HRFuser: A Multi-resolution Sensor Fusion Architecture for 2D Object Detection – Supplementary Material –

Tim Brödermann[1], Christos Sakaridis[1], Dengxin Dai[2] and Luc Van Gool[1,3]

I. Additional Architectural Details

All branches of HRFuser start with a CNN reducing the resolution by a factor of four, followed by four stages consisting of multiple identical blocks. For all branches, we use basic bottleneck blocks to build the first stage [1] and transformer blocks to build all subsequent stages and streams [2]. A transformer block consists of a local-window self-attention on 7×7 windows followed by an feed-forward network with 3×3 depth-wise convolution and an expansion ratio of 4. The additional parameters of the transformer blocks for different versions of HRFuser are displayed in Tab. VII, where D_s and H_s apply to all blocks and MWCA modules within a given stream.

TABLE VII

Parameters of the tiny (T), small (S), and base (B) versions of HRFuser. D_s denotes the number of channels and H_s the number of heads, with $s \in \{1, \ldots, 4\}$ denoting the corresponding stream. For the camera branch α, the values are displayed as: (D_1, D_2, D_3, D_4) and (H_1, H_2, H_3, H_4). The secondary branches β only have one stream: D_1 and H_1.

Model	Branch	#channels (D_s)	#heads (H_s)
HRFuser-T	α	$(18, 36, 72, 144)$	$(1, 2, 4, 8)$
	β	18	1
HRFuser-S	α	$(32, 64, 128, 256)$	$(1, 2, 4, 8)$
	β	32	1
HRFuser-B	α	$(78, 156, 312, 624)$	$(2, 4, 8, 16)$
	β	78	2

II. Overview of Characteristics of Different Sensors

We provide a more detailed overview of the characteristics of each sensor we use in our experiments:

1) **Camera:** Very high resolution and rich texture, but poor readings in low illumination and fog and no direct geometric information.
2) **Lidar:** Fair resolution, explicit range information, independent from external illumination, degradation when the optical medium is not clear (fog, rain, snow).
3) **Radar:** Robust to adverse weather and illumination, velocity information, low resolution, noisy.
4) **Gated camera:** High resolution, robust to adverse weather and illumination, still not widely adopted.

III. Additional Details on the Experimental Setup

For both datasets (DENSE and nuScenes), we use AdamW with a base learning rate of 0.001, weight decay of 0.01, and betas of 0.9 and 0.999. We apply a learning rate warm-up for 500 iterations with a ratio of 0.001.

DENSE [3]. The dataset provides 1920×1024 camera images, lidar and radar points, and 1280×720 gated camera images, captured under a variety of normal and adverse weather conditions. We process the inputs in the same way as [3]. The camera is cropped to a 1248×360 window around the center of the gated camera. The image from the gated camera is transformed into the image plane of the camera using a homography mapping as in [3]. We also crop the annotated 2D bounding boxes to the aforementioned 1248×360 window, discarding boxes for which more than 90% of the original box area lies outside the crop. The gated camera is cropped to the same window. The strongest lidar return and radar are projected onto the image plane. As the radar cross-section (RCS) data are not publicly available, we use only 2 radar channels: depth and velocity over ground. However, RCS data were used for training the method of [3], so comparing [3] to our method is actually unfair to our method. Example inputs are displayed in Fig. 7. We train on 3 classes defined as shown in Tab. VIII, and evaluate only on the car class, using the KITTI evaluation framework [4]. We run the training of HRFuser-T on 4 Nvidia Titan RTX GPUs with a batch size of 12.

nuScenes [5]. Similar to [6], we resize the recorded 1600×900 images to 640×360 and project the radar points as 3m-high pillars onto the image plane. This creates a 640×360 projected radar image with 3 channels: range, RCS and velocity over ground. Compared to [6], we do not accumulate radar data across time or filter them in any way. The lidar points are projected onto the image plane, yielding a 640×360 image with 3 channels: range, intensity and height. Example inputs are displayed in Fig. 8. All input channels are normalized over the entire dataset. We run the training of HRFuser-T on 4 Nvidia RTX 2080 TI GPUs with a batch size of 12. We follow the mmdet3d [7] framework and use a set of 10 classes for training and evaluation, which

