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

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

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

Abstract. Multimodal sensor fusion is an essential capability for autonomous robots, enabling object detection and decision-making in the presence of failing or uncertain inputs. While recent fusion methods excel in normal environmental conditions, these approaches fail in adverse weather, e.g., heavy fog, snow, or obstructions due to soiling. We introduce a novel multi-sensor fusion approach tailored to adverse weather conditions. In addition to fusing RGB and LiDAR sensors, which are employed in recent autonomous driving literature, our sensor fusion stack is also capable of learning from NIR gated camera and radar modalities to tackle low light and inclement weather.

We fuse multimodal sensor data through attentive, depth-based blending schemes, with learned refinement on the Bird's Eye View (BEV) plane to combine image and range features effectively. Our detections are predicted by a transformer decoder that weighs modalities based on distance and visibility. We demonstrate that our method improves the reliability of multimodal sensor fusion in autonomous vehicles under challenging weather conditions, bridging the gap between ideal conditions and real-world edge cases. Our approach improves average precision by 17.2 AP compared to the next best method for vulnerable pedestrians in long distances and challenging foggy scenes. Our project page is available here[1].

1 Introduction

Autonomous vehicles rely on multi-modal perception systems with sensors such as LiDAR [16, 34, 71, 84], camera [22, 73, 75], and radar [49], combining distinct modalities with complementary weaknesses and strengths to enable safe autonomous driving. Recent work [3, 13, 28, 47, 62, 77, 83] combines input from these diverse sensors to enhance environment perception with accurate localization and classification of objects in captured street scenes. As such, these systems benefit from the accuracy of LiDAR depth [77], the robustness of radar [28, 51], and the dense semantic information of cameras [13, 47, 62]. Although fusion is crucial for downstream classification and localization tasks, as was shown in [3, 7, 29], when

[*] These authors contributed equally to this work.

[1] https://light.princeton.edu/samfusion/

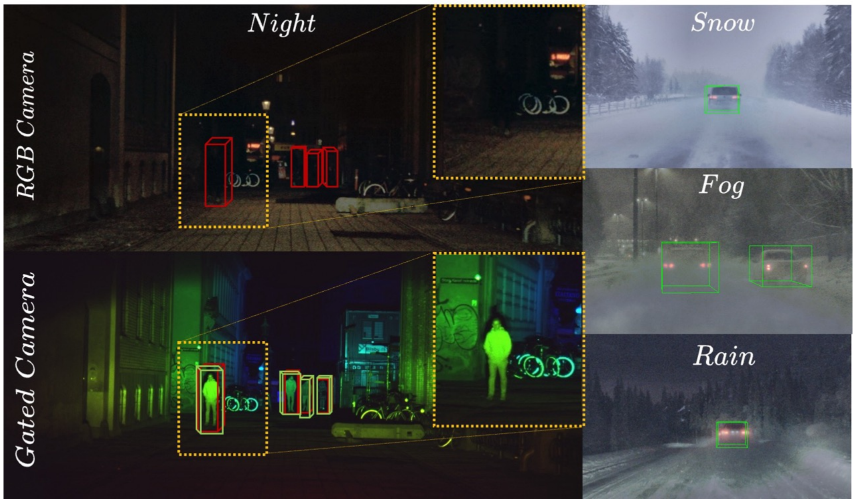


Fig. 1: In this work, we present SAMFusion, a multimodal fusion approach combining gated NIR, RGB color-imaging, LiDAR, and radar point clouds for object detection in challenging adverse weather conditions. The qualitative results above show beneficial low light detection capabilities due to the gated camera as well as example detections from our proposed approach in night-time, snowy and foggy conditions, which are achieved through attentive blending of features and multimodal querying. We depict ground truth bounding boxes in red, and predictions in green.

sensors fail, special care is required to achieve better results with fusion than with single camera networks. Examples of fusion strategies include physically-inspired entropy-driven fusion, as proposed in [3], and learned attention fusion as seen in [7]. The most effective 3D object detection methods often utilize a Bird's-Eye-View (BEV) representation, either by concatenating modality-specific feature maps [42,47] or by employing multiple attention-based modules to enhance BEV features [1,19]. However, the robustness of these techniques is typically validated only on datasets collected under favorable weather conditions [8,20], and they have not been proven effective against adverse weather-related disturbances, such as asymmetric degradation in LiDAR point clouds [4]. This vulnerability is largely attributed to the reliance on a unimodal query generator, and dependence on LiDAR-based depth projections [83], which can lead to network failures in the absence of reliable LiDAR data.

