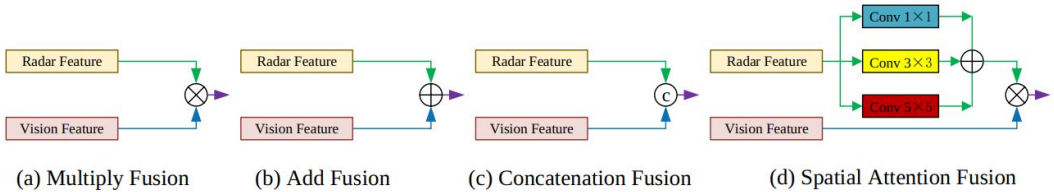


Figure 10: Various feature fusion blocks (Image source [9])

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, "Spatial attention fusion for obstacle detection using mmwave radar and vision sensor," Sensors, vol. 20, no. 4, p. 956, 2020.

Challenges

- Scarcity of adverse weather-related datasets
  - Camera, Lidar, and clear weather but no radar, adverse weather
- Limitations in fusion architecture
  - Late fusion or middle/feature fusion but no tightly-coupled fusion
- General guidelines for network architecture
  - What to fuse – camera, lidar, radar, thermal camera, ultrasonic
  - How to fuse – addition, averaging, multiplication, concatenation
  - When to fuse – early, mid, late, tightly-coupled

Challenges

- Scarcity of adverse weather-related datasets
  - Camera, Lidar, and clear weather but no radar, adverse weather
- Limitations in fusion architecture
  - Late fusion or middle/feature fusion but no tightly-coupled fusion
- General guidelines for network architecture
  - What to fuse, How to fuse, When to fuse
- Limited generalization from previous studies
  - Benchmarking on custom datasets
  - Custom classes, metrics

Challenges

- Scarcity of adverse weather-related datasets
  - Camera, Lidar, and clear weather but no radar, adverse weather
- Limitations in fusion architecture
  - Late fusion or middle/feature fusion but no tightly-coupled fusion
- General guidelines for network architecture
  - What to fuse, How to fuse, When to fuse
- Limited generalization from previous studies
  - Benchmarking on custom datasets, classes, metrics
- Computational constraints

Contributions

- Detailed study on top 3 multimodal sensor fusion methods for 2D detection
- Method comparison using a common metric, COCO
- Analyzed performance with 2 datasets, highlighting adverse weather robustness
- Standardized evaluation criteria to eliminate biases
- In-depth qualitative and quantitative analysis

Related Works

Name	Sensors*	Dataset Used	Fusion Method	Year
Nobis et al. [8]	CR	nuScenes	Early	2019
SAF-FCOS [9]		nuScenes	Middle	2020
BIRANet [10]		nuScenes	Middle	2020
Liu et al. [11]		Custom	Early	2021
RODNet [12]		CRUW	Middle	2021
Danapal et al. [13]		nuScenes	Tightly-coupled	2022
Radecki et al. [14]	CRL	KITTI	Early	2016
Bijelic et al. [1]		DENSE	Middle	2020
Rawashdeh et al. [15]		DENSE	Middle	2021
MT-DETR [16]		DENSE	Tightly-coupled	2023
HRFuser [17]		nuScenes, DENSE	Tightly-coupled	2023
FLIR System Inc. [18]	CRT	Proprietary	NA	2018
RadarNet [19]	RL	nuScenes	Early	2020
MVDNet [20]		DENSE	Middle	2021

Table 1: Multimodal sensor fusion methods for adverse weather conditions, Sensors* - C - Camera, R – Radar, L – LiDAR, T- Thermal (sorted based on sensors group)

Related Works

Name	Sensors*	Dataset Used	Fusion Method	Year
Nobis et al. [5]		nuScenes	Early	2019
SAF-FCOS [9]	**CR**	**nuScenes**	**Middle**	**2020**
BIRANet [7]		nuScenes	Middle	2020
Liu et al. [8]		Custom	Early	2021
RODNet [9]		CRUW	Middle	2021
Danapal et al. [10]		nuScenes	Tightly-coupled	2022
Radecki et al. [11]		KITTI	Early	2016
Bijelic et al. [12]		DENSE	Middle	2020
Rawashdeh et al. [13]		DENSE	Middle	2021
MT-DETR [16]	**CRL**	**DENSE**	**Tightly-coupled**	**2023**
HRFuser [17]		**nuScenes, DENSE**	**Tightly-coupled**	**2023**
FLIR System Inc. [16]	CRT	Proprietary	NA	2018
RadarNet [17]	RL	nuScenes	Early	2020
MVDNet [18]		DENSE	Middle	2021

Table 1: Multimodal sensor fusion methods for adverse weather conditions, Sensors* - C - Camera, R – Radar, L – LiDAR, T- Thermal (sorted based on sensors group)

Related Works

- Synthetic data for adverse weather conditions
  - 🟢 Synthetic data and deep learning-based de-hazing mitigate weather impacts [60, 61]
  - 🟢 GANs simulate real-world weather for model training [66, 67, 68]
  - 🔴 High computational complexity in deep de-hazing models [62]
  - 🔴 Poor generalization of models to real-world haze [62]
- Role of simulation
  - 🟢 CARLA simulator enables complex road environment creation [71]
  - 🟢 Simulations provide safe testing conditions for adverse weather [9]
  - 🔴 Simulators may not accurately mirror real-world conditions [26]
  - 🔴 Challenges in integrating real and virtual data effectively [26]

Methodology - Datasets

Name	Sensors*	Weather Cond.**	Size (GB)	Year	Citation
DENSE [1]	CRL	F, SN, R, O, N	582	2020	269
nuScenes [21]		R, N	400	2020	3459
Oxford RobotCar [22]		R, SN, F	4700	2020	317
EU Long-term [23]		SN, R, O, N	NA	2020	72
RADIATE [24]		F, SN, R, O, SL, N	NA	2021	132
K-Radar [25]		F, R, SN	13000	2022	15
Boreas [26]		SN, R, O, N	4400	2022	38
aiMotive [27]		R, O, N	85	2023	3

Table 2: Multimodal sensor fusion methods for adverse weather conditions,
Sensors*: C - Camera, R – Radar, L – LiDAR;
Weather Cond.**: F - Fog, SN - Snow, R - Rain, O - Overcast, SL - Sleet, N – Night
(sorted based on year)

Methodology - Datasets

Name	Sensors*	Weather Cond.**	Size (GB)	Year	Citation
DENSE [1]	CRL	F, SN, R, O, N	582	2020	269
nuScenes [21]		R, N	400	2020	3459
Oxford RobotCar [57]		R, SN, F	4700	2020	317
EU Long-term [72]		SN, R, O, N	NA	2020	72
RADIATE [73]		F, SN, R, O, SL, N	NA	2021	132
K-Radar [34]		F, R, SN	13000	2022	15
Boreas [74]		SN, R, O, N	4400	2022	38
aiMotive [75]		R, O, N	85	2023	3

Table 2: Multimodal sensor fusion methods for adverse weather conditions,
Sensors*: C - Camera, R – Radar, L – LiDAR;
Weather Cond.**: F - Fog, SN - Snow, R - Rain, O - Overcast, SL - Sleet, N – Night
(sorted based on year)

[1] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, “Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather,” pp. 11 682–11 692, 2020.
[21] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

Dataset - DENSE

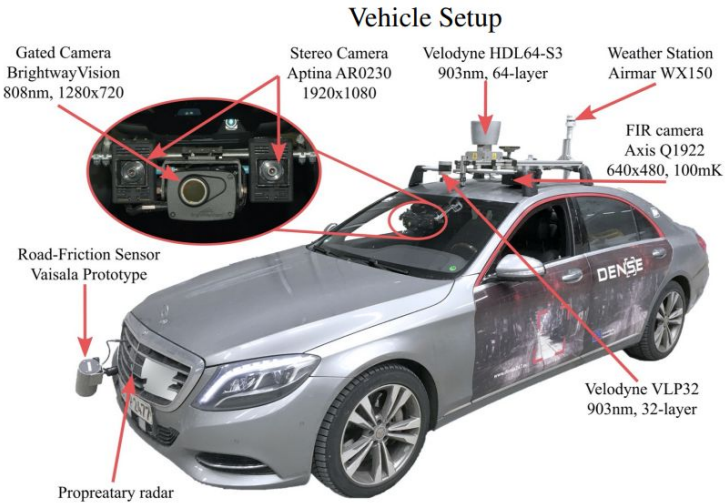


Figure 10: Test vehicle setup (Image source [1])

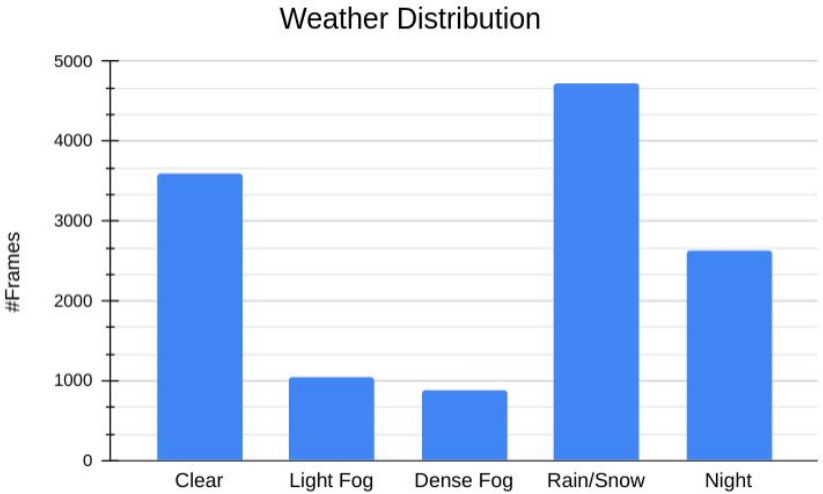


Figure 11: Distribution of weather conditions

[1] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, “Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather,” pp. 11 682–11 692, 2020.

Dataset - DENSE

- Classes (3):
  - car, pedestrian, cyclist
- Radar data:
  - In point cloud format
  - It includes 3D information – range, azimuth, and velocity
- Offers 2D annotations
- COCO style format

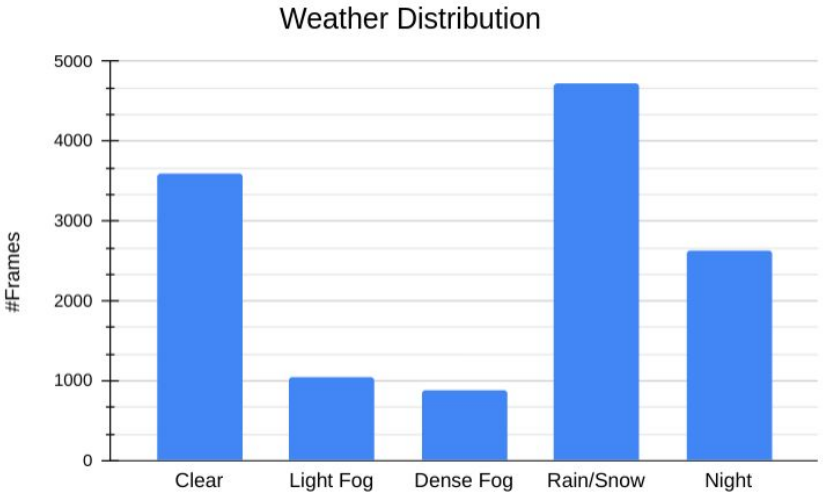


Figure 11: Distribution of weather conditions

Dataset - DENSE

Figure 12: Random samples from the DENSE dataset, where 1st column is Camera with ground truth, 2nd is LiDAR, 3rd is Radar (1st Row: Day, 2nd Row: Light Fog, 3rd Row: Dense Fog, 4th Row: Snow, 5th Row: Night).
Note: Radar sparse points are highlighted with a red ellipse.

Dataset - DENSE

Dataset - DENSE

Dataset - nuScenes

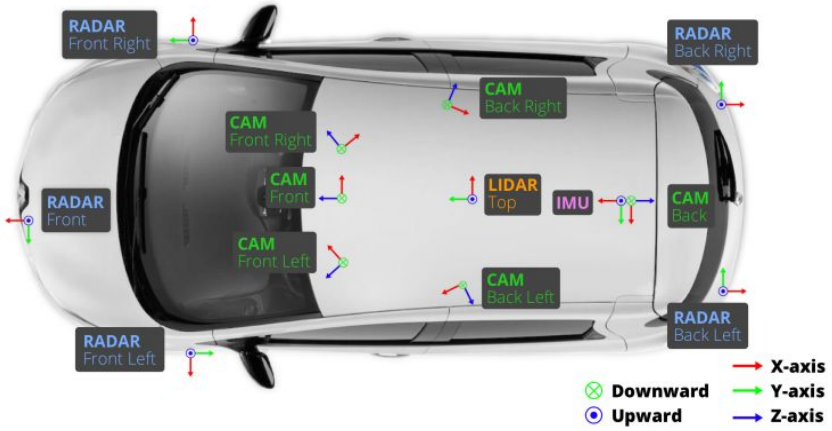


Figure 13: Test vehicle setup (Image source [21])

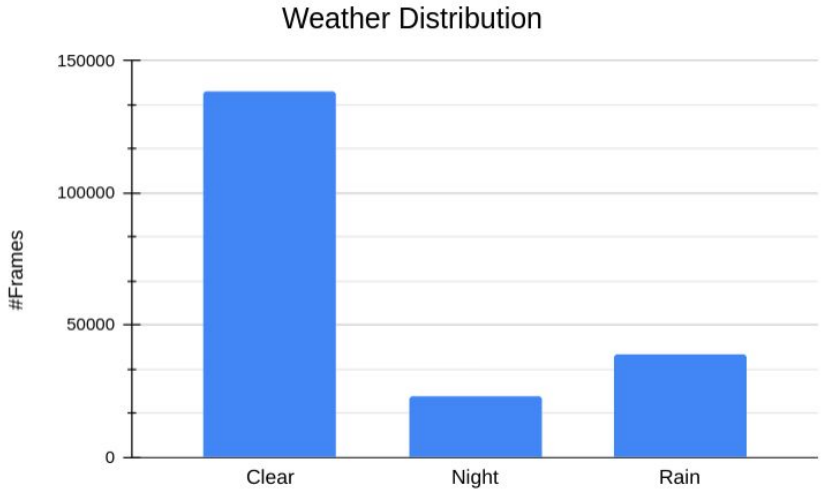


Figure 14: Distribution of weather conditions

[21] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

Dataset - nuScenes

- Classes (10):
  - car, truck, trailer, bus, construction vehicle, bicycle, motorcycle, pedestrian, traffic cone, and barrier
- Radar data:
  - In point cloud format
  - It includes 3D information – range, azimuth, and velocity
- Offers 3D annotations
- COCO style format

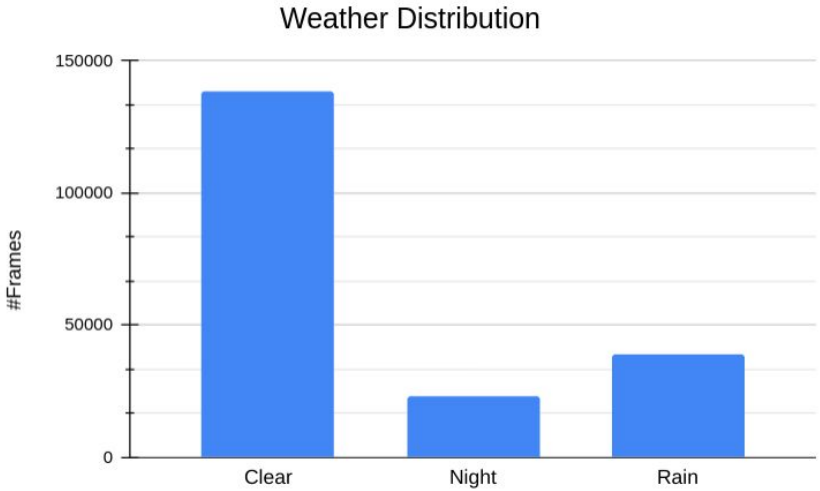
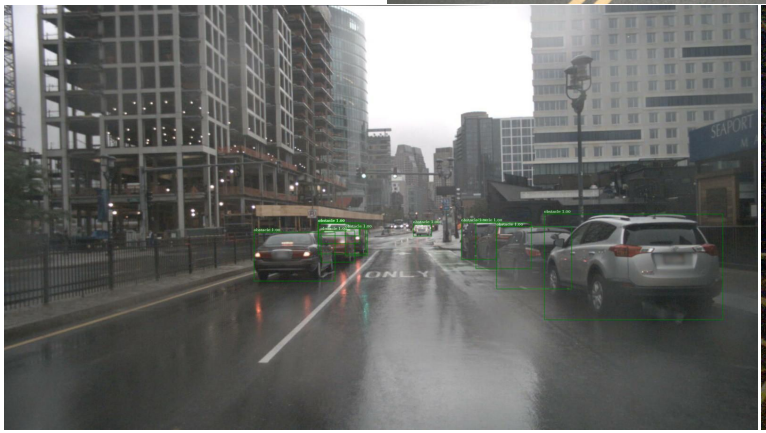


Figure 14: Distribution of weather conditions

Dataset - nuScenes



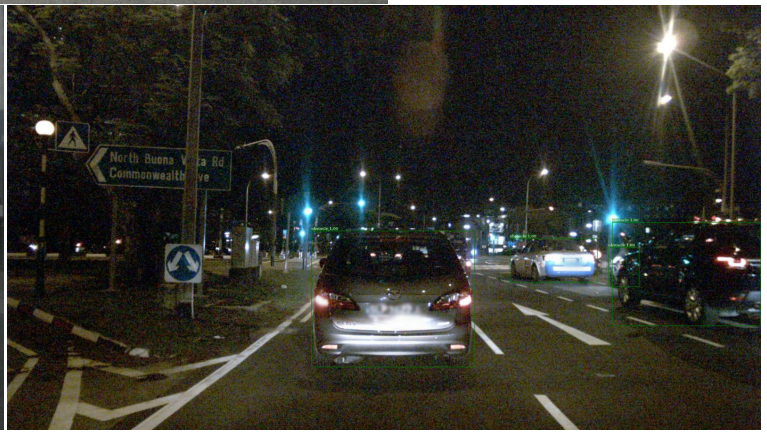


Figure 15: Random samples from the nuScenes dataset

Dataset – nuScenes vs DENSE

Dataset	nuScenes [21]	DENSE [1]
RGB Cameras	6	2
RGB Resolution	1600x900	1920x1024
LiDAR Sensors	1	2
LiDAR Resolution	32	64
Radar Sensor	5	1
Gated Camera	x	1
FIR Camera	x	1
Frame Rate	10 Hz	10 Hz
Dataset Statistics		
Labeled Frames	40K	13.5K
Labels	1.4M	100K
Scene Tags	✓	✓
Night Time	✓	✓
Light Weather	✓	✓
Heavy Weather	x	✓
Fog Chamber	x	✓

Table 3: Comparison of datasets features (Table adapted from [1])

[1] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, “Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather,” pp. 11 682–11 692, 2020.
[21] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

Evaluation Metrics

- Precision
- Recall
- Average Precision (AP)
- Average Recall (AR)
- Inference time (ms) or FPS
- Floating Point Operations (FLOPs)
- Model parameters

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

$$AP@[0.5 : 0.05 : 0.95] = \frac{AP_{0.5} + AP_{0.55} + \ldots + AP_{0.95}}{10}$$

Selected Methods

Name	Sensors*	Dataset Used	Fusion Method	Year
Nobis et al. [5]		nuScenes	Early	2019
SAF-FCOS [9]	**CR**	**nuScenes**	**Middle**	**2020**
BIRANet [7]		nuScenes	Middle	2020
Liu et al. [8]		Custom	Early	2021
RODNet [9]		CRUW	Middle	2021
Danapal et al. [10]		nuScenes	Tightly-coupled	2022
Radecki et al. [11]		KITTI	Early	2016
Bijelic et al. [12]		DENSE	Middle	2020
Rawashdeh et al. [13]		DENSE	Middle	2021
MT-DETR [16]	**CRL**	**DENSE**	**Tightly-coupled**	**2023**
HRFuser [17]		**nuScenes, DENSE**	**Tightly-coupled**	**2023**
FLIR System Inc. [16]	CRT	Proprietary	NA	2018
RadarNet [17]	RL	nuScenes	Early	2020
MVDNet [18]		DENSE	Middle	2021

Table 1: Multimodal sensor fusion methods for adverse weather conditions, Sensors* - C - Camera, R – Radar, L – LiDAR, T- Thermal (sorted based on sensors group)

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, “Spatial attention fusion for obstacle detection using mmwave radar and vision sensor,” Sensors, vol. 20, no. 4, p. 956, 2020.
[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Method 1: SAF-FCOS

Name	Sensors	Dataset Used	Fusion Method
SAF-FCOS	CR	nuScenes	Middle

SAF-FCOS: Spatial Attention Fusion with Fully COnvolutional Single-stage object detection

- Preprocessing: Radar imagery

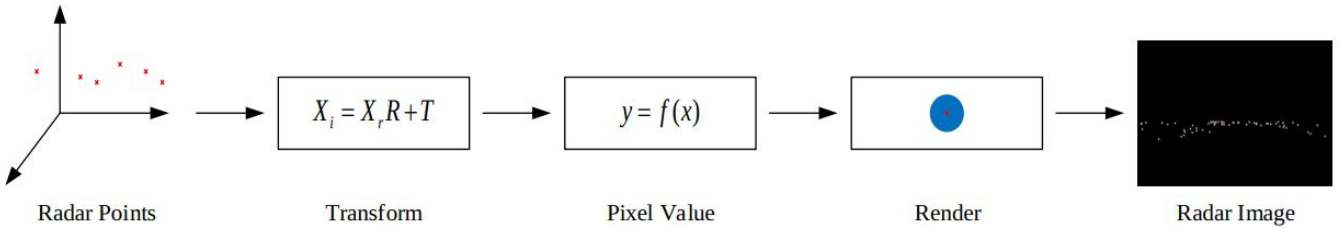


Figure 16: Radar image generation model (Image source [9])

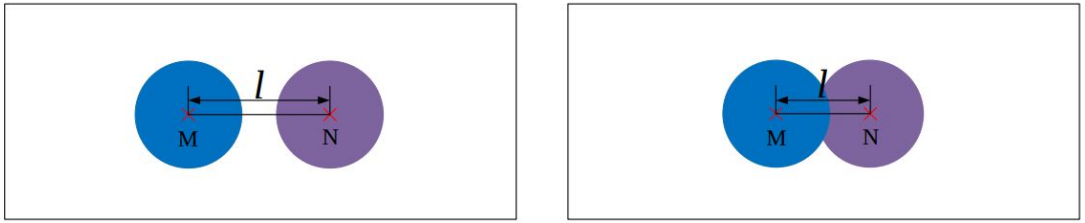
$$R = \frac{128d}{250} + 127, \quad G = \frac{128(v_x + 20)}{40} + 127, \quad B = \frac{128(v_y + 20)}{40} + 127$$

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, "Spatial attention fusion for obstacle detection using mmwave radar and vision sensor," Sensors, vol. 20, no. 4, p. 956, 2020.