[1]Computer Vision Lab, ETH Zurich, 8092 Zurich, Switzerland {timbr,csakarid,vangool}@vision.ee.ethz.ch
[2]VAS, MPI for Informatics, 66123 Saarbrücken, Germany ddai@mpi-inf.mpg.de
[2]ESAT, KU Leuven, 3001 Leuven, Belgium

TABLE VIII

MAPPING FROM ORIGINAL DENSE CLASSES TO THE SET OF CLASSES WE USE FOR TRAINING.

DENSE Class	Mapped Class
PassengerCar	Car
Pedestrian	Pedestrian
RidableVehicle	Cyclist
LargeVehicle	DontCare
Vehicle	DontCare
DontCare	DontCare

TABLE IX

MAPPING FROM ORIGINAL NUSCENES CLASSES TO OUR DEFAULT SET OF TRAINING AND EVALUATION CLASSES [7].

NuScenes Class	Mapped Class
vehicle.car	car
vehicle.truck	truck
vehicle.trailer	trailer
vehicle.bus.bendy	bus
vehicle.bus.rigid	bus
vehicle.construction	construction_vehicle
vehicle.bicycle	bicycle
vehicle.motorcycle	motorcycle
human.pedestrian.child	pedestrian
human.pedestrian.adult	pedestrian
human.pedestrian.construction_worker	pedestrian
human.pedestrian.police_officer	pedestrian
movable_object.trafficcone	traffic_cone
movable_object.barrier	barrier

are defined based on the original nuScenes classes as shown in Tab. IX.

Clarification on nuScenes classes used in experiments and class-wise performance. As we mention in the main paper, "unless otherwise stated" we indeed use the set of 10 nuScenes classes in our experiments. The only case where we use a reduced set of 6 classes is in Tab. I, in which this choice of classes was necessary for comparability to the method in [8], which only reports results on these 6 classes. This different choice of classes has been explicitly stated in the caption of Tab. I.

IV. ADDITIONAL ABLATIONS

Choice of primary modality. As mentioned in Sec. IV-C, we investigate the effect of different sensors as a primary modality and provide a comparison on DENSE in Tab. X. While we have selected the camera as the primary modality for our HRFuser architecture, in this experiment we set each available sensor besides the camera, i.e., lidar, radar, and gated camera, as the primary modality. The RGB camera becomes thereby a secondary modality and is treated as the others. We observe that having either the camera or the gated camera as the primary sensor generally attains higher performance than having lidar and radar as the primary sensor, even though the respective difference is slight. Our intuition for this finding is that the high spatial resolution of the camera and the gated camera makes them better choices for serving as the primary modality, as the primary modality is used in our cross-attention block to compute the queries and thus to determine which regions of the other modalities to attend to, a function which can be carried out with higher spatial accuracy in case high-resolution readings are available from the primary modality. However, since fusion across all modalities starts at an early stage in the network, HRFuser can learn meaningful features even when using a sparse modality such as radar as the primary modality, and in any case, HRFuser is fairly robust with respect to which modality serves as the primary one.

PVTv2 adaptations. We create the alternative attention mechanisms PVTv2-CA and PVTv2-Li-CA, which are presented in Tab. V of the main paper, by adapting the state-of-the-art transformer PVTv2 [9]. We adapt its spatial reduction attention module for cross-attention by entering the query from our primary branch and the key and value from our secondary branch, and pass this into their proposed convolutional feed-forward module employing depth-wise convolution. In contrast to PVTv2, we do not incorporate the overlapping patch embedding, in order to keep the spacial dimensions of the feature maps unchanged. No pre-training is applied when training PVTv2, which is also the case for all other presented methods.