Recent advancements in gated imaging technology offer a promising alternative to conventional imaging modalities, and were explored in [5,22,31,66,67]. This work demonstrates the capability of gated cameras to actively eliminate backscatter [5], provide accurate depth [66,67], and achieve high signal-to-noise ratios (SNR) in adverse scenarios such as night-time, fog, snowy or rainy conditions, all due to their active gated scene illumination. We will therefore use gated cameras in addition to more conventional camera, LiDAR and radar data to further increase robustness.

In summary, we tackle the challenge of robust object detection in inclement weather by addressing two key problems in sensor fusion: modality projection quality and robustness against sensor distortions in adverse weather. To this end, we propose a sensor-adaptive multi-modal fusion method – SAMFusion. We introduce a novel encoder structure with a depth-guided camera-LiDAR transformation and additional early fusion between both camera modalities, incorporating distance-wise precise cross-modal projections. Additionally, we introduce a novel multi-modal, distance-based query generation approach to avoid relying solely on the LiDAR modality to generate detection proposals, as in [1, 83]. Specifically, we make the following contributions:

- We propose a novel transformer-based multi-modal sensor fusion approach, improving object detection in the presence of severe sensor degradation.
- We introduce an encoder architecture combining early camera fusion, depth-based cross-modal transformation, and adaptive blending in conjunction with learned distance-weighted multimodal decoder proposals to increase the reliability of object detection across lighting and weather conditions.
- We design a transformer decoder that aggregates multimodal information in BEV through multimodal proposal initialization.
- We validate the method on automotive adverse weather scenes [4] and improve 3D-AP, especially for the pedestrian class by more than **17.2 AP** in dense fog and **15.62 AP** in heavy snow on the most challenging distance category from 50 m-80 m relative to the state of the art.

2 Related Work

3D Object Detection. The task of 3D object detection evolved from 2D object detection, requiring the prediction of 3D-bounding boxes (bboxes) and orientations of objects [21,34,41,54,82]. Unimodal LiDAR methods, such as [34,88], have been explored to leverage the depth accuracy of the LiDAR sensor to predict 3D bboxes based on LiDAR point clouds. Point-based methods [54, 55, 59, 82] therefore generate detections from raw point cloud features. Other methods group LiDAR points into 3D voxels [14, 15] or pillars [74, 84]. Voxel and point-based methods can also be chained together, such as in [58, 60, 76], which implement additional refinement steps to improve 3D object detection performance based on region of interest pooling [23, 57]. Camera-based methods were investigated in [44–46, 73], which work in the image space itself. However, camera data has proven to be a good candidate for fusion with LiDAR, as the former can be mapped to a BEV representation, and the latter natively lives in the BEV space. Therefore, the camera representation space has since evolved from camera coordinates [46, 73] to joint multi-view setups and predicted BEV representations [26, 39], improving 3D detection accuracy.

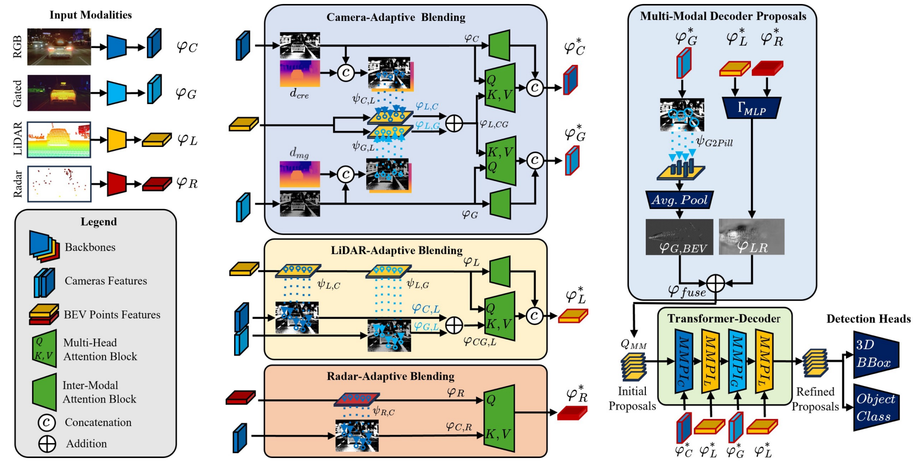
Multi-modal Sensor Fusion. While a common BEV map is not necessarily the default choice, several multi-modal sensor fusion approaches have incorporated semantic camera information to enrich individual LiDAR points, as described in [65, 69, 85]. Subsequent studies, such as [78, 86], have investigated how to extract detailed information from camera data for LiDAR point clouds, which is heavily dependent on the quality of projection and was further refined by [85]. These approaches introduced virtual 3D camera points to provide a more dense environmental context for enhancing sparse point clouds at long distances. Li et al. [38] extended this approach by integrating deformable attention [89] to create a unified representation of both modalities in the 3D voxel space.