Method 1: SAF-FCOS

Name	Sensors	Dataset Used	Fusion Method
SAF-FCOS	CR	nuScenes	Middle

- Preprocessing: Radar imagery



Rendering Case A Rendering Case B

Figure 16: Radar image rendering condition (Image source [9])

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, “Spatial attention fusion for obstacle detection using mmwave radar and vision sensor,” Sensors, vol. 20, no. 4, p. 956, 2020.

Method 1: SAF-FCOS

Name	Sensors	Dataset Used	Fusion Method
SAF-FCOS	CR	nuScenes	Middle

- Model architecture

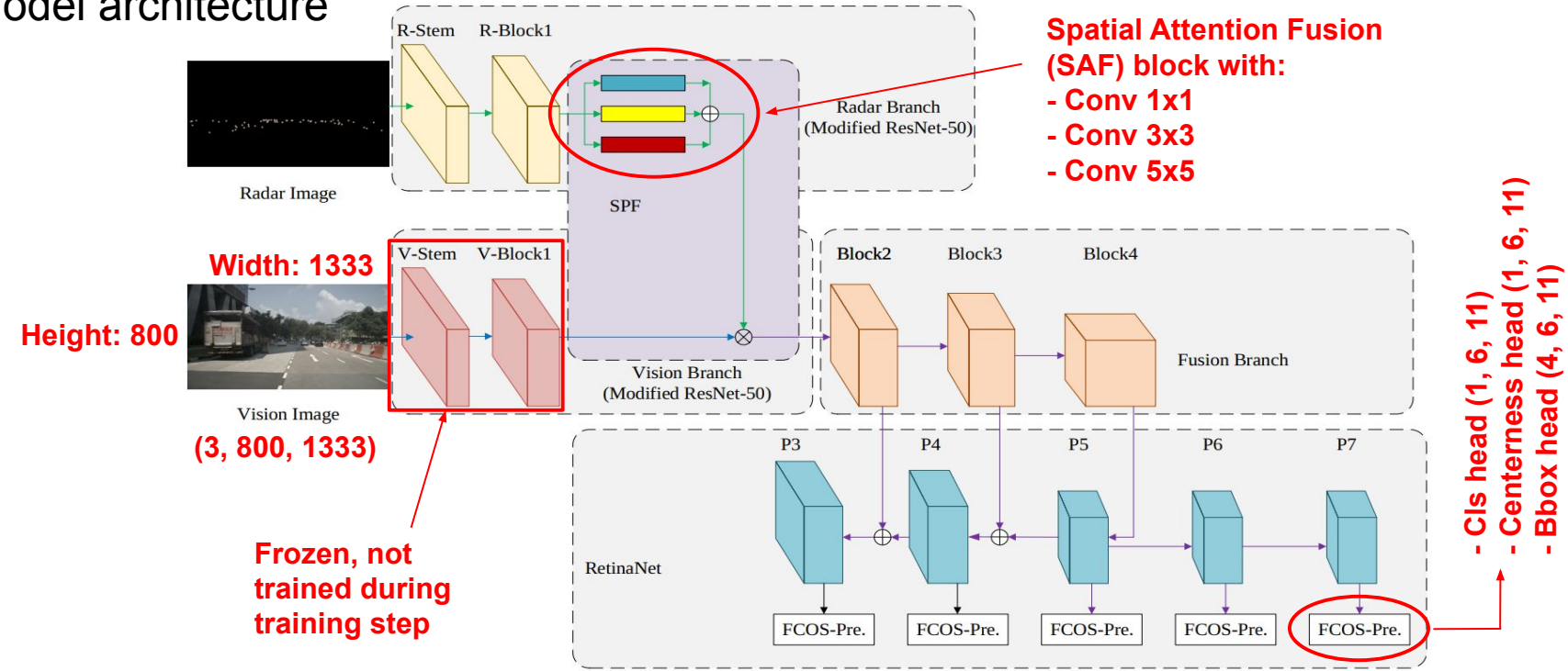


Figure 17: SAF-FCOS model architecture (Image source [9])

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, "Spatial attention fusion for obstacle detection using mmwave radar and vision sensor," Sensors, vol. 20, no. 4, p. 956, 2020.

Method 1: SAF-FCOS

Name	Sensors	Dataset Used	Fusion Method
SAF-FCOS	CR	nuScenes	Middle

- Model architecture

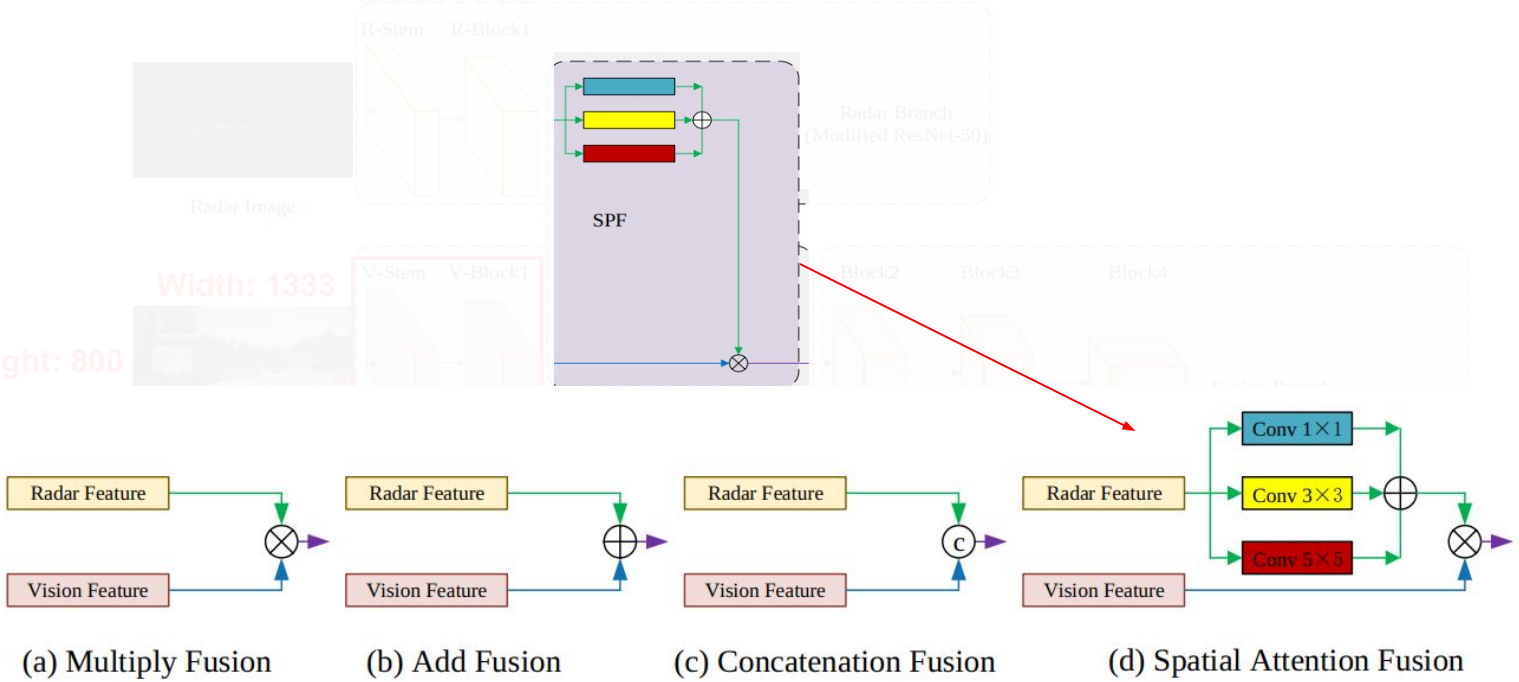


Figure 18: Various feature fusion blocks (Image source [9])

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, "Spatial attention fusion for obstacle detection using mmwave radar and vision sensor," Sensors, vol. 20, no. 4, p. 956, 2020.

Method 1: SAF-FCOS

Name	Sensors	Dataset Used	Fusion Method
SAF-FCOS	CR	nuScenes	Middle

- Loss function [9]
  - Classification loss: Focal loss [29]
  - BBox regression loss: IOU loss [30]
  - Balancing weight: λ
  - Positive sample indicator

$$L(c_i, t_i) = \frac{1}{N_{\text{pos}}} \left(\sum_i L_{\text{cls}}(c_i, c_i^*) + \lambda \sum_i \mathbb{1}_{c_i^* > 0} L_{\text{reg}}(t_i, t_i^*) \right)$$

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, “Spatial attention fusion for obstacle detection using mmwave radar and vision sensor,” Sensors, vol. 20, no. 4, p. 956, 2020
[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Doll´ar, “Focal loss for dense object detection,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
[30] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, “Unitbox: An advanced object detection network,” in Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 516–520.

Method 2: HRFuser

Name	Sensors	Dataset Used	Fusion Method
HRFuser	CRL	nuScenes, DENSE	Tightly-coupled

HRFuser: High Resolution Fuser

- Preprocessing: LiDAR & Radar imagery

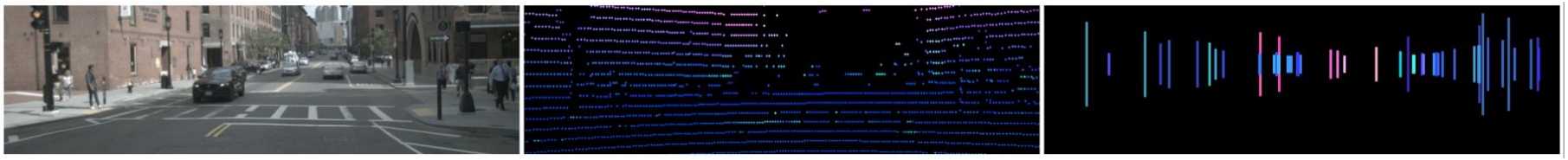


Figure 19: An example projection image from nuScenes. Viewing sequence: RGB image, LiDAR and Radar point projections (Image adapted from [17])

Table 4: Overview of sensor projection parameters. RCS stands for Radar Cross Section

	DENSE	nuScenes
Sensor	**Projection Parameters**	
Radar Imagery	Distance, Velocity	Distance, Velocity, RCS
Lidar Imagery	Distance, Intensity, Height	Distance, Intensity, Height

[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Method 2: HRFuser

Name	Sensors	Dataset Used	Fusion Method
HRFuser	CRL	nuScenes, DENSE	Tightly-coupled

HRFuser: High Resolution Fuser

- Preprocessing: Radar processing

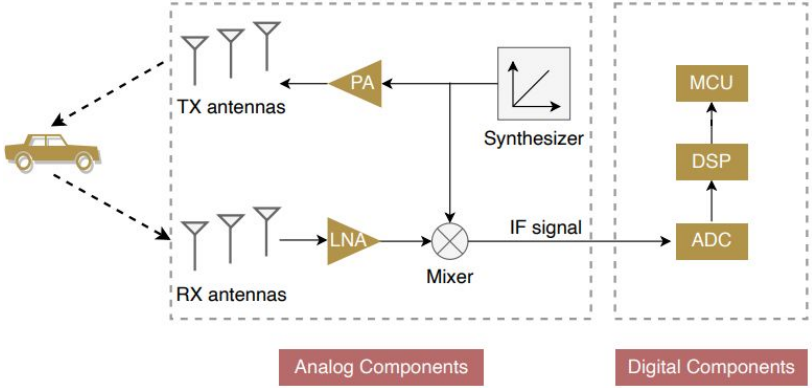


Figure x: Overview of radar working pipeline (Image source [4])

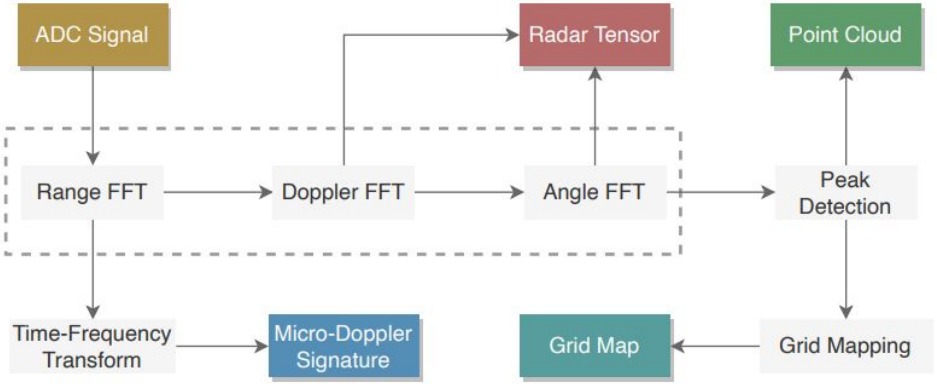


Figure x: Generation progress of five radar data representations (i.e., ADC signal, radar tensor, point cloud, grid map, and micro-Doppler signature) (Image source [4])

[4] Yao, Shanliang, et al. "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review." arXiv preprint arXiv:2304.10410 (2023).

Method 2: HRFuser

Name	Sensors	Dataset Used	Fusion Method
HRFuser	CRL	nuScenes, DENSE	Tightly-coupled

HRFuser: High Resolution Fuser

- Preprocessing: Radar data representation

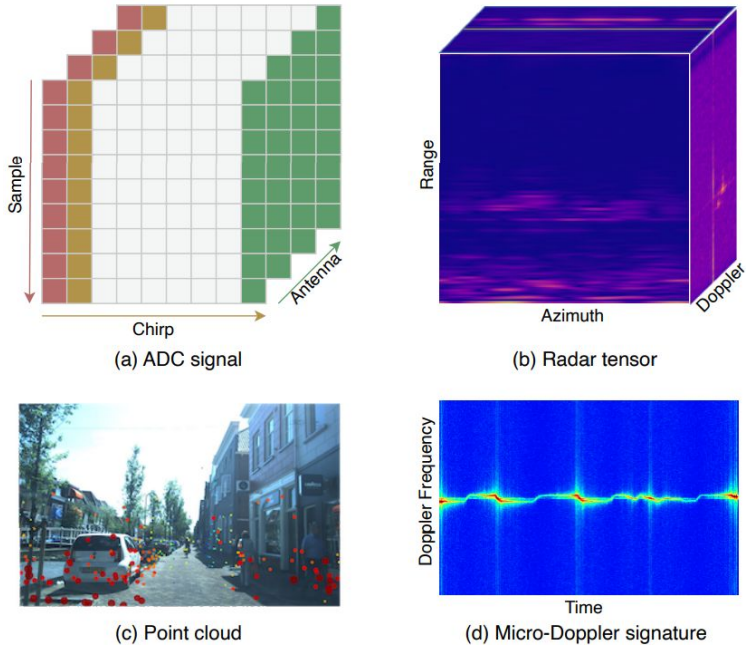


Figure x: Radar data representations. (a) ADC signal in the format of SimpleChirp-Antenna tensor. (b) Radar tensor represented by a 3D Range-AzimuthDoppler tensor. Image is generated from the CARRADA [33] dataset. (c) Point cloud projected on a 2D image plane. Image is generated from the View-of-Delft [34] dataset. (d) Micro-Doppler signature showing a pedestrian walking. Image is generated from the Open Radar Datasets [35] [4]

[4] Yao, Shanliang, et al. "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review." arXiv preprint arXiv:2304.10410 (2023).

Method 2: HRFuser

Name	Sensors	Dataset Used	Fusion Method
HRFuser	CRL	nuScenes, DENSE	Tightly-coupled

HRFuser: High Resolution Fuser

- Model architecture (1/3)

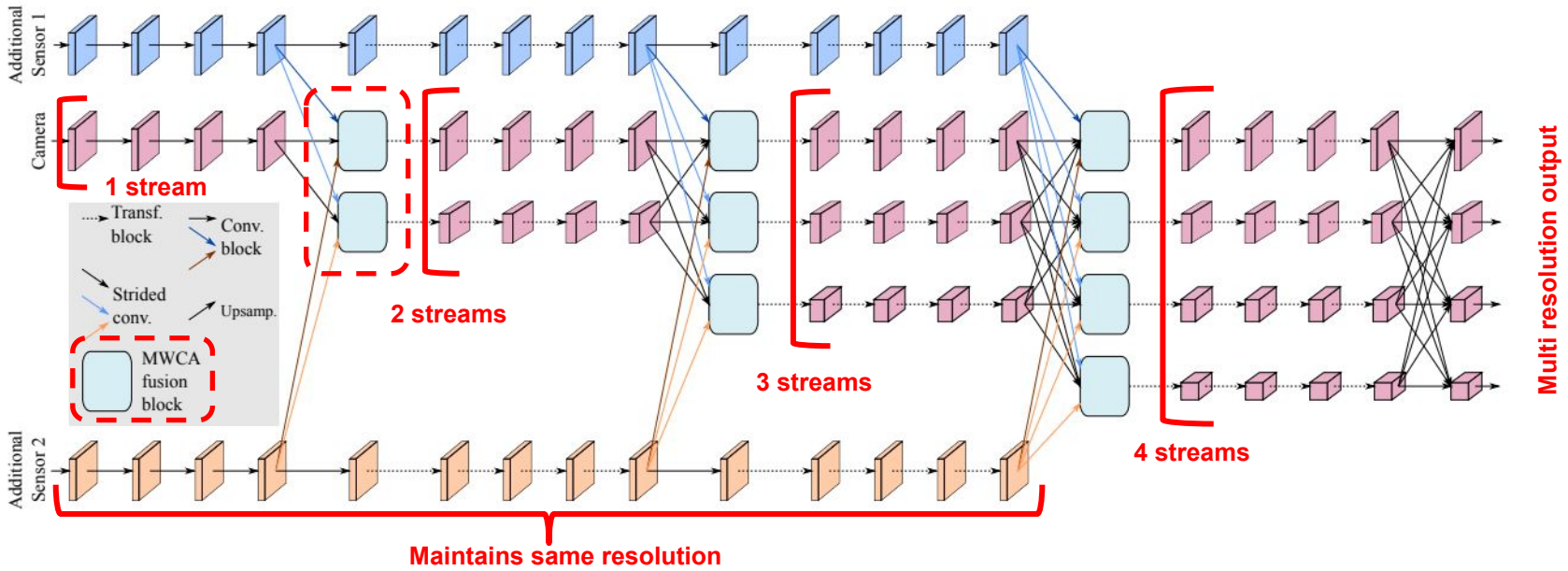


Figure 20: HRFuser architecture (Image adapted from [17])

[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Method 2: HRFuser

Name	Sensors	Dataset Used	Fusion Method
HRFuser	CRL	nuScenes, DENSE	Tightly-coupled

HRFuser: High Resolution Fuser

- Model architecture (2/3)

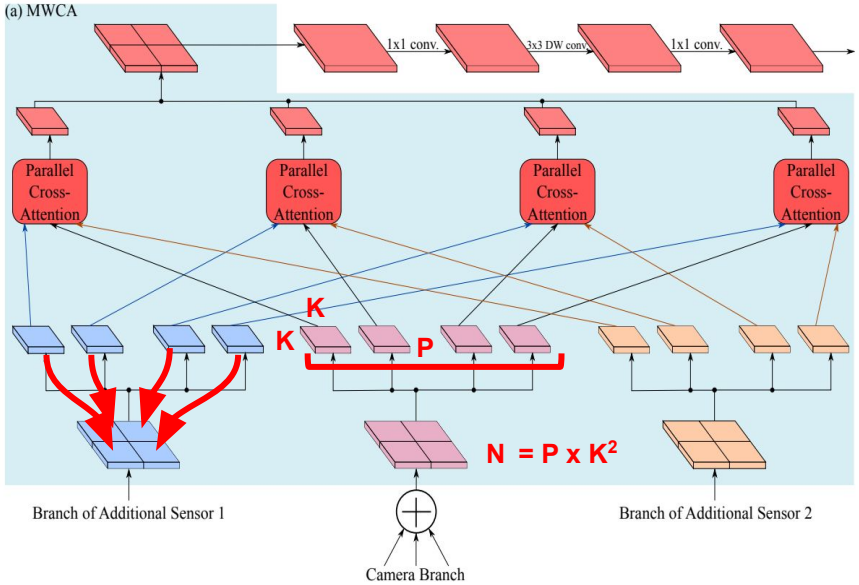


Figure 21: Proposed multi-window cross-attention (MWCA) block followed by a feed-forward network. DW conv. denotes depth-wise convolution. (Image adapted from [17]).