Parallel cross-attention skip connection. To examine the benefit of the skip connections for the secondary modalities which are involved in our parallel CA block as shown in Fig. 4 of the main paper, we remove these skip connections (blue and orange in Fig. 4) and observe in Tab. XI a drop of 0.6% in AP relative to our default MWCA. This finding indicates that skip connections from all modalities are beneficial for cross-attention, as they allow the network to attend to details without having to learn the identity function.

Amount of training data. We investigate whether the performance gain of HRFuser is due to the increased number of "inputs" or due to general network strength and better learned features. We trained HRFuser-T with two modalities (lidar + camera) on a subset of nuScenes containing half the training data. As seen in the Tab. XII, this model substantially outperforms the camera-only model trained on the complete training set, recovering 90% of the performance gain of the fusion model that sees the complete training set. Thus, even with the same quantity of "inputs" as the camera-only model trained on the full nuScenes training set, HRFuser delivers a significant performance improvement.

Note that the volume of information is not the same across modalities. In nuScenes, an average of only 0.71% of the pixels per radar image and 1.61% per lidar image have a measurement. Thus, adding a second modality does not double the volume of information but only increases it slightly, which means that the comparison of a camera-only model to a fusion model is reasonably fair.

V. CLASS-WISE PERFORMANCE ON NUSCENES

We report in Tab. XIII the average precision of our HRFuser-T versus the baseline camera-only HRFormer-T

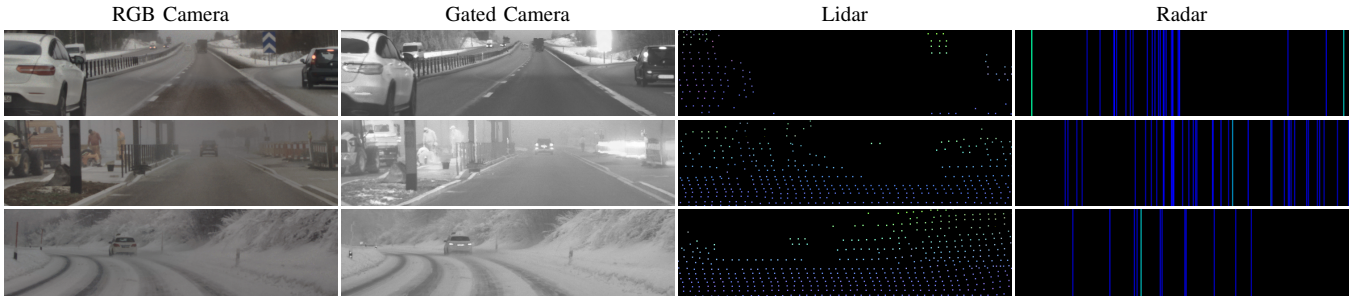

Fig. 7. Example inputs to HRFuser from DENSE. From left to right: RGB image, warped gated camera image, projected lidar points, projected radar points. The radar and lidar projections are highlighted and enlarged for better visualization. Best viewed on a screen at full zoom.