Recently, another line of research operating in the BEV space has shown remarkable effectiveness. This approach fuses features that are aggregated in a reference frame (e.g., the LiDAR BEV perspective) and then processed by task decoders performing various perception tasks such as 3D object detection [9, 27, 30, 40, 45, 70, 79], lane estimation [37, 45, 53], tracking [25], semantic segmentation [40, 45, 47], and planning [25]. Such a framework supports multitasking and multimodal models that benefit from the additional supervision and regularization provided by these configurations. However, even the most recent BEV representation approaches [32, 47] still face challenges in projecting detailed camera features into the BEV world coordinate system and preventing error propagation in the case of sensor distortions.

Sensor Fusion in Adverse Weather. In this work, we specifically aim to tackle the degradation of individual sensors under adverse weather conditions, which drastically reduces object detection performance as shown previously in [28, 50, 63, 64, 72]. Multi-modal sensor fusion emerged as a viable approach to achieve robustness under these scenarios [2,3,7,18,50,87]. In detail, [2,13,29,43] fuse the camera modality with radar information, while [3, 7] introduce additional sensing modalities and exploit novel, physically-grounded fusion techniques. However, these only allow for the prediction of 2D object detections. Our approach projects to a common BEV plane, with attention-based feature fusion and the incorporation of dense depth to allow for more performant 3D object detection.

3 SAMFusion

In this section, we introduce the SAMFusion architecture for multimodal 3D object detection. SAMFusion leverages the complementary strengths of LiDAR, radar, RGB, and gated cameras. Gated cameras excel in foggy and low-light conditions, while radar is effective in rain and at long distances. By integrating these sensors into a depth-based feature transformation, a multi-modal query proposal network and a decoder head, SAMFusion ensures robust and reliable 3D object detection across diverse scenarios. The architecture is illustrated in Fig. 2.



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

Fig. 2: SAMFusion Architecture. 2a First, we extract features from each modality. 2b Then, we refine them, fusing modalities through attention and depth-based blending. 2c Finally, refined gated and range (LiDAR and radar) features are agglomerated in BEV, and are combined in a weighted manner that is aware of distance and weather, before being refined further and sent to detection heads to produce bounding box outputs. The gated camera and radar sensors complement the high-definition RGB camera and LiDAR to better handle poor illumination and adverse weather.

The inputs - RGB/gated camera, LiDAR, radar - are transformed into features through their respective feature extractors 2a. These features are blended in the Multi-Modal encoder 2b in an attentive fashion, and are combined with camera-specific feature maps to produce enriched features φ^* - we refer to this as “early fusion”.

Features φ^* are now passed to the multi-modal decoder proposal module 2c where they are refined with another level of fusion in the Bird’s Eye View representation to combine the image features (gated camera) and the range features (LiDAR, radar) in an adaptive, distance-weighted fashion for initial object proposals. Additionally, the enriched features φ^* are sent to the transformer decoder that refines the initial object proposals to attentively produce detection outputs. The decoder proposals include optimizations to adaptively weight distance through a learned weighting scheme that is aware of the physical properties of ranging sensors while fusing with the information-dense camera modality.

3.1 Cross-Modal Adaptive Blending

This section describes the early attention fusion schemes of individual sensor features. An illustration of the methodology is shown in Fig. 2b.

In the SAMFusion encoder, early attention fusion integrates information from different modalities. To achieve this, we first create a weighted context from the features of the primary modality, which aligns with the features of the secondary modality. This context (key) is then queried with data from the second modality (query), resulting in a rich mix of aligned features.

Our early fusion approach supports queries from both camera and LiDAR modalities, creating two parallel instances of pair-wise (query, key) attentive fusion. In "Camera-Adaptive Blending," queries from RGB and gated cameras are compared against weighted LiDAR context samples (RGB camera against Sampled LiDAR and gated camera against Sampled LiDAR). This blending accounts for objects visible in one modality but not in the other. Similarly, in "LiDAR-Adaptive Blending," LiDAR queries are scored with sampled weighted camera context features blended across RGB and gated images (LiDAR against Sampled camera).