[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Method 2: HRFuser

Name	Sensors	Dataset Used	Fusion Method
HRFuser	CRL	nuScenes, DENSE	Tightly-coupled

HRFuser: High Resolution Fuser

- Model architecture (3/3)

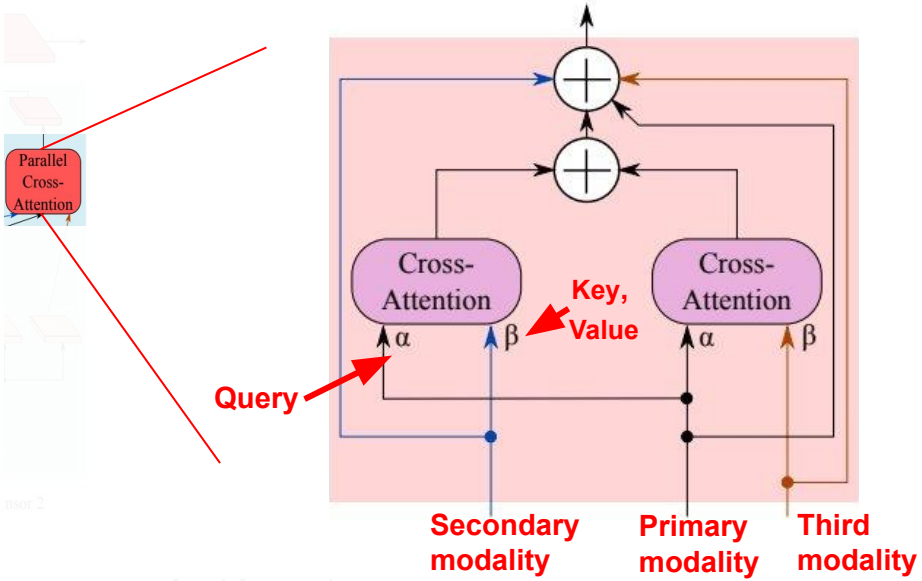


Figure 22: Parallel cross-attention block (Image adapted from [17])

[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Method 2: HRFuser

Name	Sensors	Dataset Used	Fusion Method
HRFuser	CRL	nuScenes, DENSE	Tightly-coupled

HRFuser: High Resolution Fuser

- Loss function [17]
  - Classification loss: Crossentropy loss
  - BBox regression loss: Smooth L1
  - Balancing weight: λ
  - Positive sample indicator

$$L(c_i, t_i) = \frac{1}{N_{\text{pos}}} \left(\sum_i L_{\text{cls}}(c_i, c_i^*) + \lambda \sum_i \mathbb{1}_{c_i^* > 0} L_{\text{reg}}(t_i, t_i^*) \right)$$

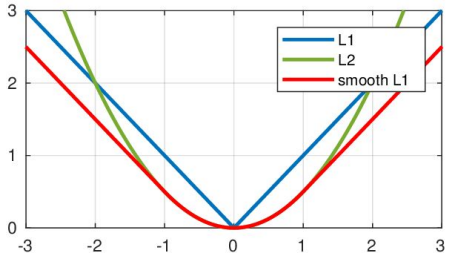
[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Method 2: HRFuser

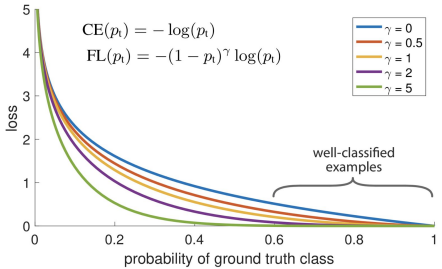
Name	Sensors	Dataset Used	Fusion Method
HRFuser	CRL	nuScenes, DENSE	Tightly-coupled

HRFuser: High Resolution Fuser

- Loss function [17]
  - Classification loss: Crossentropy loss
  - BBox regression loss: Smooth L1
  - Balancing weight: λ
  - Positive sample indicator



$$L(c_i, t_i) = \frac{1}{N_{\text{pos}}} \left(\sum_i L_{\text{cls}}(c_i, c_i^*) + \lambda \sum_i \mathbb{1}_{c_i^* > 0} L_{\text{reg}}(t_i, t_i^*) \right)$$



[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Method 3: MT-DETR

Name	Sensors	Dataset Used	Fusion Method
MT-DETR[16]	CRL	DENSE	Tightly-coupled

MT-DETR: MulTi-sensor MulTimodal DEtection TRansformer

- Preprocessing: LiDAR & Radar imagery

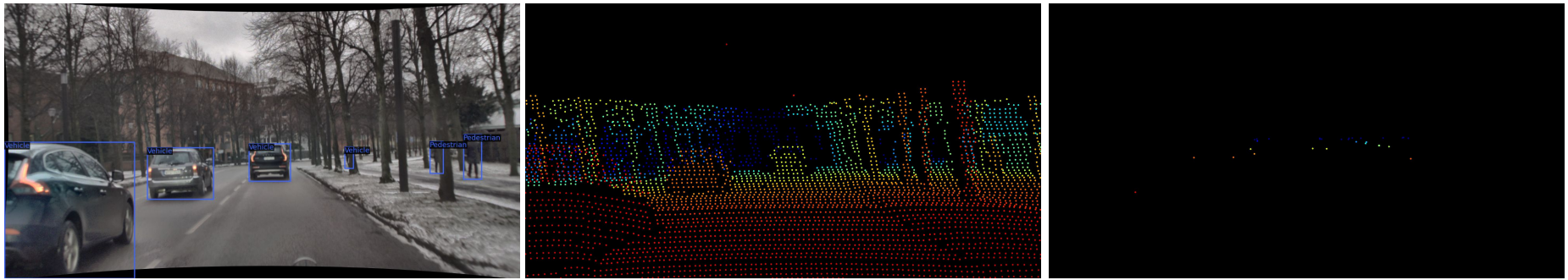


Figure 23: Sample projection images for LiDAR and Radar from DENSE dataset [2]. Each point is colored according to its depth value.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Method 3: MT-DETR

Name	Sensors	Dataset Used	Fusion Method
MT-DETR	CRL	DENSE	Tightly-coupled

MT-DETR: MulTi-sensor MulTimodal DEtection TRansformer

- Preprocessing: LiDAR & Radar imagery

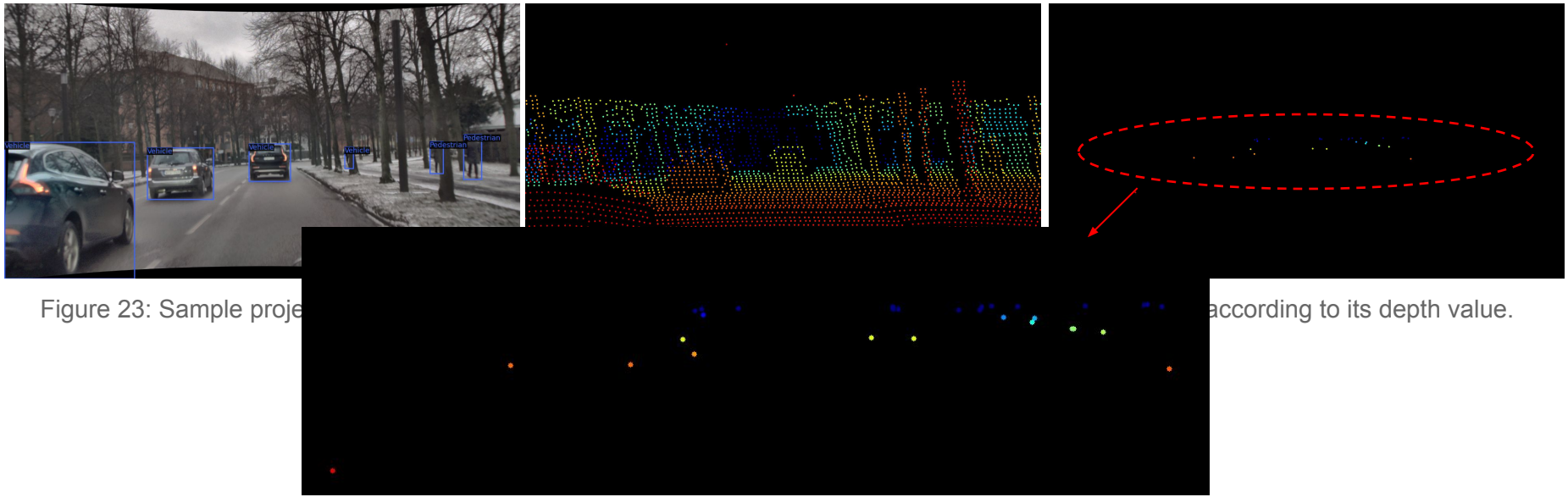


Figure 23: Sample proje ...according to its depth value.

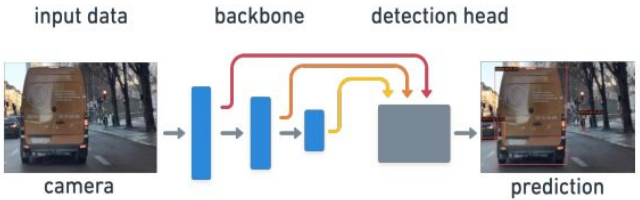
[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Method 3: MT-DETR

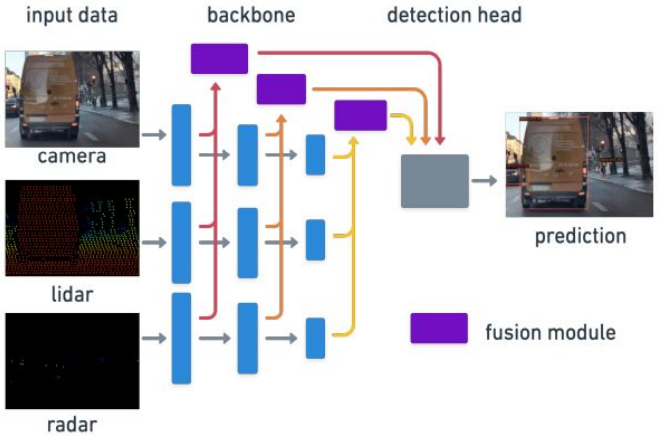
Name	Sensors	Dataset Used	Fusion Method
MT-DETR	CRL	DENSE	Tightly-coupled

MT-DETR: MulTi-sensor MulTimodal DEtection TRansformer

- Model architecture (1/3)



(a) Unimodal



(b) Middle fusion

Figure 24: Comparison of two types of object detection model architecture. (a) Unimodal architecture with camera image. (b) Middle fusion based multimodal architecture with three modalities (Image adapted from [16])

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Method 3: MT-DETR

Name	Sensors	Dataset Used	Fusion Method
MT-DETR	CRL	DENSE	Tightly-coupled

MT-DETR: MulTi-sensor MulTimodal DEtection TRansformer

- Model architecture (2/3)

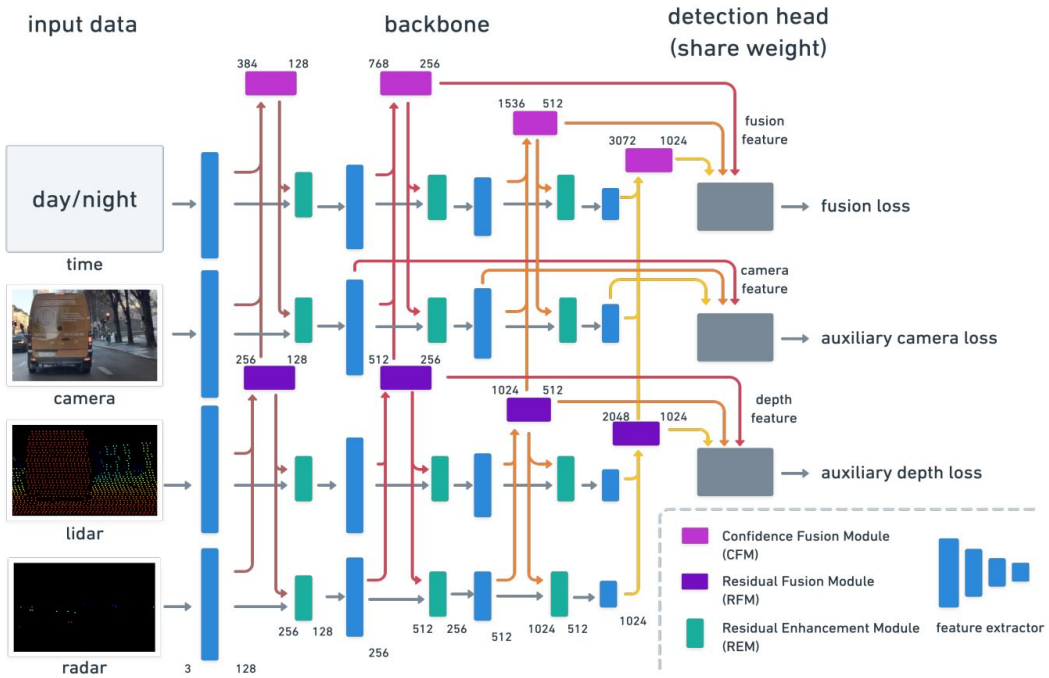


Figure 25: MT-DETR model architecture (Image source [16])

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Method 3: MT-DETR

Name	Sensors	Dataset Used	Fusion Method
MT-DETR	CRL	DENSE	Tightly-coupled

MT-DETR: MulTi-sensor MulTimodal DEtection TRansformer

- Model architecture (2/3)

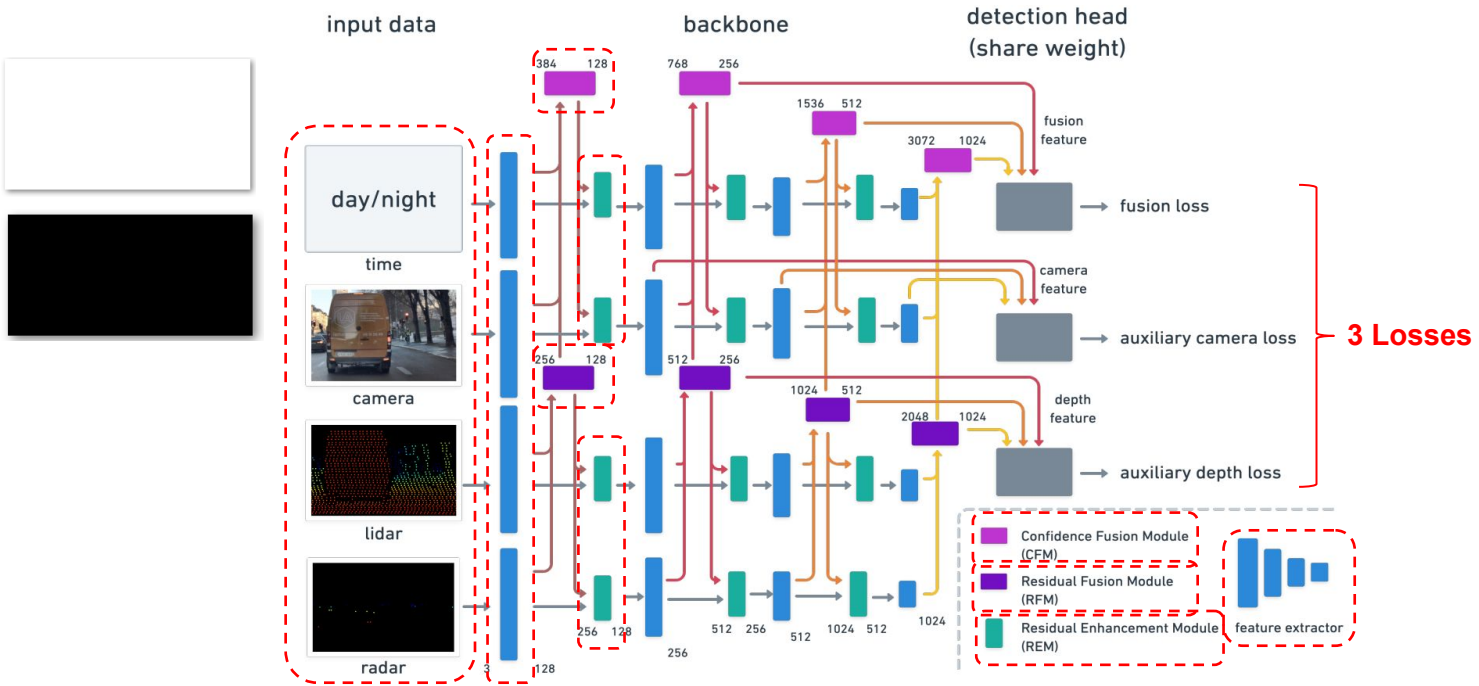


Figure 25: MT-DETR model architecture (Image source [16])

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Method 3: MT-DETR

Name	Sensors	Dataset Used	Fusion Method
MT-DETR	CRL	DENSE	Tightly-coupled

MT-DETR: MulTi-sensor MulTimodal DEtection TRansformer

- Model architecture (3/3)

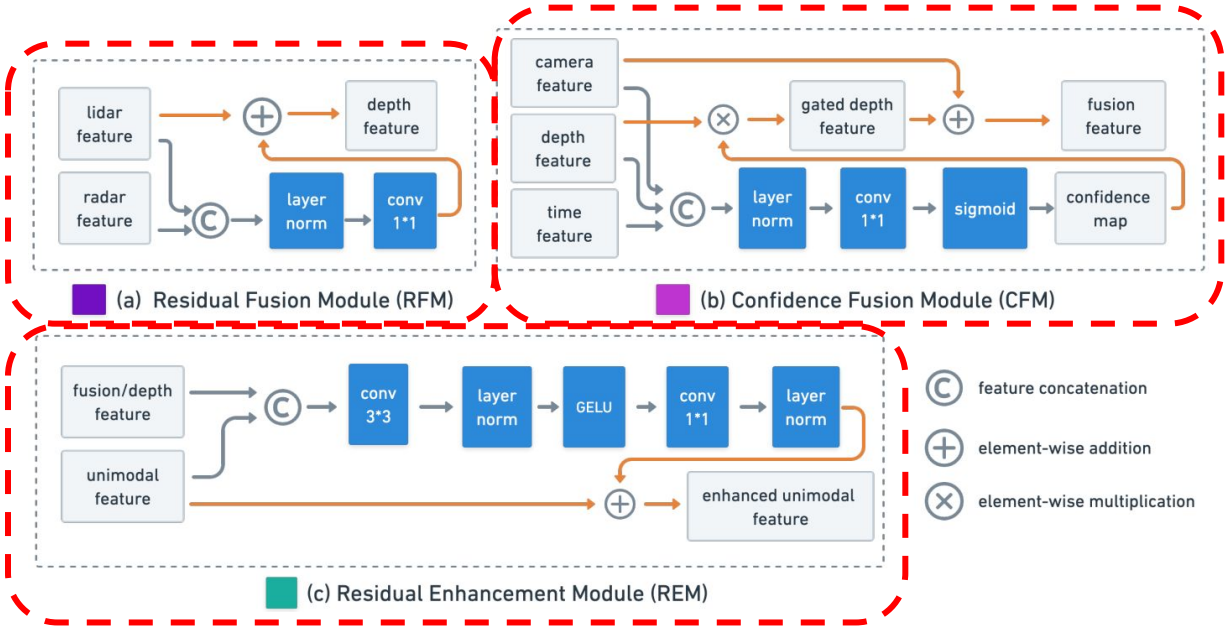


Figure 26: Fusion modules and enhancement module design (Image source [16])

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Method 3: MT-DETR

Name	Sensors	Dataset Used	Fusion Method
MT-DETR	CRL	DENSE	Tightly-coupled

MT-DETR: MulTi-sensor MulTimodal DEtection TRansformer

- Loss function [16]
  - Classification loss: Focal loss
  - BBox regression loss: L1 & GIoU
  - Balancing weight: λ
  - Positive sample indicator

$$L(c_i, t_i) = \frac{1}{N_{\text{pos}}} \left(\sum_i L_{\text{cls}}(c_i, c_i^*) + \lambda \sum_i \mathbb{1}_{c_i^* > 0} L_{\text{reg}}(t_i, t_i^*) \right)$$

$$\mathcal{L}_m = 2\mathcal{L}_{\text{focal}}(P_m, \hat{P}) + 5\mathcal{L}_1(P_m, \hat{P}) + 2\mathcal{L}_{\text{GIoU}}(P_m, \hat{P})$$

$$\mathcal{L}_{total} = \lambda_{fusion}\mathcal{L}_{fusion} + \lambda_{camera}\mathcal{L}_{camera} + \lambda_{depth}\mathcal{L}_{depth}$$

3 Losses

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[31] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 658–666.