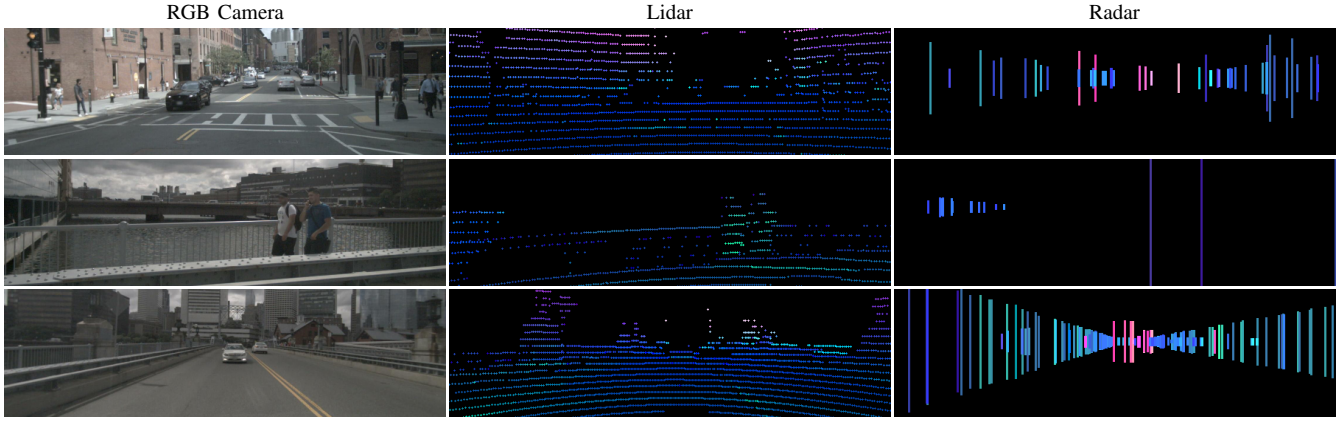

Fig. 8. Example inputs to HRFuser from nuScenes. From left to right: RGB image, projected lidar points, projected radar points. The radar and lidar projections are highlighted and enlarged for better visualization. Best viewed on a screen at full zoom.

TABLE X

ABLATION ON DENSE ON THE SELECTION OF DIFFERENT PRIMARY MODALITIES. RESULTS ARE IN AP.

Primary Sensor	clear			light fog			dense fog			snow/rain		
	easy	mod.	hard	easy	mod.	hard	easy	mod.	hard	easy	mod.	hard
RGB camera (ours)	90.15	**87.10**	79.48	90.60	89.34	**86.50**	87.93	80.27	78.21	**90.05**	**85.35**	**78.09**
Lidar	90.02	86.89	79.35	90.61	**89.45**	86.28	88.64	80.60	78.59	89.86	84.98	77.44
Radar	89.99	86.96	79.47	**90.65**	89.30	80.89	88.45	80.53	72.33	89.86	85.08	77.46
Gated camera	**90.20**	86.82	**79.54**	90.58	89.33	80.92	**88.73**	**80.87**	**79.05**	90.03	85.29	77.82

TABLE XI

ADDITIONAL ABLATION OF MWCA ON NUSCENES. MWCA: OUR MWCA FUSION BLOCK, $\text{MWCA}_{\text{w/o skip}}$: OUR FUSION BLOCK WITHOUT SKIP CONNECTIONS FOR SECONDARY MODALITIES IN THE PARALLEL CA BLOCK.

Method	AP	$\text{AP}_{0.5}$	$\text{AP}_{0.75}$	AP_m	AP_l	AR
HRFuser-T ($\text{MWCA}_{\text{w/o skip}}$)	30.9	56.6	30.0	22.1	41.9	41.9
HRFuser-T (MWCA)	**31.5**	**57.4**	**31.1**	**22.7**	**42.5**	**42.3**

TABLE XII

ABLATION STUDY ON USING A SMALLER TRAINING SET. VALIDATION SET RESULTS OF HRFUSER-T TRAINED ON A SUBSET OF NUSCENES CONTAINING EITHER ALL OR HALF THE TRAINING SPLIT ARE REPORTED. C: CAMERA, L: LIDAR.

	nuScenes (%)	Modalities	AP
HRFuser-T	100	C	26.5
HRFuser-T	50	CL	30.8 (+4.3)
HRFuser-T	100	CL	31.2 (+4.7)

on each individual class on nuScenes. HRFuser consistently outperforms HRFormer across all classes by large margins.

VI. HRFUSER WITH AN HRNET-BASED BACKBONE

Tab. XIV includes *HRFuser-w18 (HRNet)*, a variant of HRFuser built upon HRNetV2-w18 [1]. In this variant, we keep the same transformer-based MWCA fusion mechanism with the same parameters as for the default, HRFormer-based HRFuser. However, the camera branch of this variant, in which our MWCA fusion blocks are inserted, resembles HRNetV2p-w18 and follows the HRNet architecture using "Basic" blocks introduced in [1]. The secondary modality branches we introduce follow analogously the design of

TABLE XIII

CLASS-WISE PERFORMANCE ON NUSCENES. RESULTS ARE IN AP.