Finally, we refine radar features in a similar fashion, where the radar proposals are scored with weighted context provided from the RGB camera.

Camera-Adaptive Blending. In this module, we use attention to score the camera features φ_C, φ_G (query) against the weighted context $\varphi_{L,CG}$ (keys, values) derived from the LiDAR modality. To generate such a context, we gather LiDAR BEV features φ_L corresponding to the camera features. We note that the LiDAR feature encoder outputs are available in the form of a BEV image. Therefore, we transform all the camera pixels (u, v) onto the LiDAR coordinate frame. In order to achieve this, we need pixel-wise depth $\mathbf{d}(u, v)$ for each camera feature coordinate. In Fig. 2b we denote the concatenation with the symbol Ⓒ that assigns the corresponding depth to each pixel.

Together with depth, we use known camera intrinsics and extrinsics (with respect to LiDAR) to lift image points into the 3D (x, y, z) LiDAR coordinate space. In our setup, we compute depth differently for RGB and gated cameras. For RGB cameras, we use stereo RGB pairs from the dataset and predict depth utilizing [35], while for gated cameras, the depth (d_{MG}) is attained from a mono-RGB method [56], which is fine-tuned on the gated camera data following [68].

The projection - $\psi_{C,L}$ for RGB camera, $\psi_{G,L}$ for gated camera, $\psi_{C;G,L}$ - is attained by lifting the pixels into a point cloud using

$$\begin{cases} z = \mathbf{d}(u, v), \\ x = (u - C_x) \times z/f_x, \\ y = (v - C_y) \times z/f_y, \end{cases} \tag{1}$$

where (fx, fy) are the horizontal and vertical focal lengths of the camera and (Cx, Cy) is the pixel location corresponding to the camera center, and then applying a change of frame of reference to bring the 3D points into the LiDAR coordinate frame.

The reprojected 3D camera points (x, y, z) are then squashed along the height coordinate y onto the LiDAR BEV grid. Further, we resolve the discretization of the LiDAR feature map $\varphi_L(x, z)$ by bilinear interpolation of the corresponding BEV coordinates. Subsequently, the found correspondences are used to enrich each 3D camera point (x, y, z) with extracted LiDAR features φ_L, which are backprojected into the camera image and paired with image features prior to scoring with attention. Through this procedure, for each RGB and gated camera pixel $\varphi_C(u,v)$ and $\varphi_G(u,v)$ we obtain corresponding LiDAR feature points $\varphi_{L,C}(u,v)$ and $\varphi_{L,G}(u,v)$.

Finally, these two independent weighted LiDAR contexts are blended together to get a composite representation $\varphi_{L,CG}$ that is aware of both camera modalities. This composition is obtained by summing up the two feature maps, where we drop the positional dependence in $\varphi_{L,C}(u,v)$ and $\varphi_{L,G}(u,v)$ for notational convenience:

$$\varphi_{L,CG} = \varphi_{L,C} \oplus \varphi_{L,G}, \tag{2}$$

where $\oplus$ is the element-wise addition operation.

The described process is introduced to integrate detailed camera-specific information into $\varphi_{L,CG}$, avoiding the case when either modality fails due to reduced visibility of the sensors in adverse lighting conditions.

Having obtained the associated LiDAR feature points to compare with, we integrate cross-modal attention to learn enriched modality-specific feature maps, including object features from the LiDAR modality that can be occluded in the camera frames due to the physical position of the sensors. We carry out an attention computation between the respective camera and LiDAR modalities $(\varphi_C, \varphi_{L,CG})$ and $(\varphi_G, \varphi_{L,CG})$ to produce the final enriched camera-specific feature maps φ_C^* and φ_G^*, to guide the decoder object proposals. We write the cross-modal attentive blending equation with LiDAR (key, value) $\varphi_{L,CG}$, abbreviating the extracted RGB and gated features φ_C, φ_G as $\varphi_{C;G}$ and the enriched maps φ_C^*, φ_G^* as $\varphi_{C;G}^*$, as

$$\varphi_{C;G}^* = \sum_{\varphi_{L,CG} \in J_s} softmax\left(\frac{\varphi_{C;G}\, \varphi_{L,CG}^T}{\sqrt{d}}\right)\varphi_{L,CG}. \tag{3}$$

The attention computation is performed over a local window J_s around the sampled point (i,j), with a window size of k and a softmax normalization factor of d, representing the dimensionality of the point cloud features.