Experiment Overview

- 3 methods
  - SAF-FCOS, HRFuser, and MT-DETR
  - Middle and Tightly-coupled fusion
- 2 datasets
  - nuScenes and DENSE
- 3 sensors
  - Camera, LiDAR, and Radar
- Weather conditions
  - Day, night, light fog, dense fog, rain, snow
- Metrics
  - COCO style

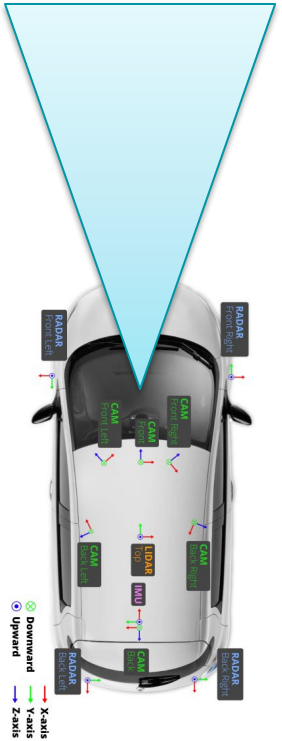


Figure 27: Vehicle front view (Image adapted [21])

[21] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

Experiment Setup

- Computational resource
  - NVIDIA V100 GPU
  - NVIDIA A100 GPU
  - NVIDIA RTX 3090
- Software environment

Table 5: PyTorch and CUDA versions

Methods	PyTorch	CUDA
SAF-FCOS	1.12	11.6
HRFuser	1.10	10.2 / 11.1
MT-DETR	1.10	11.1

Experiment Setup – Dataset Split

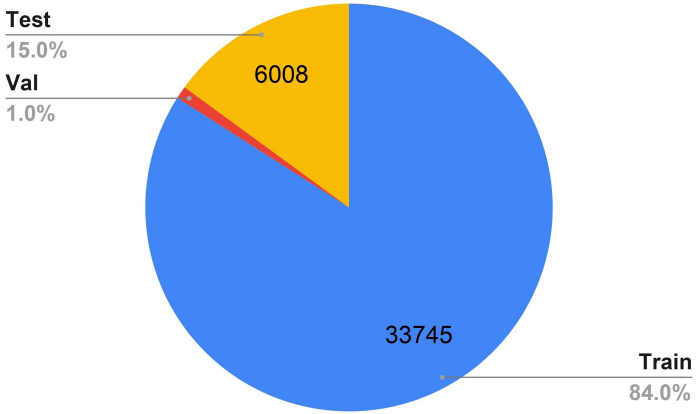


Figure 28: Data split of nuScenes [21] dataset.
Total samples: 40157.
Note: this is a custom dataset split chosen by the authors.

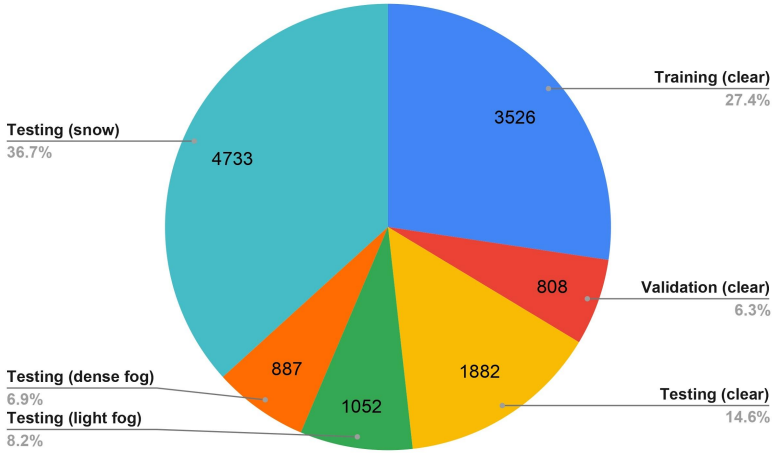


Figure 29: Data split for DENSE [1] dataset.
Total samples: 12888.
Note: all methods use clear weather data for training and validation and use adverse weather data for testing.

[1] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, “Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather,” pp. 11 682–11 692, 2020.
[21] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

Experiment Setup – Model Configuration & Training

Table 5: Training parameters. (*) denotes total batch size for models trained on multiple GPUs.

Methods	Optimizer	Learning Rate	Batch Size	Epochs	Input (WxH)
SAF-FCOS [9]	SGD	0.001	8	12	1333 x 800
HRFuser [17] (nuScenes)	AdamW	0.0003	12*	12	640x384
HRFuser [17] (DENSE)	AdamW	0.001	12*	60	1284x384
MT-DETR [16]	AdamW	0.0001	1	36	1333 x 800

- Learning rate decay and warmup schemes
- Checkpoint based on validation loss
- Inference confidence threshold 0.3
- Augmentation used:
  - RandomFlip, Crop, Resize, RandomDrop

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, “Spatial attention fusion for obstacle detection using mmwave radar and vision sensor,” Sensors, vol. 20, no. 4, p. 956, 2020.
[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Experiment Setup – Model Configuration & Training

Table 6: Available configurations for each method

	Fusion Arch.			Modalities			
Methods	**Early**	**Middle**	**Tightly-coupled**	**Camera**	**Radar**	**Lidar**	**Extra**
SAF-FCOS[9]	×	✓	×	✓	✓	×	×
HRFuser[17]	×	×	✓	✓	✓	✓	Gated
MT-DETR[16]	✓	✓	✓	✓	✓	✓	Time

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, “Spatial attention fusion for obstacle detection using mmwave radar and vision sensor,” Sensors, vol. 20, no. 4, p. 956, 2020.
[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Experiment Setup – Model Configuration & Training

Sr.	Method		Fusion Architecture			Sensors*			
	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
1	SAF-FCOS	HRFuser	x	✓	✓	✓	x	✓	x
2	HRFuser	-	x	x	✓	✓	✓	✓	x
3	HRFuser	-	x	x	✓	✓	✓	✓	G
4	MT-DETR	-	✓	x	x	✓	✓	✓	x
5	MT-DETR	-	x	✓	x	✓	✓	✓	x
6	MT-DETR	-	x	x	✓	✓	✓	✓	x
7	MT-DETR	-	✓	✓	✓	✓	x	✓	x
8	MT-DETR	-	✓	✓	✓	✓	✓	✓	x
9	MT-DETR	-	x	x	✓	✓	✓	✓	T
10	HRFuser	MT-DETR	x	x	✓	✓	x	✓	x
11	HRFuser	MT-DETR	x	x	✓	✓	✓	x	x
12	HRFuser	MT-DETR	x	x	✓	✓	✓	✓	x
13	HRFuser	MT-DETR	x	x	✓	✓	✓	✓	G, T

Table 7: Overview of experiment combinations. C, L, R, G, and T denote Camera, LiDAR, Radar, Gated camera, and Time

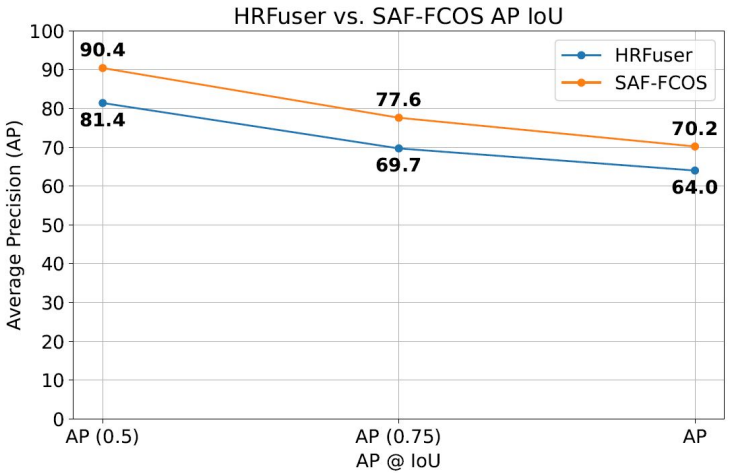
NOTE: Most of the experiments were carried out utilizing the DENSE dataset in conjunction with the MT-DETR method. Due to computational complexities and resource constraints, it is not feasible to evaluate every possible combination.

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, “Spatial attention fusion for obstacle detection using mmwave radar and vision sensor,” Sensors, vol. 20, no. 4, p. 956, 2020.
[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

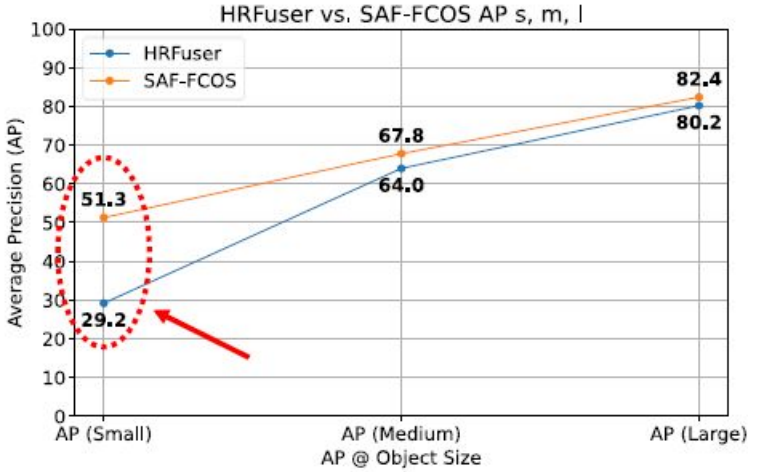
Results – Exp. 1/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
1	SAF-FCOS	HRFuser	x	✓	✓	✓	x	✓	x

Middle vs Tightly-c. | CR



(a) Average Precision (AP)



(b) Average Precision (AP) at S, M, L

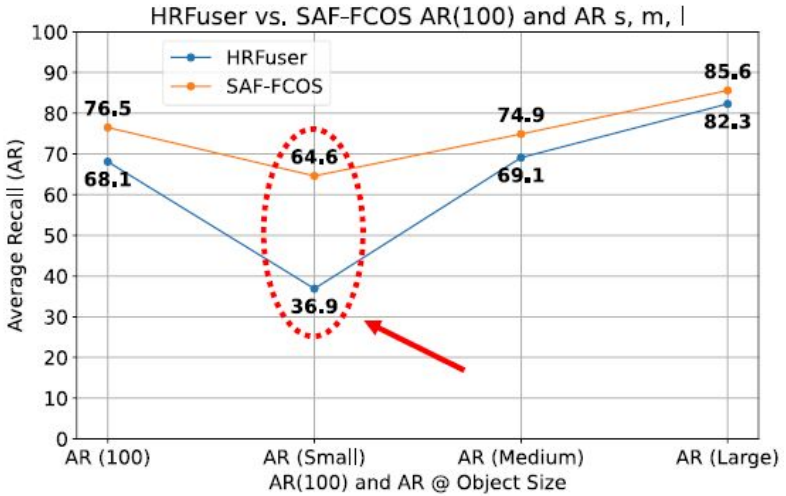
Figure 30: Comparative Analysis, HRFuser[17] vs SAF-FCOS [9]

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, "Spatial attention fusion for obstacle detection using mmwave radar and vision sensor," Sensors, vol. 20, no. 4, p. 956, 2020.
[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

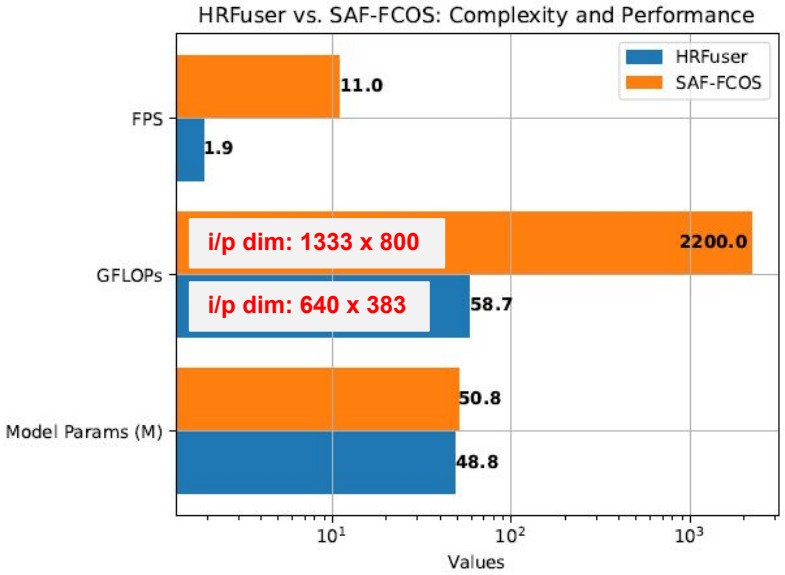
Results – Exp. 1/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
1	SAF-FCOS	HRFuser	x	✓	✓	✓	x	✓	x

Middle vs Tightly-c. | CR



(c) Average Recall (AR)



(d) Model Complexity

Figure 30: Comparative Analysis, HRFuser[17] vs SAF-FCOS [9]

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, "Spatial attention fusion for obstacle detection using mmwave radar and vision sensor," Sensors, vol. 20, no. 4, p. 956, 2020.
[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Results – Exp. 1/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
1	SAF-FCOS	HRFuser	x	✓	✓	✓	x	✓	x

Middle vs Tightly-c. | CR

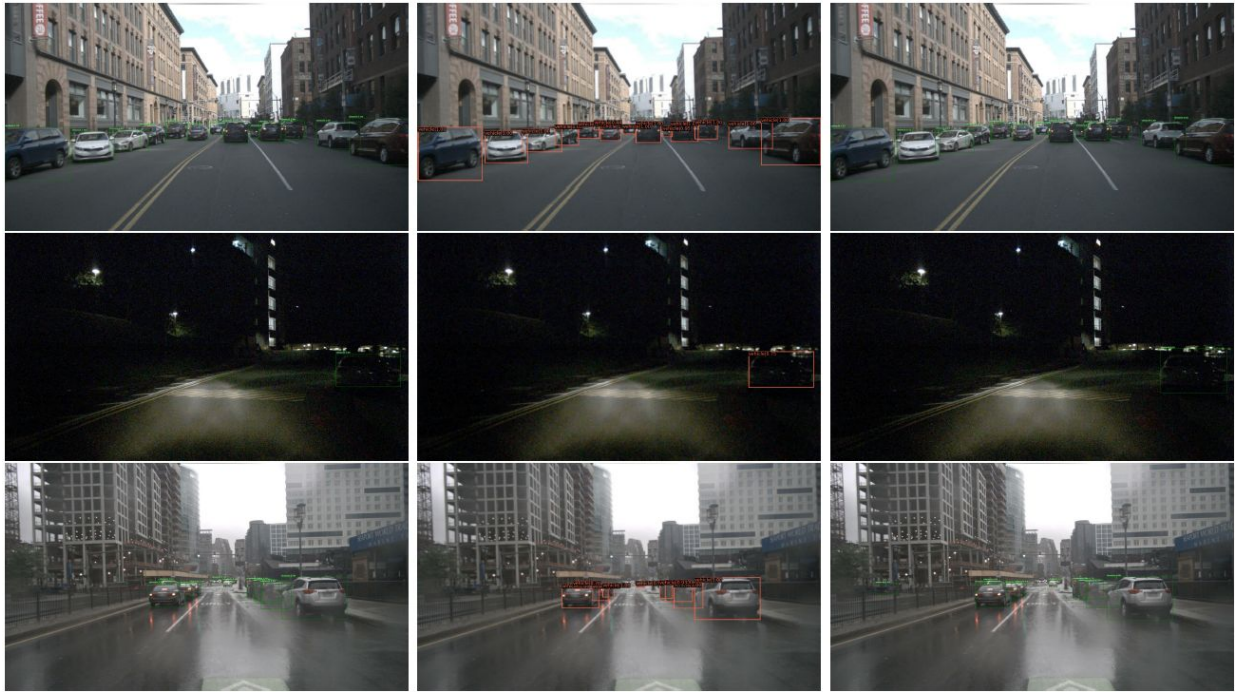


Figure 31: Detection results of HRFuser and SAF-FCOS. Where the 1st column is ground truth, 2nd is HRFuser, and 3rd is SAF-FCOS (1st Row: Day, 2nd: Night and 3rd : Rain)

Results – Exp. 1/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
1	SAF-FCOS	HRFuser	x	✓	✓	✓	x	✓	x

Middle vs Tightly-c. | CR

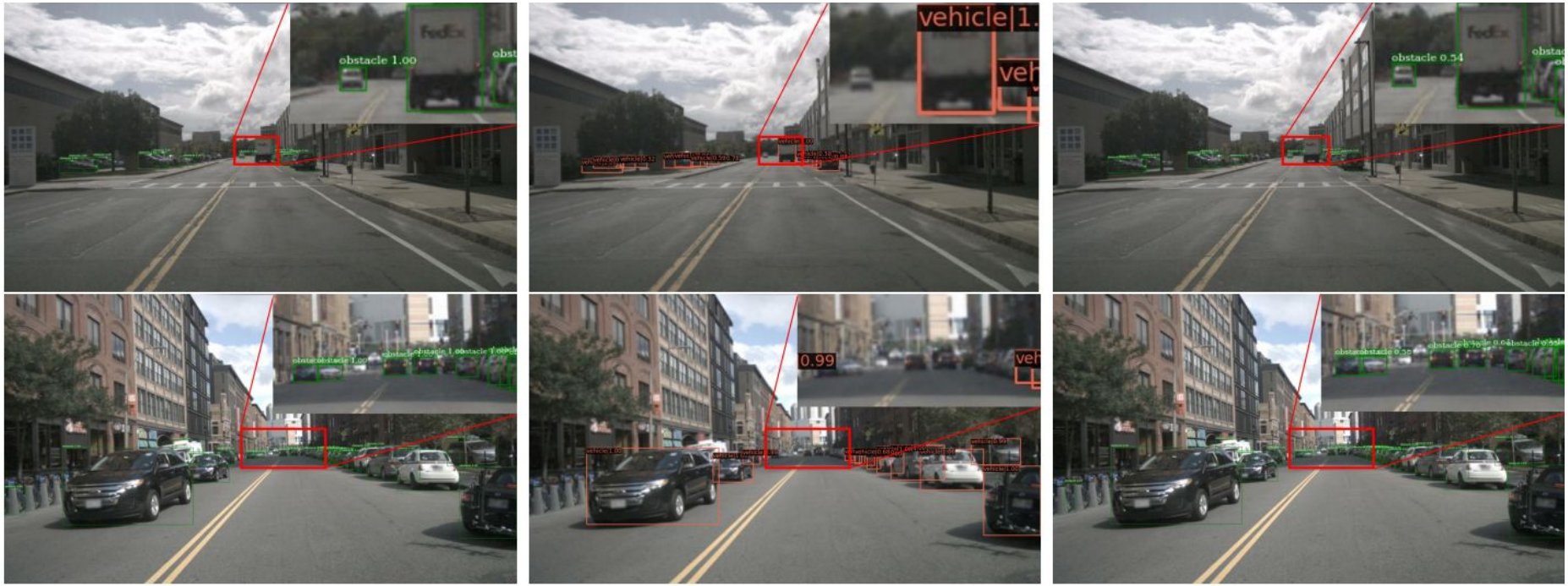


Figure 32: Small object detection results of HRFuser and SAF-FCOS. Where the 1st col is ground truth, 2nd HRFuser, and 3rd SAF-FCOS