Class	HRFormer-T	HRFuser-T
car	50.2	**53.1**
truck	29.2	**36.8**
trailer	14.9	**20.6**
bus	39.4	**48.1**
construction vehicle	7.0	**9.6**
bicycle	20.8	**26.6**
motorcycle	22.3	**28.8**
pedestrian	25.4	**30.7**
traffic cone	26.8	**28.6**
barrier	28.8	**32.0**

TABLE XIV

COMPARISON OF 2D DETECTION METHODS ON NUSCENES EVALUATED ON 6 CLASSES: CAR, TRUCK, BUS, BICYCLE, MOTORCYCLE AND PEDESTRIAN. C: CAMERA, R: RADAR, L: LIDAR.

Method	Modalities	AP	$AP_{0.5}$	$AP_{0.75}$	AP_m	AP_l	AR
HRNetV2p-w18 [1]	C	32.4	56.6	33.5	21.0	43.7	43.4
HRFuser-w18 (HRNet)	CRL	**36.7**	**63.1**	**38.1**	**24.9**	**48.6**	**47.0**

the highest-resolution branch of HRNetV2p-w18. Tab. XIV shows a 4.3% improvement in AP of our HRFuser-w18 over the camera-only HRNetV2-w18, which demonstrates the generality of the components introduced in HRFuser, as they benefit various dense prediction networks such as HRNet and HRFormer.

VII. COMPARISON TO CRF-NET USING ONLY RADAR

In Tab. XV, we compare on nuScenes the CRF-Net [6] to a version of HRFuser which only uses radar besides the camera, i.e., omitting lidar. This comparison serves in investigating whether HRFuser can leverage information from the radar better than the competing CRF-Net, which focuses explicitly on the radar modality. Indeed, HRFuser-T_{radar} yields a 4.9% improvement in AP over CRF-Net, even when the radar is the only secondary modality.

VIII. MULTIPLE MODALITIES AND OVERFITTING

Fusing multiple modalities not only allows to build more robust features but also helps against overfitting. This is demonstrated in Tab. I of the main paper, where HRFormer-T outperforms the much larger HRFormer-B by 0.5% in AP, but HRFuser-B outperforms the smaller HRFuser-T by 0.5% in AP.

IX. MWCA ATTENTION MAP ANALYSIS

To investigate what our proposed MWCA learns, we visualize some exemplary attention maps of the cross-attention fusion. As seen in Fig. 9, MWCA attends to each modality individually and in a sensor-specific fashion. The sampling pattern thereby reflects the characteristics of the sensors:

- The lidar attention map is sparse and highlights only individual points. Areas without lidar returns–such as the sky–are ignored.
- The radar attention map follows the vertical columns of the radar input, but puts additional highlight on horizontal areas corresponding to cars.
- The gated camera attention map highlights the dominant edges and structures in the image. The high resolution and density of the gated camera allows to attend to fine details.

MWCA learns continuous reasoning across windows, as demonstrated by the continuous nature of the attention maps. This is especially noticeable in the gated camera attention map, where the highlighted edges are continuous and uninterrupted by the boundaries of each local window.

X. ADDITIONAL QUALITATIVE RESULTS

We show further qualitative results on DENSE in Fig. 10 with example failure cases in Fig. 11 and on nuScenes in Fig. 12 with example failure cases in Fig. 13.

REFERENCES

[1] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *TPAMI*, 2021.

[2] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "HRFormer: High-resolution vision transformer for dense predict," in *NeurIPS*, 2021.

[3] 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," in *CVPR*, 2020.

[4] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *CVPR*, 2012.

[5] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *CVPR*, June 2020.

[6] 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 *SDF*, 2019.