Results – Exp. 2/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
2	HRFuser	-	x	x	✓	✓	✓	✓	x

Tightly-c. | C vs CR vs CL vs CLR

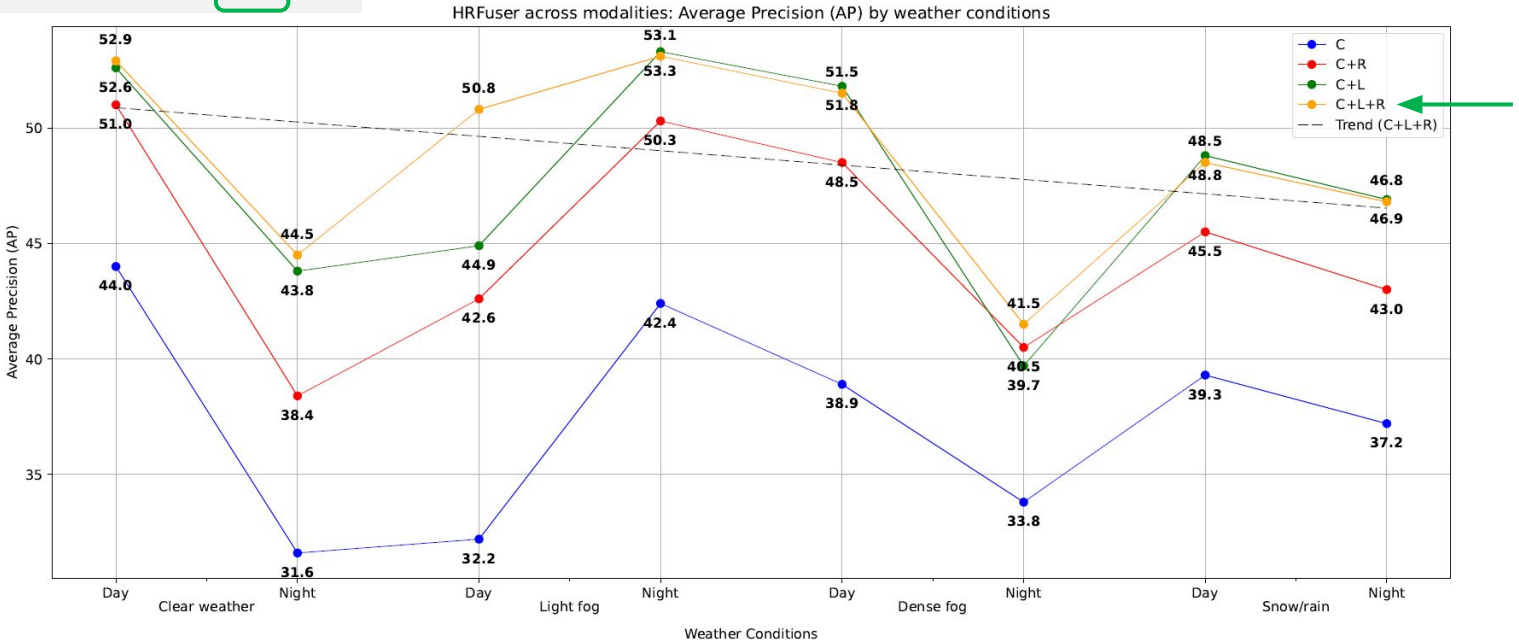


Figure 33: Performance of the HRFuser[17] model in terms of Average Precision (AP) across various sensor modalities under different weather conditions. Note: C+L+R data point values are shown above the line and rest are below the line.

[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Results – Exp. 3/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
3	HRFuser	-	x	x	✓	✓	✓	✓	G

Tightly-c. | C vs CR vs CL vs CLR vs CLRG

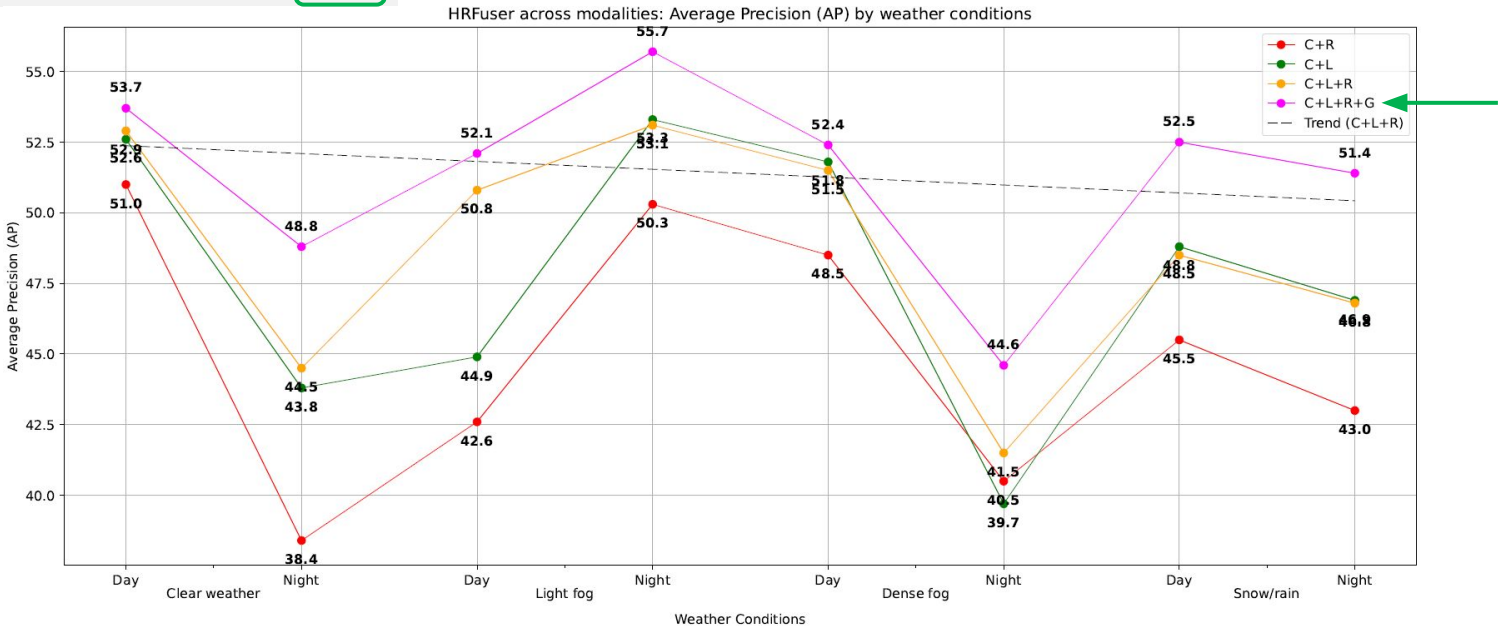


Figure 34: Performance of the HRFuser[17] model in terms of Average Precision (AP) across various sensor modalities under different weather conditions with an additional sensor Gated infrared camera. Note: C+L+R+G data point values are shown above the line and rest are below the line.

[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Results – Exp. 3/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
3	HRFuser	-	x	x	✓	✓	✓	✓	x

Tightly-c. | C vs CR vs CL vs CLR vs CLRG

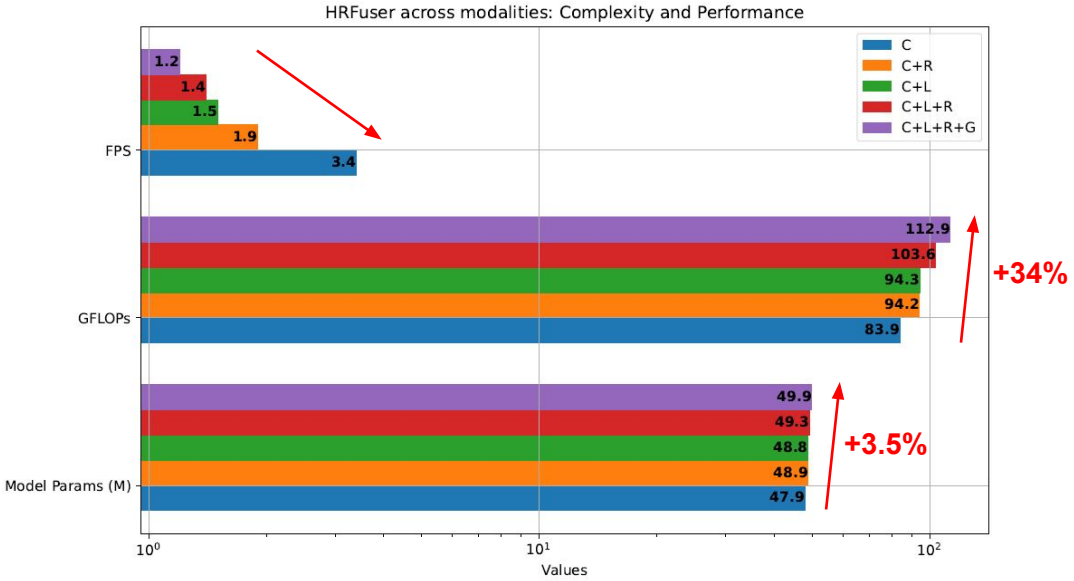


Figure 35: HRFuser[17] model's complexity and performance, displayed through model parameters (in M), GFLOPs, and FPS, across different sensor modalities. Note: this is a common plot with an additional modality, Gated infrared camera.

[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Results – Exp. 3/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
3	HRFuser	-	x	x	✓	✓	✓	✓	x

Tightly-c. | C vs CLR

Figure 35: A few samples to showcase the detection results of HRFuser. Where the 1st column is ground truth, 2nd is Camera only, and 3rd is Tightly-Coupled Fusion with C+L+R (Rows show the challenging weather conditions).

Results – Exp. 3/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
3	HRFuser	-	x	x	✓	✓	✓	✓	x

Tightly-c. | C vs CLR

Figure 37: A few samples to showcase the detection results of HRFuser. Where the 1st column is ground truth, 2nd is Camera only, and 3rd is Tightly-Coupled Fusion with C+L+R (Rows show the challenging weather conditions).

Results – Exp. 4/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
4	MT-DETR	-	✓	x	x	✓	✓	✓	x

Early | C vs CR vs CLR

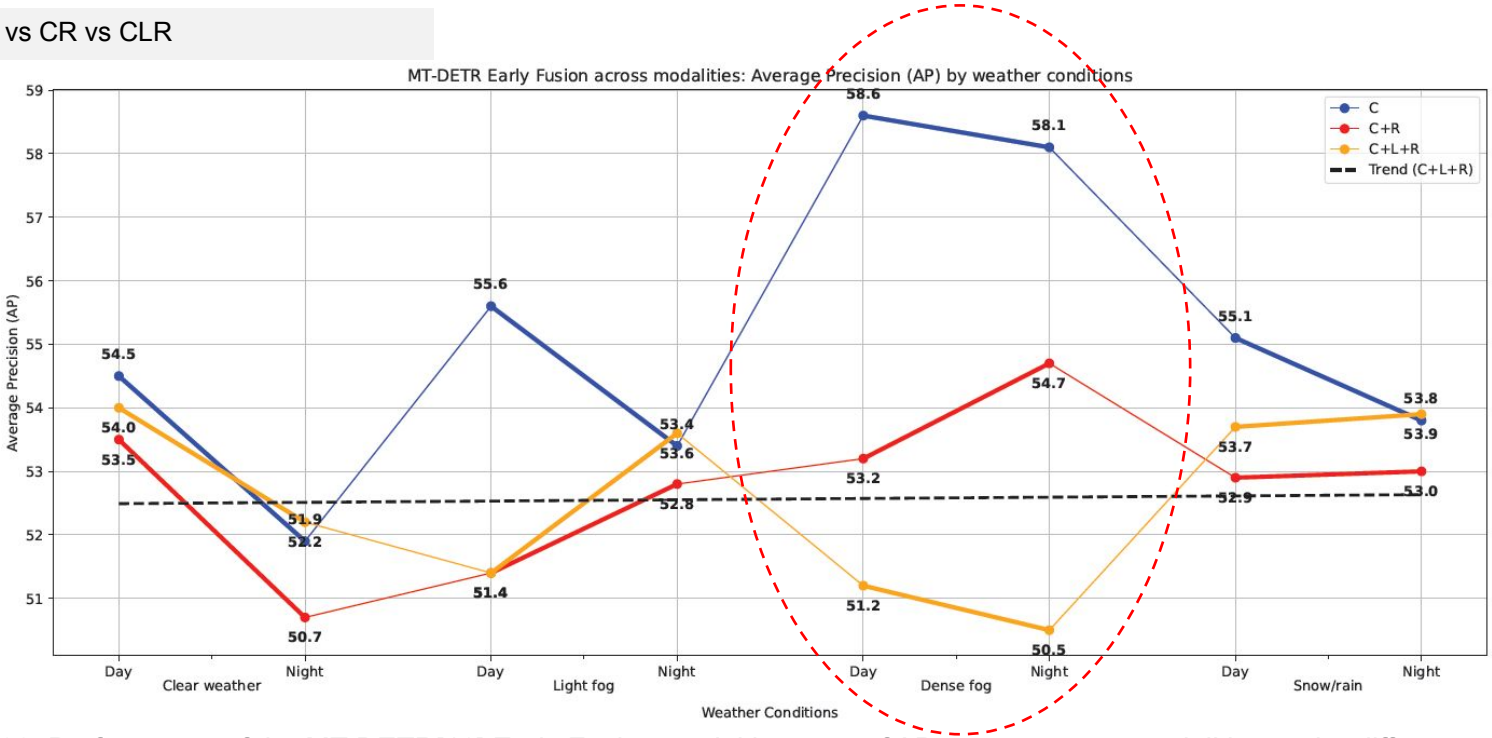


Figure 38: Performance of the MT-DETR[16] Early Fusion model in terms of AP across sensor modalities under different weather conditions. Note: C data point values are shown above the line and rest are below the line.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Results – Exp. 4/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
4	MT-DETR	-	✓	x	x	✓	✓	✓	x

Early | C vs CR vs CLR

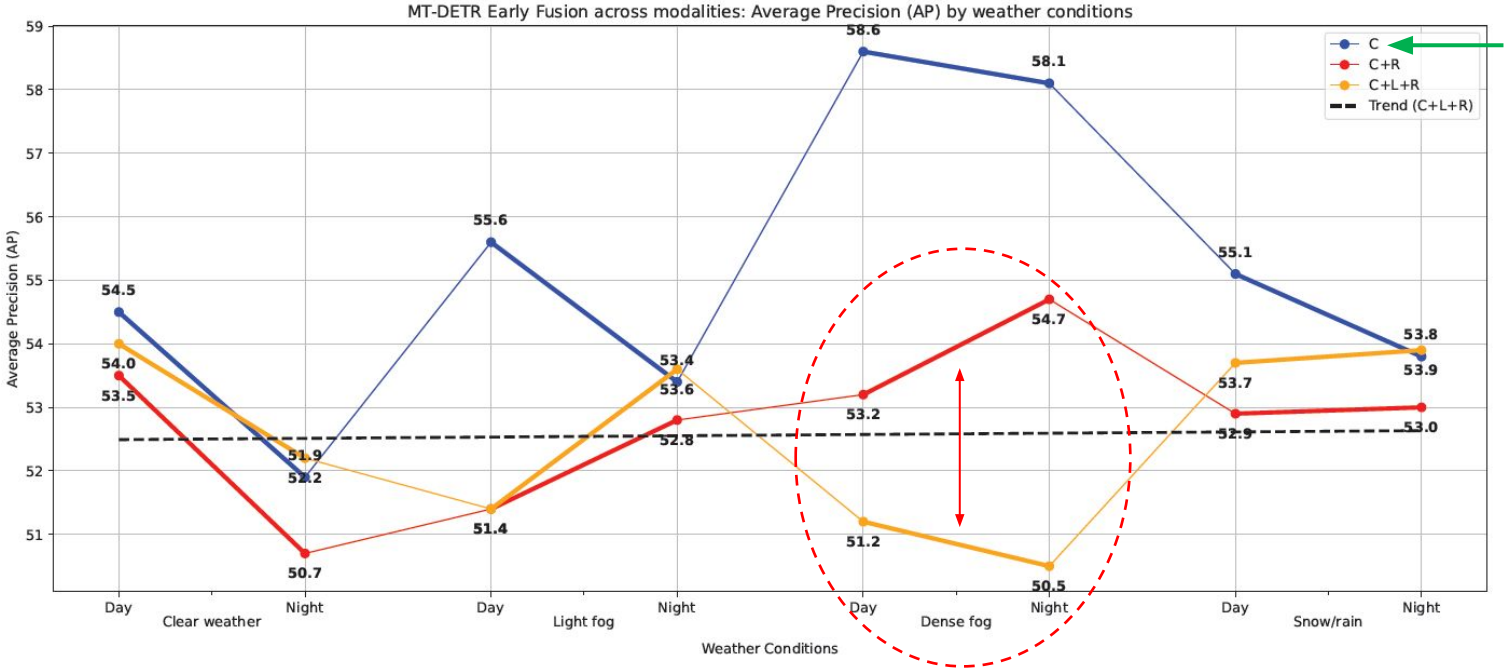


Figure 39: Performance of the MT-DETR Early Fusion model in terms of AP across sensor modalities under different weather conditions. Note: C data point values are shown above the line and rest are below the line.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Results – Exp. 4/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
4	MT-DETR	-	✓	x	x	✓	✓	✓	x

Early | C vs CR vs CLR

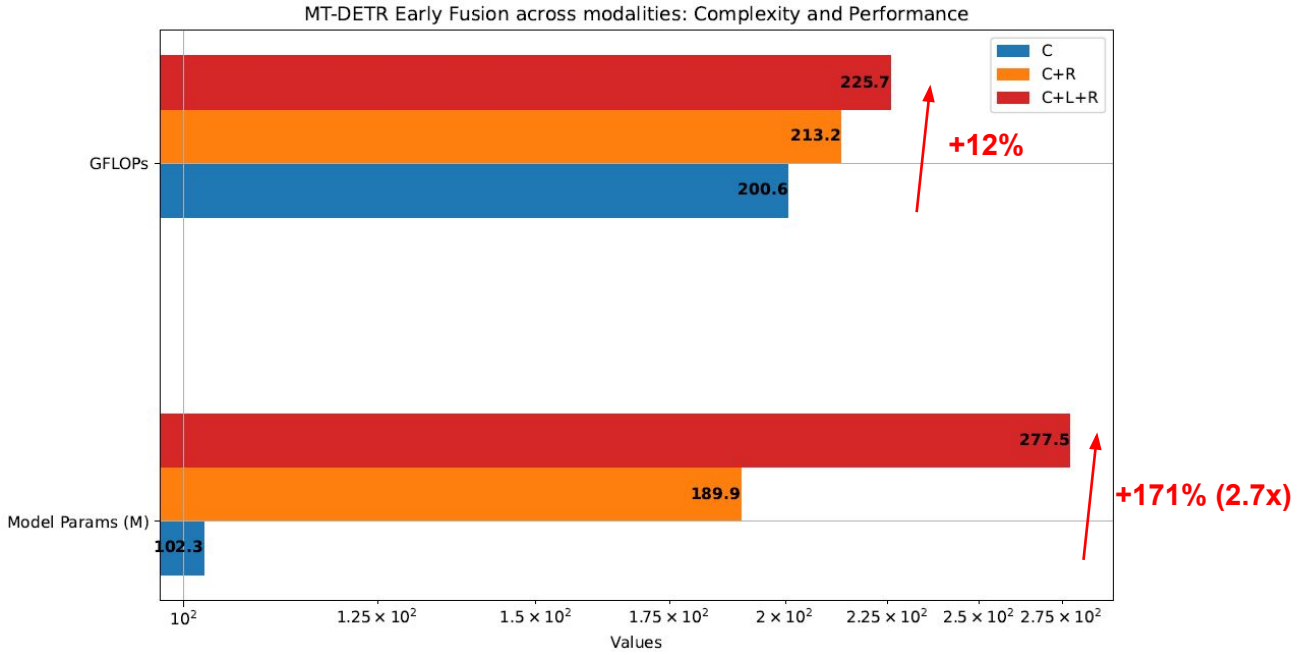


Figure 40: MT-DETR Early Fusion model's complexity, displayed through model parameters (in M), and GFLOPs across different sensor modalities.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Results – Exp. 5/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
5	MT-DETR	-	x	✓	x	✓	✓	✓	x

Middle | C vs CR vs CLR

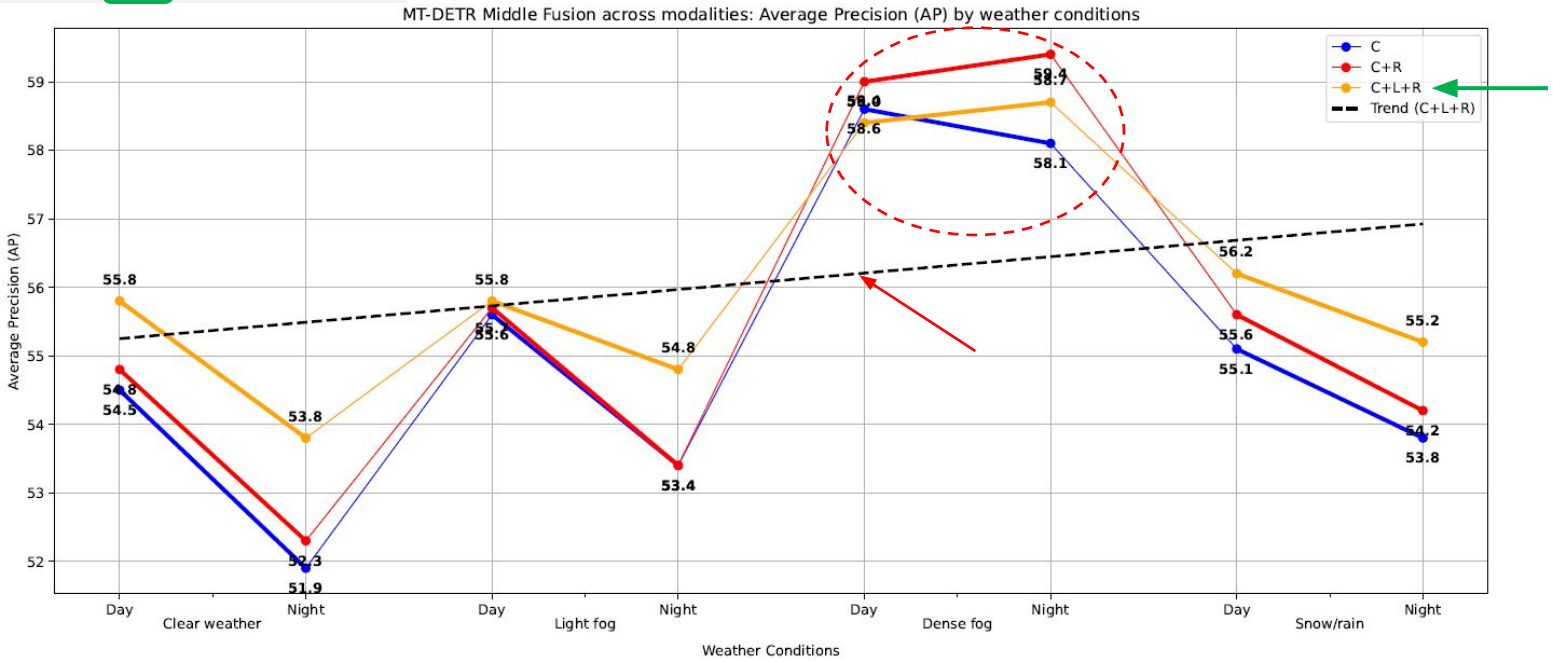


Figure 41: Performance of the MT-DETR Middle Fusion model in terms of AP across sensor modalities under different weather conditions. Note: C+L+R data point values are shown above the line and rest are below the line.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Results – Exp. 5/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
5	MT-DETR	-	x	✓	x	✓	✓	✓	x

Middle | C vs CR vs CLR

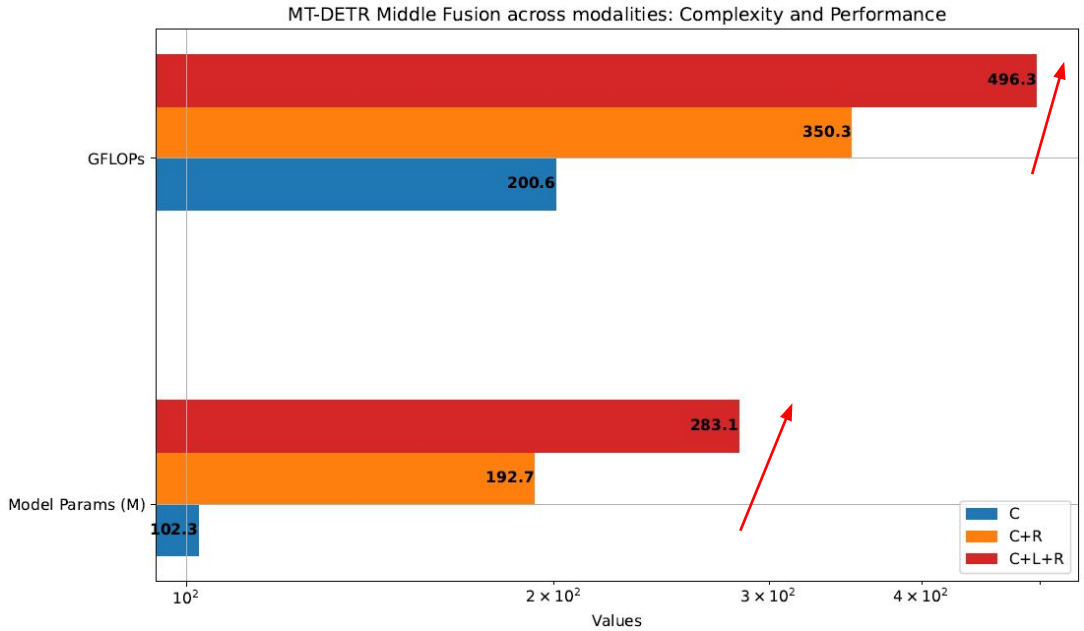


Figure 42: MT-DETR Middle Fusion model's complexity, displayed through model parameters (in M), and GFLOPs across different sensor modalities.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Results – Exp. 6/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
6	MT-DETR	-	x	x	✓	✓	✓	✓	x

Tightly-Coupled | C vs CR vs CLR

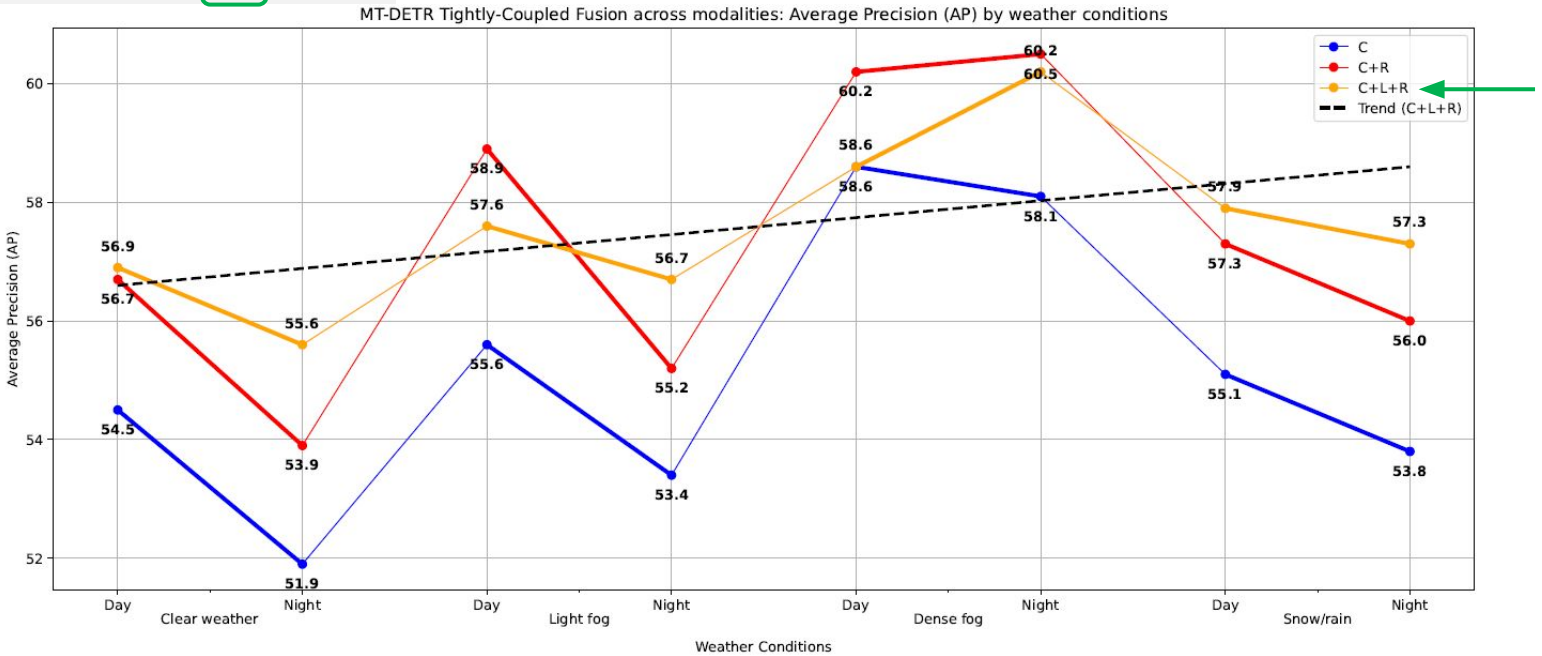


Figure 43: Performance of the MT-DETR Tightly-Coupled Fusion model in terms of AP across sensor modalities under different weather conditions. Note: C+L+R data point values are shown above the line and rest are below the line.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Results – Exp. 7/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
7	MT-DETR	-	✓	✓	✓	✓	x	✓	x

Early vs Middle vs Tightly-Coupled | CR

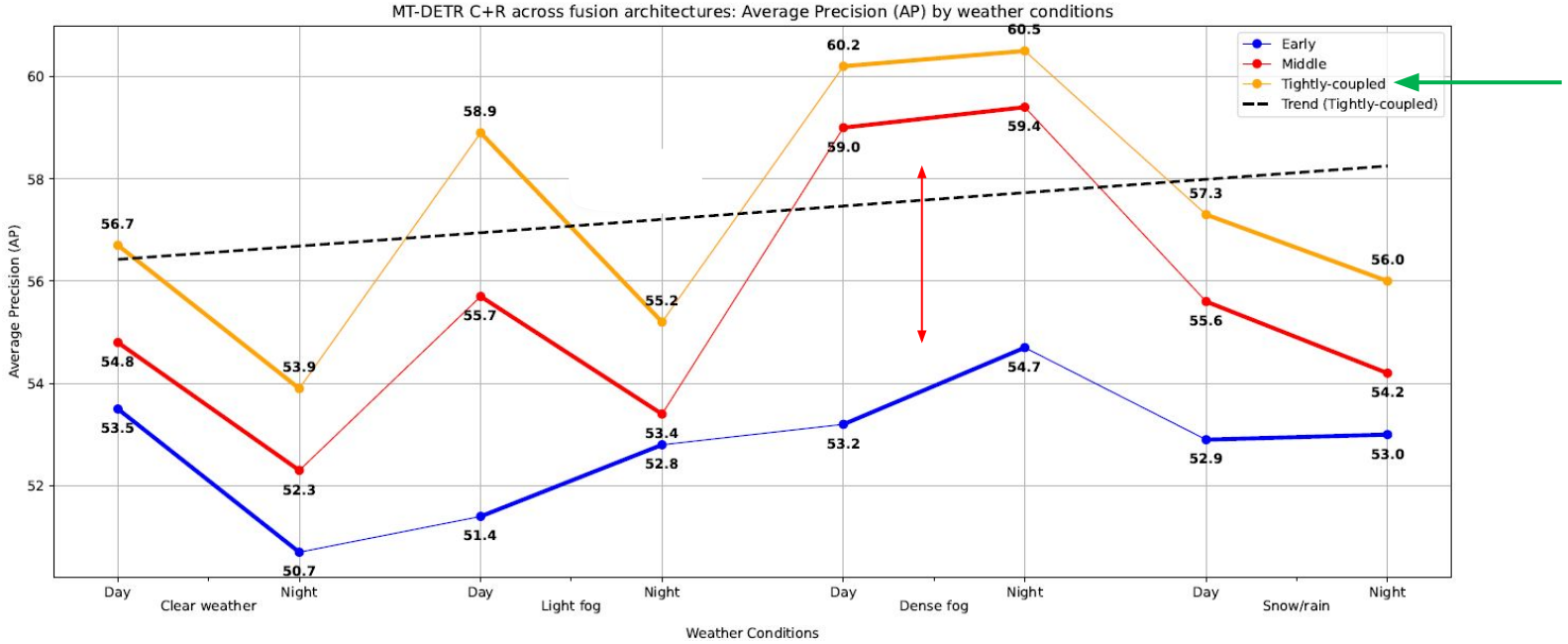


Figure 44: Performance of the MT-DETR across the Early, Middle, and Tightly-Coupled Fusion model with C+R in terms of AP across sensor mod. under diff. weather cond. Note: Tightly- Coupled data point values are shown above the line and rest are below the line.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Results – Exp. 7/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
7	MT-DETR	-	✓	✓	✓	✓	x	✓	x

Early vs Middle vs Tightly-Coupled | CR

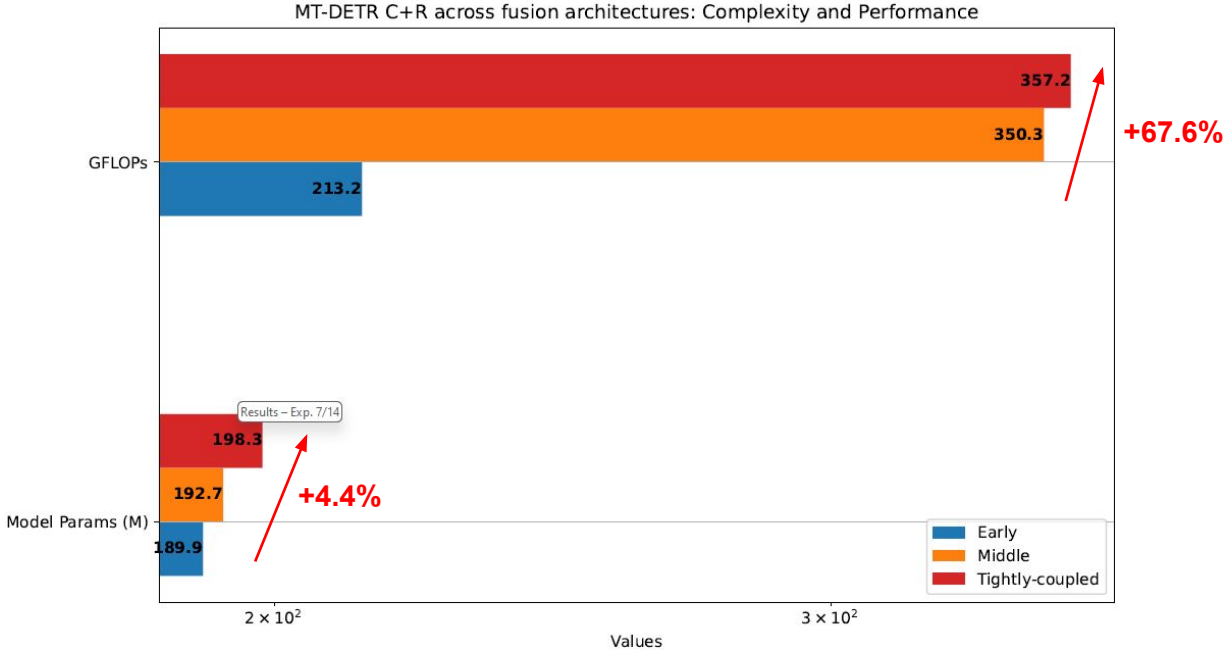


Figure 45: Performance of the MT-DETR across the Early, Middle, and Tightly-Coupled Fusion model with C+R in terms of AP across sensor mod. under diff. weather cond. Note: Tightly- Coupled data point values are shown above the line and rest are below the line.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Results – Exp. 8/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
8	MT-DETR	-	✓	✓	✓	✓	✓	✓	x

Early vs Middle vs Tightly-Coupled | CLR

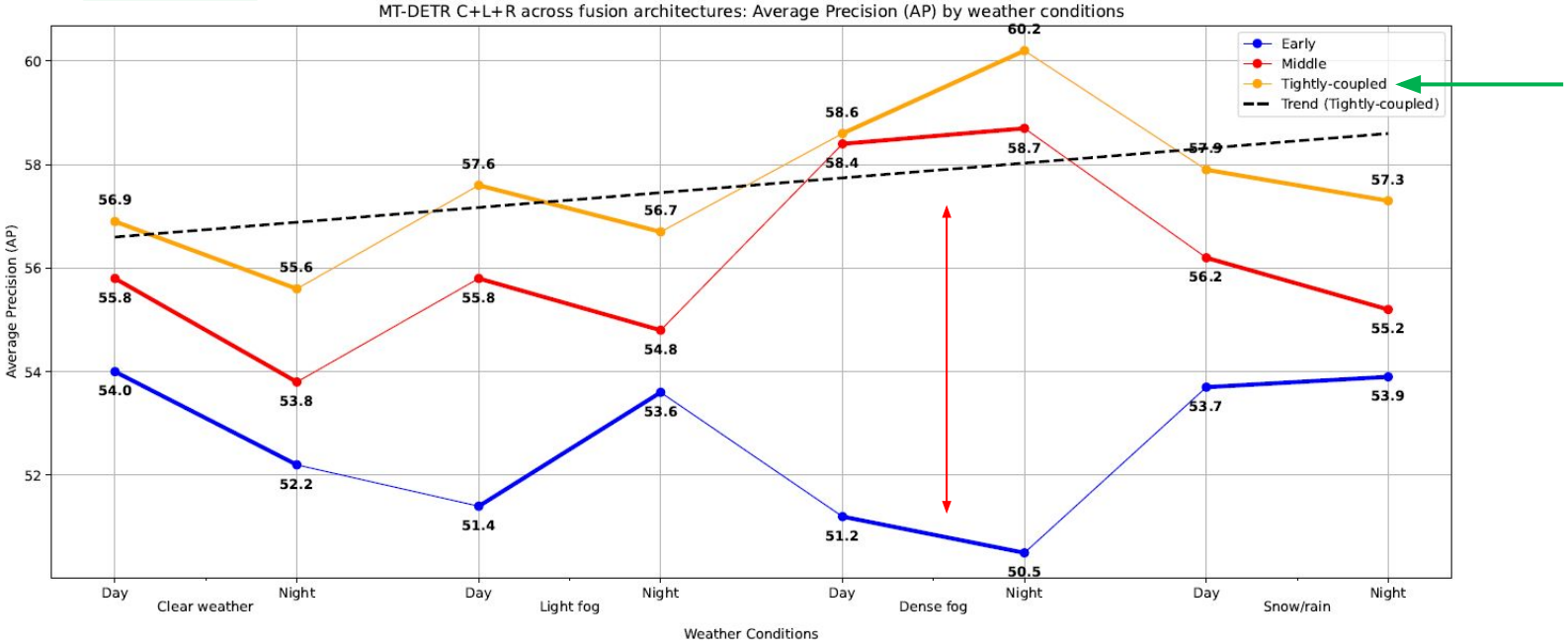


Figure 46: Performance of the MT-DETR across the Early, Middle, and Tightly-Coupled Fusion model with C+L+R in terms of AP across sensor mod. under diff. weather cond. Note: Tightly- Coupled data point values are shown above the line and rest are below the line.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Results – Exp. 9/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
9	MT-DETR	-	x	x	✓	✓	✓	✓	Time