[7] M. Contributors, "MMDetection3D: OpenMMLab next-generation platform for general 3D object detection," https://github.com/open-mmlab/mmdetection3d, 2020.

[8] R. Nabati and H. Qi, "Radar-camera sensor fusion for joint object detection and distance estimation in autonomous vehicles," *arXiv e-prints*, vol. abs/2009.08428, 2020.

[9] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *CVM*, 2022.

[10] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *ICRA*, 2023.

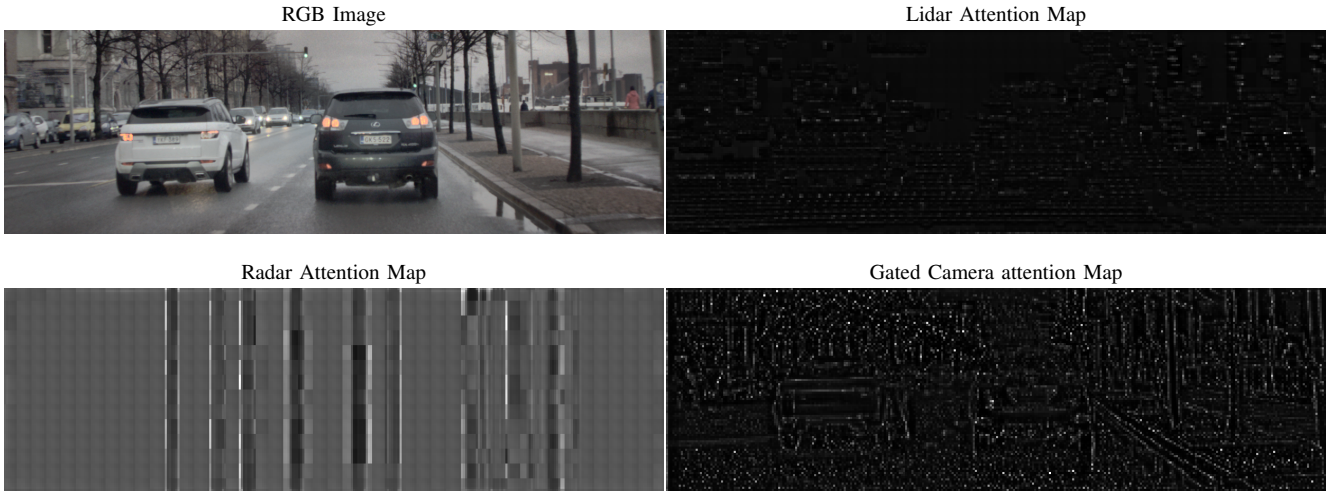

Fig. 9. Visualization of attention maps from MWCA at the first fusion stage into the highest-resolution stream. Top left: RGB image, top right: lidar attention map, bottom left: radar attention map, bottom right: gated camera attention map. Best viewed on a screen at full zoom.