Tightly-Coupled | CR vs CL vs CLR vs CLRT

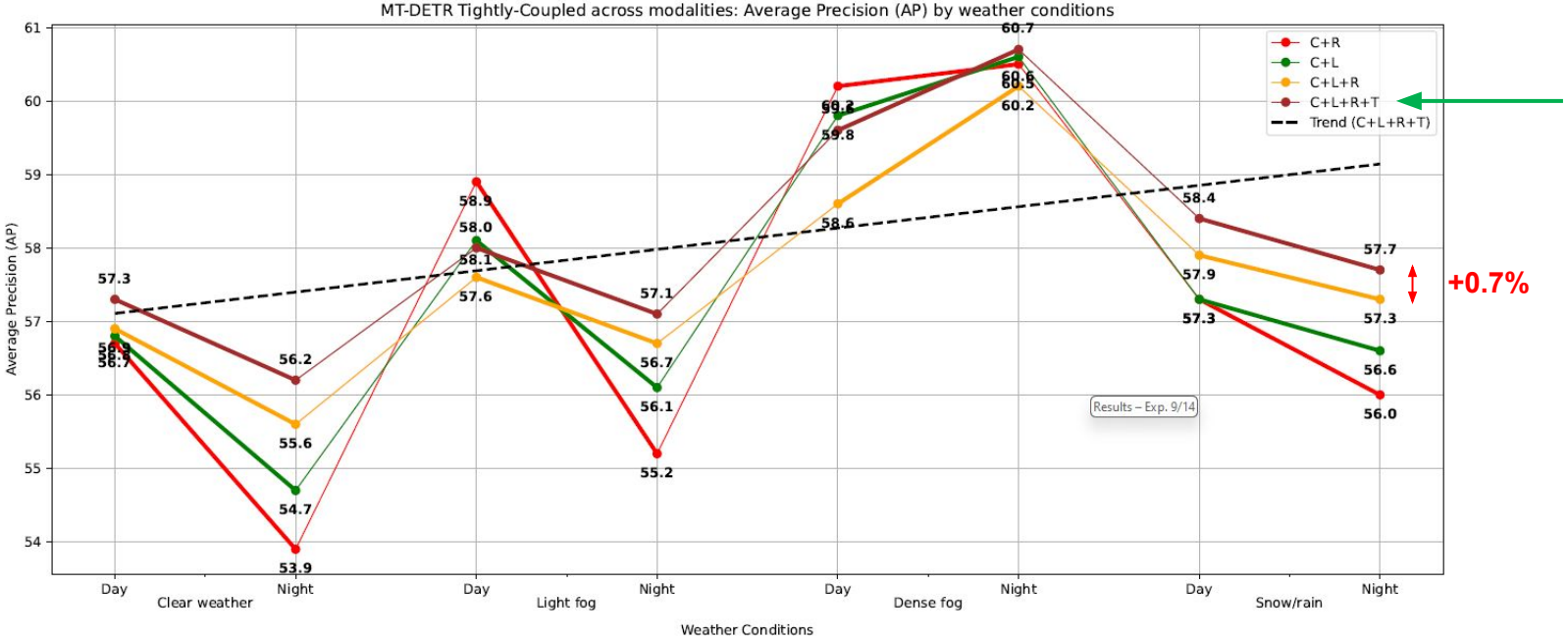


Figure 47: Performance of the MT-DETR across the Early, Middle, and Tightly-Coupled Fusion model with C+L+R in terms of AP across sensor mod. under diff. weather cond. Note: Tightly- Coupled data point values are shown above the line and rest are below the line.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Results – Exp. 9/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
9	MT-DETR	-	x	x	✓	✓	✓	✓	Time

Tightly-Coupled | CR vs CL vs CLR vs CLRT

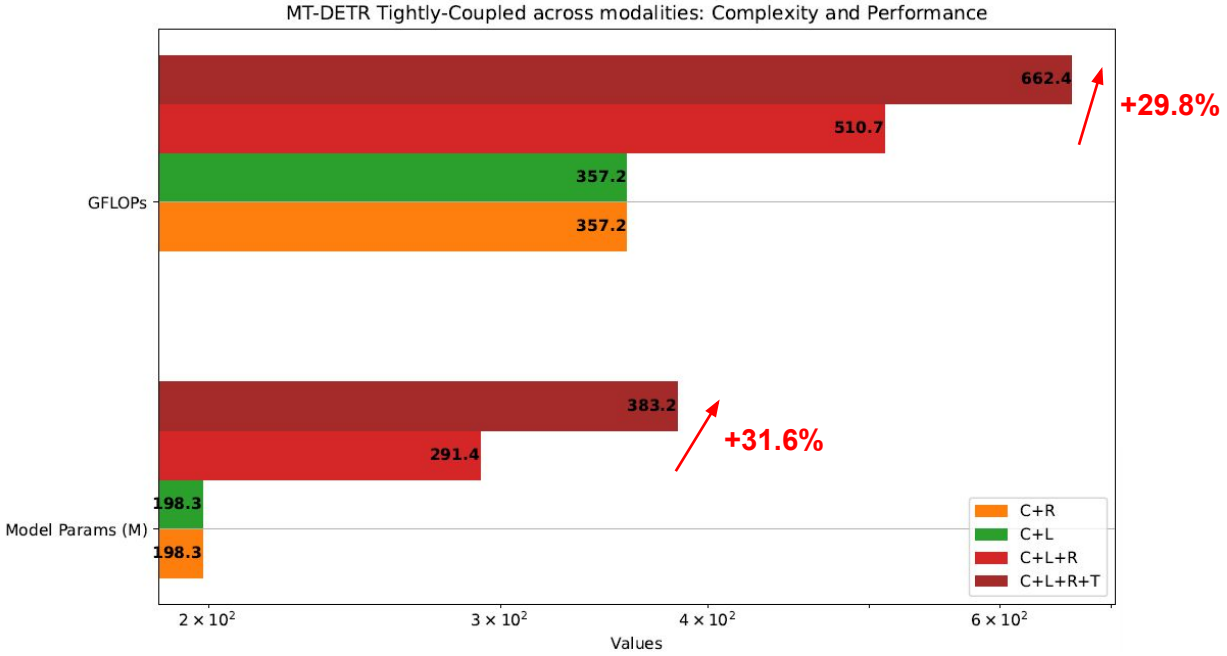


Figure 48: Performance of the MT-DETR across the Early, Middle, and Tightly-Coupled Fusion model with C+L+R in terms of AP across sensor mod. under diff. weather cond. Note: Tightly- Coupled data point values are shown above the line and rest are below the line.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

Results – Exp. 9/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
9	MT-DETR	-	x	x	✓	✓	✓	✓	Time

Middle vs Tightly-Coupled | CLR

Figure 49: A few samples to showcase the detection results of MT-DETR. Where the 1st column is ground truth, 2nd is Middle Fusion, and 3rd is Tightly-Coupled Fusion, both are using the same C+L+R (Rows show the extremely challenging weather conditions).

Results – Exp. 9/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
9	MT-DETR	-	x	x	✓	✓	✓	✓	Time

Middle vs Tightly-Coupled | CLR

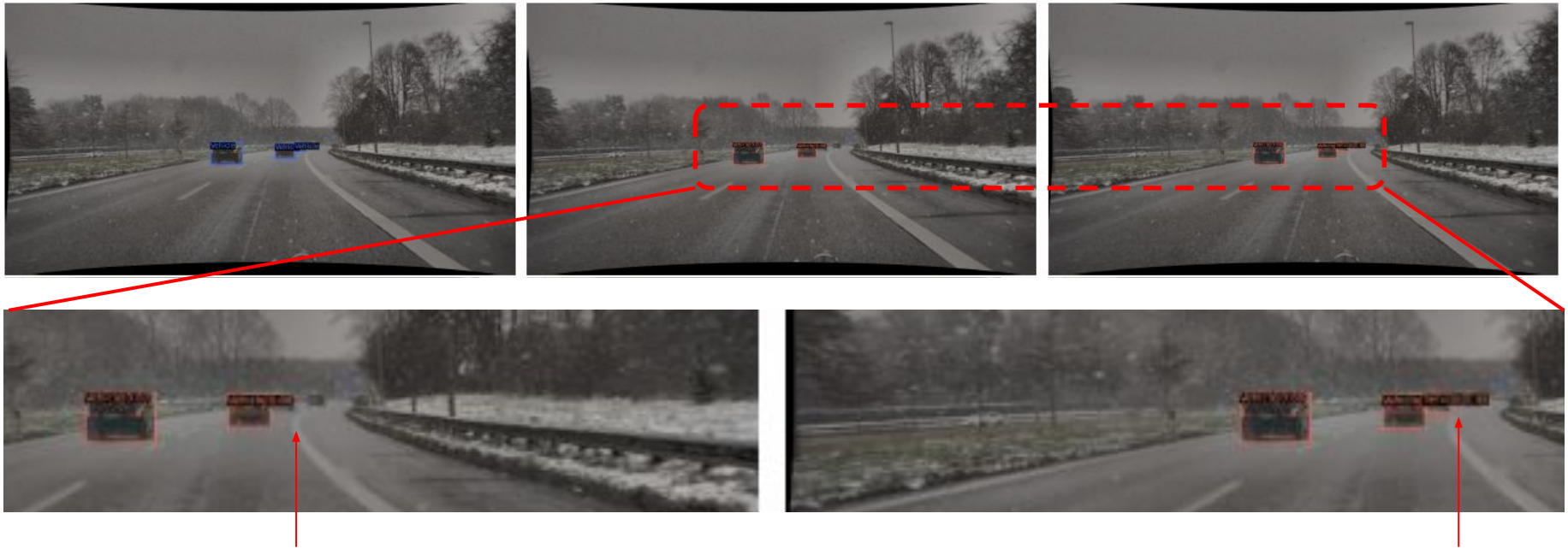


Figure 50: A few samples to showcase the detection results of MT-DETR. Where the 1st column is ground truth, 2nd is Middle Fusion, and 3rd is Tightly-Coupled Fusion, both are using the same C+L+R (Rows show the extremely challenging weather conditions).

Results – Exp. 9/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
9	MT-DETR	-	x	x	✓	✓	✓	✓	Time

Middle vs Tightly-Coupled | CLR

Figure 51: A few samples to showcase the detection results of MT-DETR. Where the 1st column is ground truth, 2nd is Middle Fusion, and 3rd is Tightly-Coupled Fusion, both are using the same C+L+R (Rows show the extremely challenging weather conditions).

Results – Exp. 9/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
9	MT-DETR	-	x	x	✓	✓	✓	✓	Time

Middle vs Tightly-Coupled | CLR

Figure 52: A few samples to showcase the detection results of MT-DETR. Where the 1st column is ground truth, 2nd is Middle Fusion, and 3rd is Tightly-Coupled Fusion, both are using the same C+L+R (Rows show the extremely challenging weather conditions).

Results – Exp. 9/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
9	MT-DETR	-	x	x	✓	✓	✓	✓	Time

Middle vs Tightly-Coupled | CLR

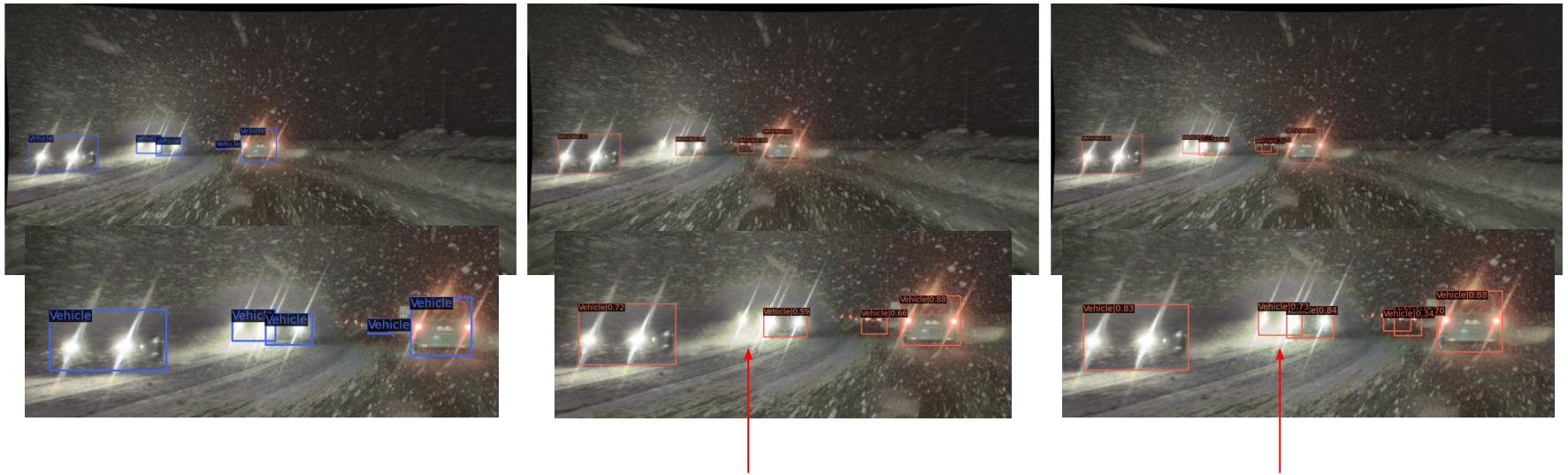


Figure 53: A few samples to showcase the detection results of MT-DETR. Where the 1st column is ground truth, 2nd is Middle Fusion, and 3rd is Tightly-Coupled Fusion, both are using the same C+L+R (Rows show the extremely challenging weather conditions).

Results – Exp. 9/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
9	MT-DETR	-	x	x	✓	✓	✓	✓	Time

Middle vs Tightly-Coupled | CLR

Figure 54: A few samples to showcase the detection results of MT-DETR. Where the 1st column is ground truth, 2nd is Middle Fusion, and 3rd is Tightly-Coupled Fusion, both are using the same C+L+R (Rows show the extremely challenging weather conditions).

Results – Exp. 10/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
10	HRFuser	MT-DETR	x	x	✓	✓	x	✓	x

Tightly-Coupled | CR | HRFuser vs MT-DETR

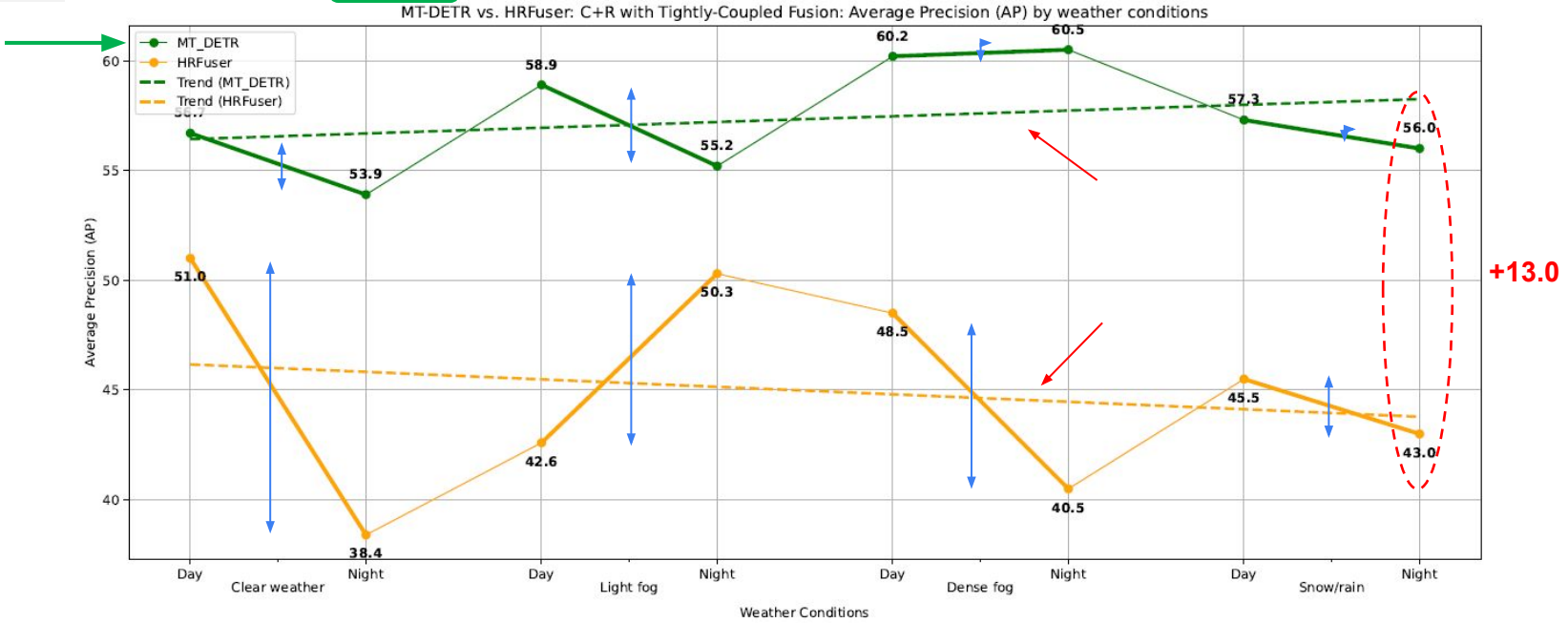


Figure 55: Performance of the MT-DETR vs. HRFuser with Tightly-Coupled Fusion on C+R in terms of AP under different weather conditions. Note: MT-DETR data point values are shown above the line

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Results – Exp. 11/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
11	HRFuser	MT-DETR	x	x	✓	✓	✓	x	x

Tightly-Coupled | CL | HRFuser vs MT-DETR

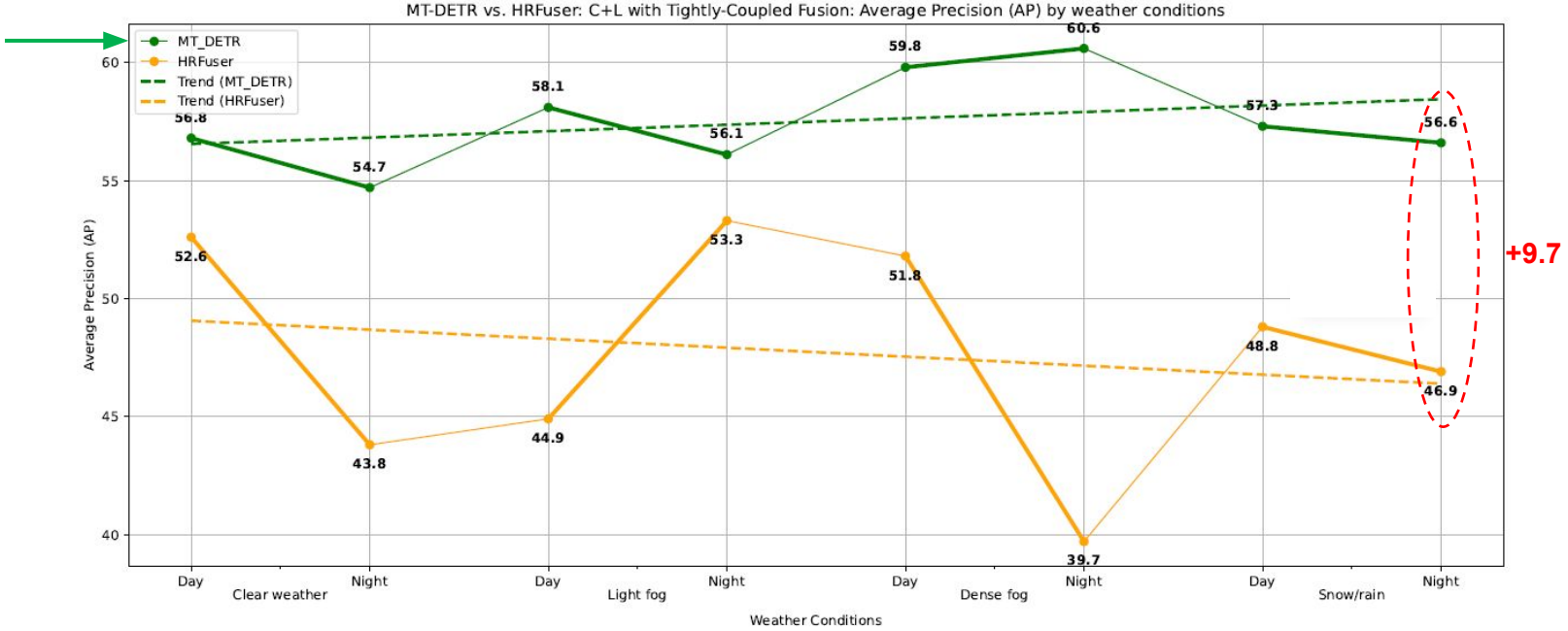


Figure 56: Performance of the MT-DETR vs. HRFuser with Tightly-Coupled Fusion on C+L in terms of AP under different weather conditions. Note: MT-DETR data point values are shown above the line.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Results – Exp. 12/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
12	HRFuser	MT-DETR	x	x	✓	✓	✓	✓	x