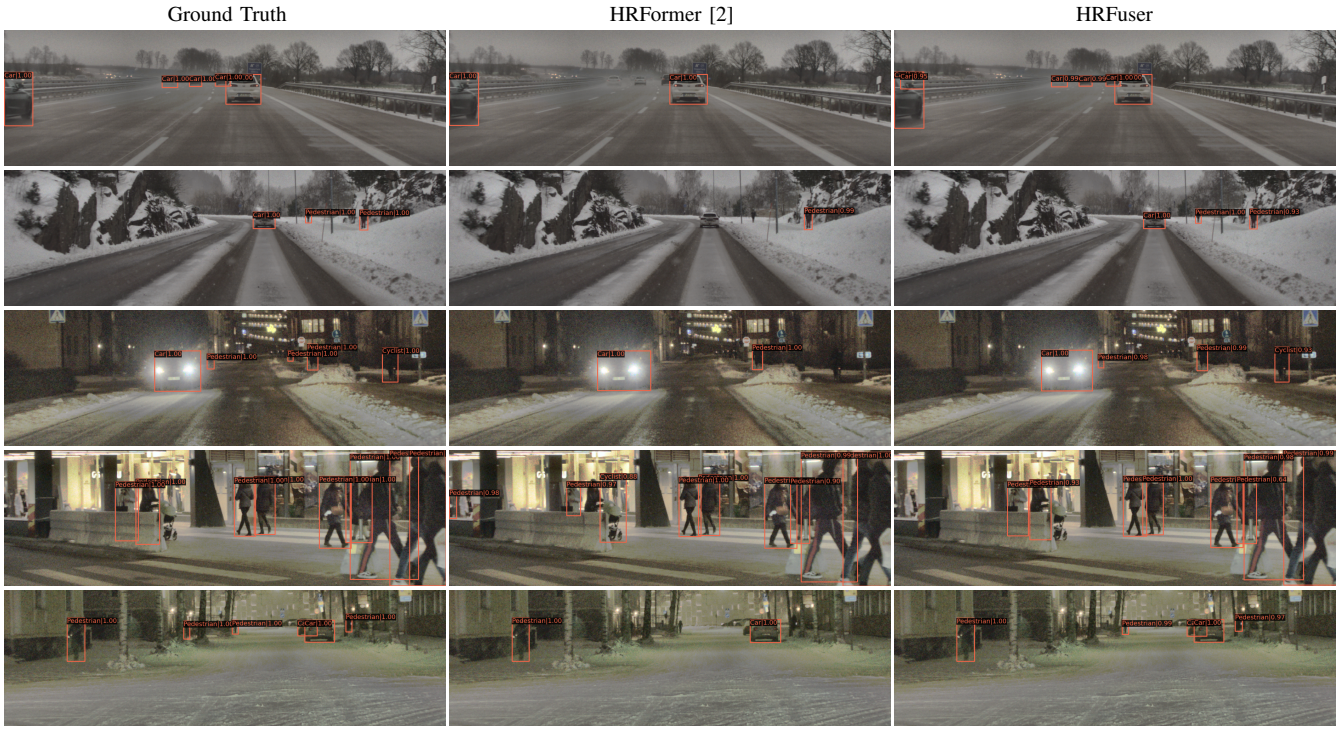

Fig. 10. Further qualitative detection results on DENSE. From left to right: image with ground-truth annotation, prediction of HRFormer, prediction of HRFuser. Best viewed on a screen at full zoom.

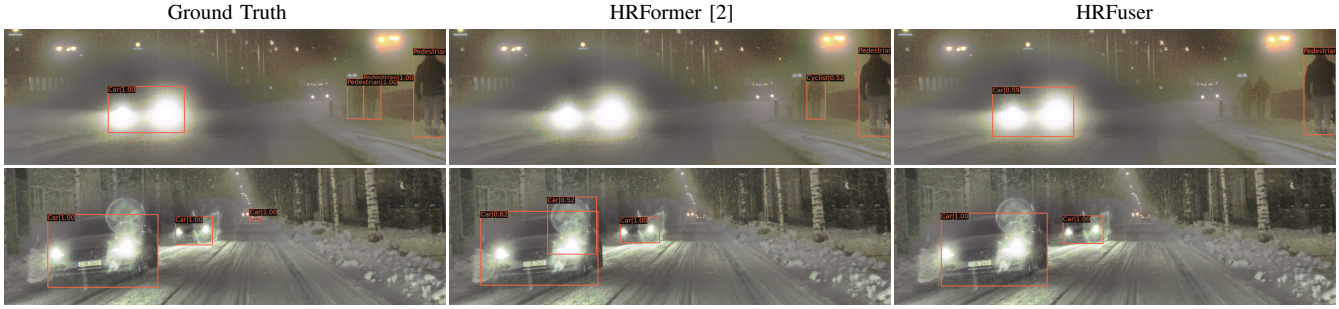

Fig. 11. Further qualitative detection results on DENSE showing failure cases of HRFuser. From left to right: image with ground-truth annotation, prediction of HRFormer, prediction of HRFuser. Best viewed on a screen at full zoom.

TABLE XV

ADDITIONAL COMPARISON ON NUSCENES OF CRF-NET [6] AGAINST A RADAR-ONLY HRFUSER, BOTH EVALUATED ON 6 CLASSES: CAR, TRUCK, BUS, BICYCLE, MOTORCYCLE AND PEDESTRIAN, USING THE SPLIT FROM [6] FOR TRAINING AND EVALUATION. C: CAMERA, R: RADAR.

Method	Modalities	AP	$AP_{0.5}$	$AP_{0.75}$	AP_m	AP_l	AR
CRF-Net [6]	CR	27.0	42.7	29.0	22.7	35.6	31.3
HRFuser-T_{radar} (HRFormer)	CR	**31.9**	**58.2**	**31.6**	**23.9**	**45.2**	**42.6**

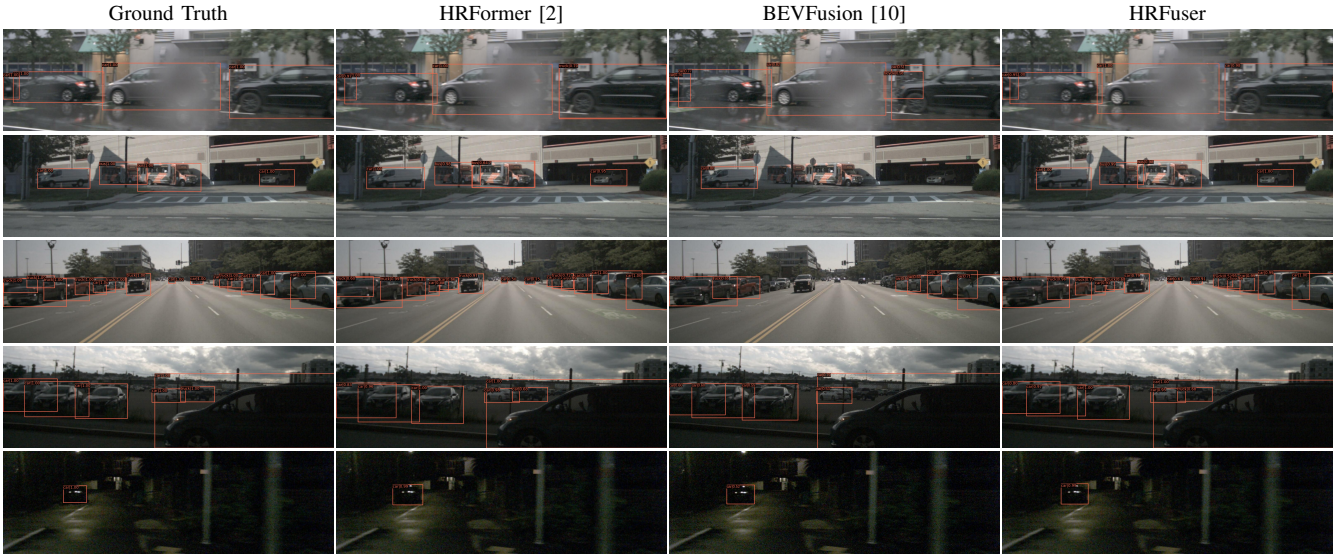

Fig. 12. Further qualitative detection results on nuScenes with 6 classes visualized: car, truck, bus, bicycle, motorcycle and pedestrian. From left to right: image with ground-truth annotation, prediction of HRFormer, 3D→2D projected predictions of BEVFusion, prediction of HRFuser. Best viewed on a screen at full zoom.

Fig. 13. Further qualitative detection results on nuScenes showing failure cases of HRFuser with 6 classes visualized: car, truck, bus, bicycle, motorcycle and pedestrian. From left to right: image with ground-truth annotation, prediction of HRFormer, 3D→2D projected predictions of BEVFusion, prediction of HRFuser. Best viewed on a screen at full zoom.