Tightly-Coupled | CLR | HRFuser vs MT-DETR

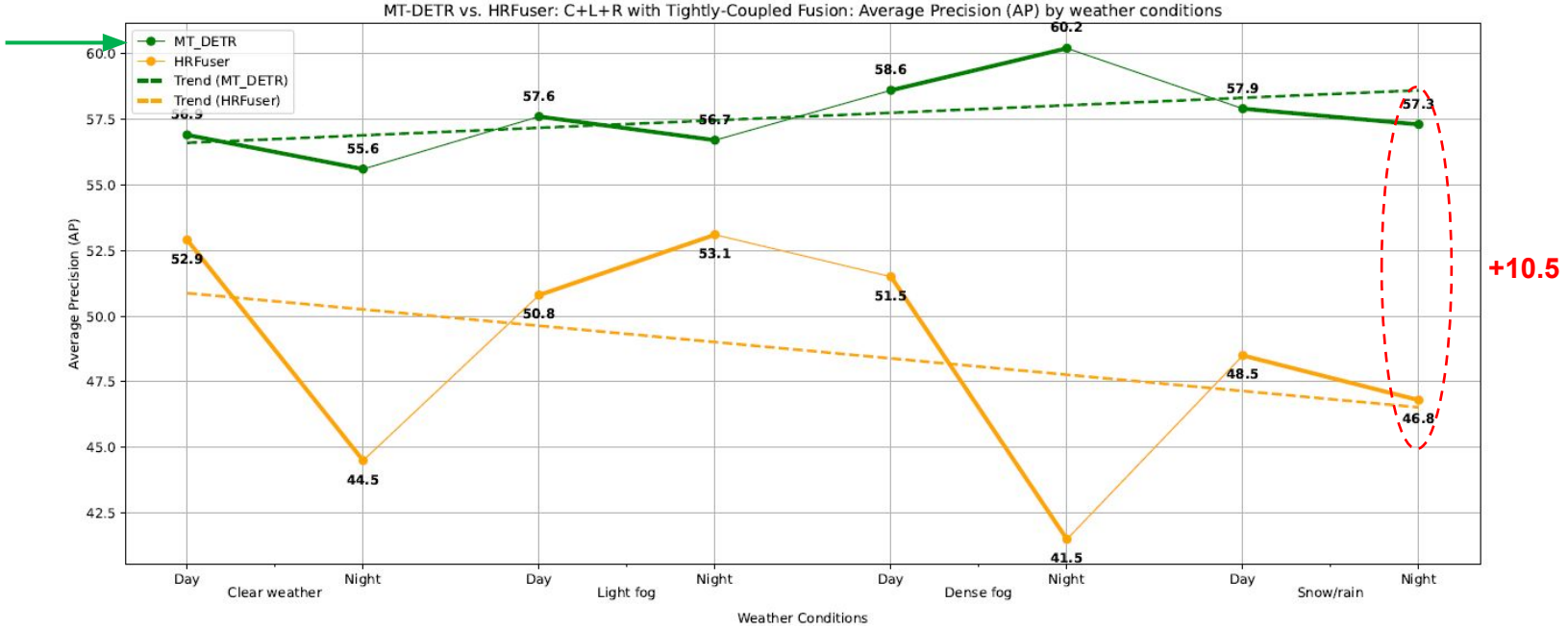


Figure 57: Performance of the MT-DETR vs. HRFuser with Tightly-Coupled Fusion on C+L+R in terms of AP under different weather conditions. Note: MT-DETR data point values are shown above the line.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Results – Exp. 13/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
13	HRFuser	MT-DETR	x	x	✓	✓	✓	✓	G,T

Tightly-Coupled | CLRG vs CLRT | HRFuser vs MT-DETR

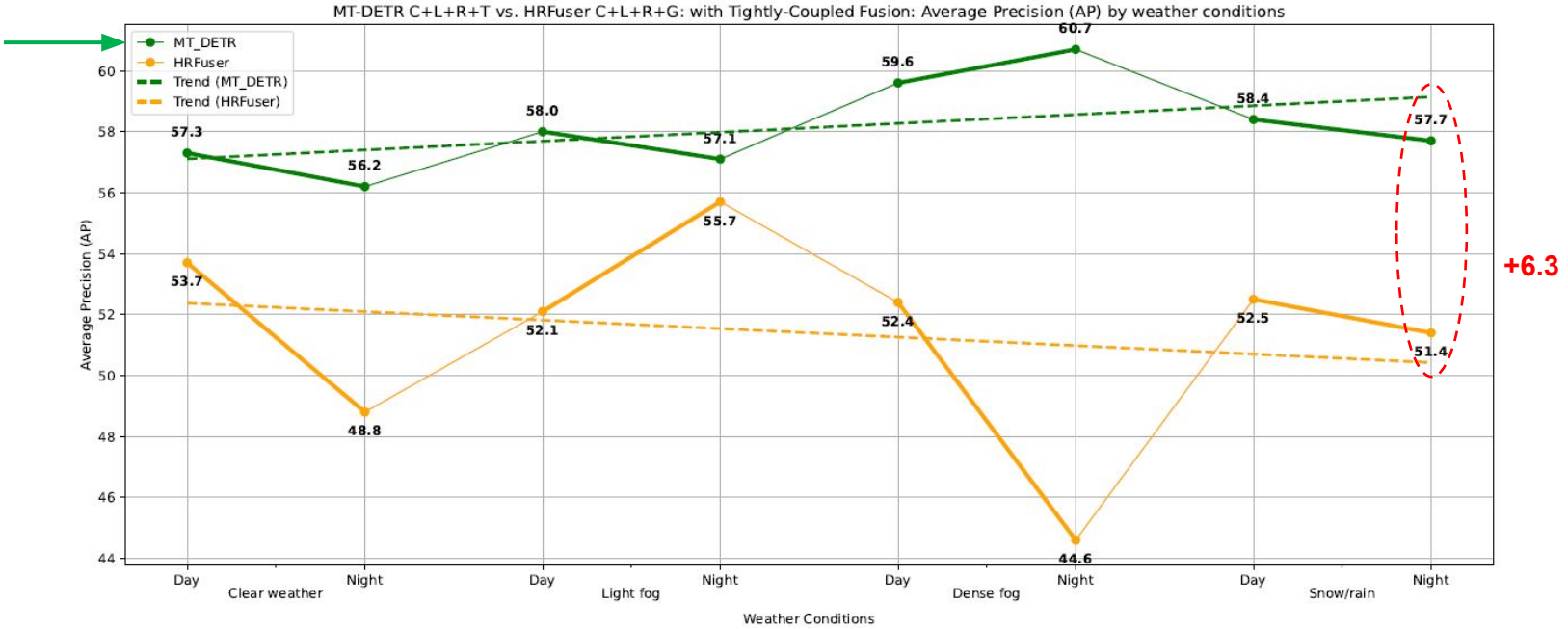


Figure 58: Performance of the MT-DETR vs. HRFuser with Tightly-Coupled Fusion on C+L+R + additional input(Time vs. Gated) in terms of AP under different weather conditions. Note: MT-DETR data point values are shown above the line.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Results – Exp. 13/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
13	HRFuser	MT-DETR	x	x	✓	✓	✓	✓	G,T

Tightly-Coupled | CLRG vs CLRT | HRFuser vs MT-DETR

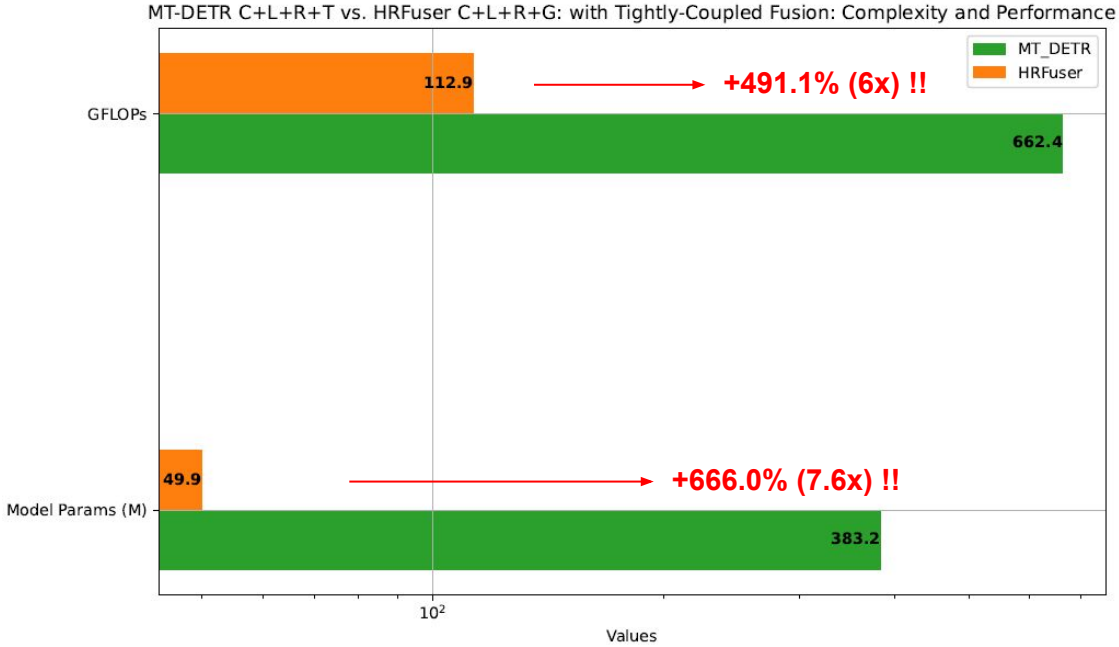


Figure 59: MT-DETR vs. HRFuser: Tightly-Coupled Fusion model's complexity on C+L+R + additional input(Time vs. Gated) displayed through model parameters (in M), and GFLOPs.

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

Results – Exp. 12/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
12	HRFuser	MT-DETR	x	x	✓	✓	✓	✓	x

Tightly-Coupled | CLR | HRFuser vs MT-DETR

Figure 60: Detection results of HRFuser vs. MT-DETR. Both with Tightly-Coupled Fusion and on the C+L+R. Where the 1st column is ground truth, 2nd is HRFuser, and 3rd is MT-DETR (Rows show the extremely challenging weather conditions).

Results – Exp. 12/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
12	HRFuser	MT-DETR	x	x	✓	✓	✓	✓	x

Tightly-Coupled | CLR | HRFuser vs MT-DETR

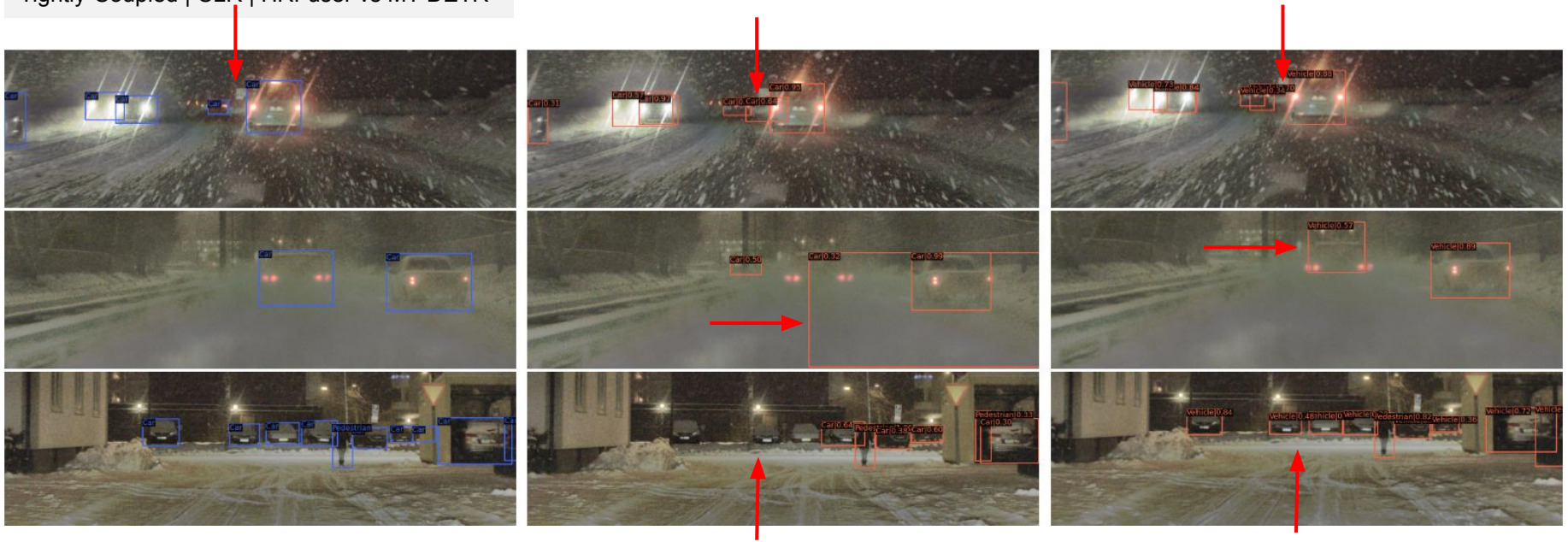


Figure 61: Detection results of HRFuser vs. MT-DETR. Both with Tightly-Coupled Fusion and on the C+L+R. Where the 1st column is ground truth, 2nd is HRFuser, and 3rd is MT-DETR (Rows show the extremely challenging weather conditions).

Results – Exp. 12/13

Sr.	1	2	Early	Middle	Tightly-c.	C	L	R	Extra
12	HRFuser	MT-DETR	x	x	✓	✓	✓	✓	x

Tightly-Coupled | CLR | HRFuser vs MT-DETR

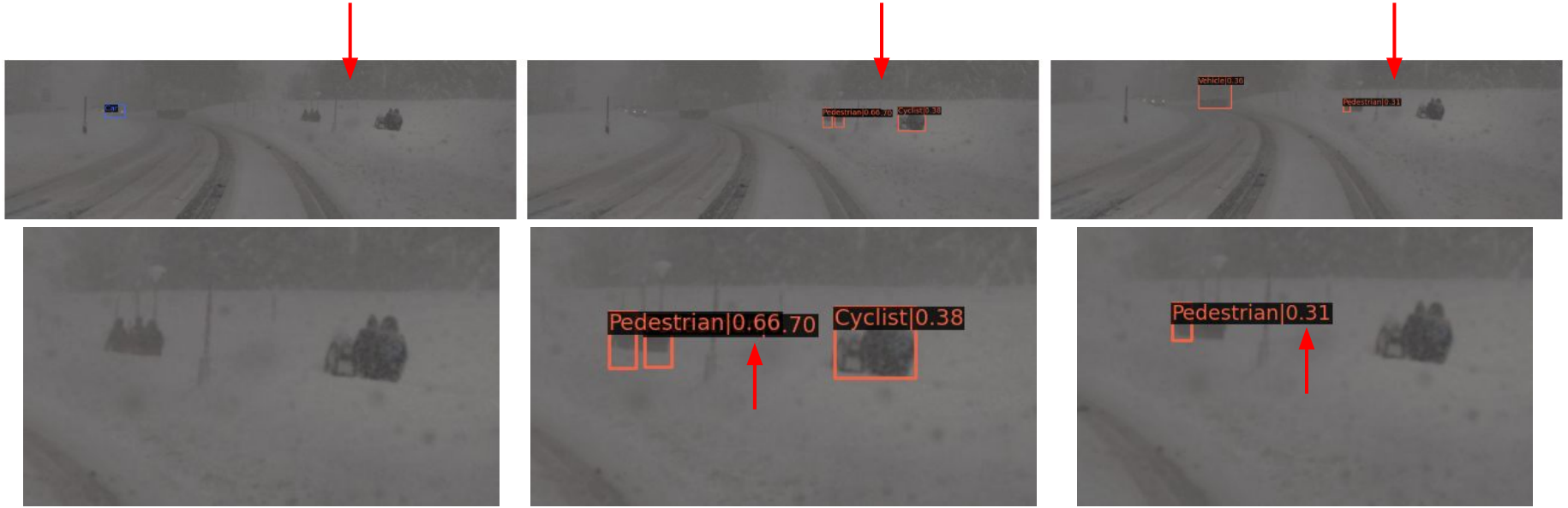
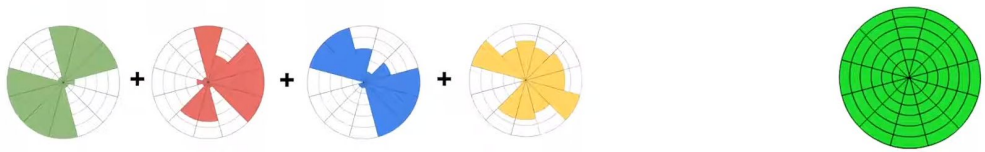


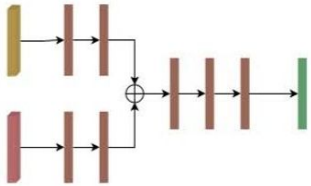
Figure 62: Detection results of HRFuser vs. MT-DETR. Both with Tightly-Coupled Fusion and on the C+L+R. Where the 1st column is ground truth, 2nd is HRFuser, and 3rd is MT-DETR (Rows show the extremely challenging weather conditions).

Conclusion

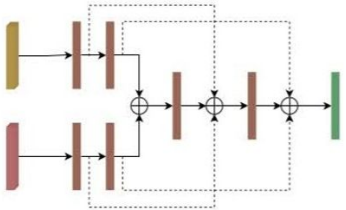
Single Sensor (Unimodal) < Complementary Sensors (Multimodal)



Middle fusion < Tightly-coupled fusion



c) Middle Fusion



d) Tightly-Coupled Fusion

Future Work

- Integration of BEV Fusion and 360-Degree View

Figure 63: BEVFusion [28]

[28] Liang, Tingting, et al. "Bevfusion: A simple and robust lidar-camera fusion framework." Advances in Neural Information Processing Systems 35 (2022): 10421-10434.

Future Work

- Integration of BEV fusion and 360-Degree view
- Point cloud and image fusion
- Utilizing 4D Radar data
- Exploring fusion architecture for advanced perception tasks

References

[8] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, “A deep learning-based radar and camera sensor fusion architecture for object detection,” in Sensor Data Fusion: Trends, Solutions, Applications (SDF). IEEE, 2019, pp. 1–7.

[9] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei, “Spatial attention fusion for obstacle detection using mmwave radar and vision sensor,” Sensors, vol. 20, no. 4, p. 956, 2020.

[10] Yadav, Ritu, Axel Vierling, and Karsten Berns. "Radar+ RGB fusion for robust object detection in autonomous vehicle." 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020.

[11 Z. Liu, Y. Cai, H. Wang, L. Chen, H. Gao, Y. Jia, and Y. Li, “Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions,” IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 7, pp. 6640–6653, 2021.

[12] Wang, Yizhou, et al. "RODNet: A real-time radar object detection network cross-supervised by camera-radar fused object 3D localization." IEEE Journal of Selected Topics in Signal Processing 15.4 (2021): 954-967.

[13] Danapal, Gokulesh, et al. "Attention Empowered Feature-level Radar-Camera Fusion for Object Detection." 2022 Sensor Data Fusion: Trends, Solutions, Applications (SDF). IEEE, 2022.

[14] P. Radecki, M. Campbell, and K. Matzen, “All weather perception: Joint data association, tracking, and classification for autonomous ground vehicles,” arXiv preprint arXiv:1605.02196, 2016.

[15] N. A. Rawashdeh, J. P. Bos, and N. J. Abu-Alrub, “Drivable path detection using cnn sensor fusion for autonomous driving in the snow,” in Autonomous Systems: Sensors, Processing, and Security for Vehicles and Infrastructure 2021, vol. 11748. SPIE, 2021, pp. 36–45.

References

[16] Chu, Shih-Yun, and Ming-Sui Lee. "MT-DETR: Robust End-to-end Multimodal Detection with Confidence Fusion." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

[17] Broedermann, Tim, et al. "HRFuser: A multi-resolution sensor fusion architecture for 2D object detection." arXiv preprint arXiv:2206.15157 (2022).

[18] “Flir. fused aeb with thermal can save lives.” [Online]. Available: https://www.flir.com/globalassets/industrial/oem/adas/flir-thermal-aeb-white-paper---final-v1.pdf.

[19] Yang, Bin, et al. "Radarnet: Exploiting radar for robust perception of dynamic objects." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. Springer International Publishing, 2020.

[20] Qian, Kun, et al. "Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[21] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[22] Barnes, Dan, et al. "The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset." 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020.

[23] Yan, Zhi, et al. "EU long-term dataset with multiple sensors for autonomous driving." 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020.

[24] Sheeny, Marcel, et al. "RADIATE: A radar dataset for automotive perception in bad weather." 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021.

References

[25] Paek, Dong-Hee, Seung-Hyun Kong, and Kevin Tirta Wijaya. "K-radar: 4d radar object detection dataset and benchmark for autonomous driving in various weather conditions." arXiv preprint arXiv:2206.08171 (2022).

[26] Burnett, Keenan, et al. "Boreas: A multi-season autonomous driving dataset." The International Journal of Robotics Research 42.1-2 (2023): 33-42.

[27] Matuszka, Tamás, et al. "aiMotive Dataset: A Multimodal Dataset for Robust Autonomous Driving with Long-Range Perception." arXiv preprint arXiv:2211.09445 (2022).

[28] Liang, Tingting, et al. "Bevfusion: A simple and robust lidar-camera fusion framework." Advances in Neural Information Processing Systems 35 (2022): 10421-10434.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Doll´ar, “Focal loss for dense object detection,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[30] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, “Unitbox: An advanced object detection network,” in Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 516–520.

[31] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 658–666.

THANK YOU