We note that, besides the cross-modal attention mechanism, we execute intra-modal-attention in parallel on the queried modality, described by

$$\varphi_{C;G}^* = \sum_{\varphi_{C;G} \in J_s} softmax\left(\frac{\varphi_{C;G}\, \varphi_{C;G}^T}{\sqrt{d}}\right)\varphi_{C;G}. \tag{4}$$

Afterwards, $\varphi_{C;G}^*$ feature maps, cross-modal-attention and intra-modal-attention results are fused with a learned weighting scheme (independently for RGB φ_C and gated φ_G).

LiDAR-Adaptive Blending. In this module, we blend LiDAR features φ_L with a weighted context from RGB and gated camera features $\varphi_{CG,L}$ using attention, with LiDAR features serving as queries and camera features as keys and values. Unlike camera-adaptive blending, depth is inherently included in the LiDAR BEV features $\varphi_L(x_L, z_L)$. Therefore, before projecting into the camera feature map, we assign the LiDAR points (x_L, y_L, z_L) to columns at the respective feature map grid positions (x_L, z_L).

Furthermore, the 3D LiDAR features $\varphi_L(x_L, y_L, z_L)$ are mapped onto the corresponding 2D image points $(u_{C;G,L}, v_{C;G,L})$ by projection, analogous to Eq. 1, through the $\psi_{L,C;G}$ LiDAR-to-camera (RGB; gated) projection matrix. The camera features corresponding to relevant LiDAR feature coordinates $(u_{C;G,L}, v_{C;G,L})$ are acquired by sampling from the image modalities through bilinear interpolation.

Next, we blend the LiDAR-aware sampled image features from the two camera modalities

$$\varphi_{CG,L} = \varphi_{C,L} \oplus \varphi_{G,L}, \tag{5}$$

before scoring against corresponding LiDAR queries. As before, we drop the positional dependence in $\varphi_C(u_{C,L}, v_{C,L}), \varphi_G(u_{G,L}, v_{G,L})$ for notational convenience.

The enriched LiDAR feature map φ_L^* is obtained similarly to the Camera-Adaptive-Blending in Sec. 3.1, blending the output of the cross-modal attention between LiDAR queries and LiDAR aware image features (similarly to Eq. 3) to the output of the intra-modal attention over LiDAR features (as per Eq. 4).

Radar-Adaptive Blending. In the radar branch, we rely on the same principle as for the LiDAR-Adaptive Blending described in Sec. 3.1, with the only difference being that we calculate the weighted context from the RGB camera modality only and don't perform intra-modal attention due to the sparseness of radar point clouds.

3.2 Multi-Modal Decoder Proposals

SAMFusion generates initial object proposals Q_{MM} based on a multi-modal BEV feature map with an additional learned weighting scheme, prioritizing modalities based on distance and weather. The distance weighting is encoded in the BEV-based fusion of radar and LiDAR while additional weather robustness is gained by enriching the multimodal queries with the gated modality. An example is rainy weather, where LiDAR is compromised and can be enhanced by proposals from camera and radar modalities.

In particular Q_{MM} are generated from LiDAR, radar and gated camera features. An illustration of the methodology is presented in Fig. 2c.

Weighted Radar And LiDAR Feature Map Fusion. We leverage distance-dependent sensor-specific ranging characteristics and employ a weighted fusion approach to combine the enriched feature maps φ_L^* and φ_R^* into a joint feature map φ_{LR} described by

$$\varphi_{LR} = \mathit{\Gamma}_{MLP}(f(d,\sigma)\varphi_L^* + (1 - f(d,\sigma))\varphi_R^*) \tag{6}$$

where $f = \exp((-\frac{d}{2\sigma^2})^2)$, and d is the distance of each feature point from the ego veichle and σ is a learned parameter.

The learned $\mathit{\Gamma}_{MLP}$ weighs LiDAR and radar features through a gaussian mask with learned variance, which amplifies LiDAR at close range and suppresses it at longer ranges to favor radar. The range is dependent on the learned guassian variance. The resulting features φ_{LR} are thus modulated to contain LiDAR and radar, weighted by their relative importance across the ROI.

Late Gated Camera Features Fusion. To generate the final object proposal, our method encodes the initial proposals extracted from the gated camera. Due to the time-of-flight principle of the sensor, they encode distance within the captured intensity profiles. To encode detailed gated camera features φ_G^* a pillar-based conditioning approach is used to transform the camera feature map into a common BEV representation matching the distance-weighted feature map φ_{LR}. The original LiDAR coordinates are transformed according to the 3D LiDAR points into the camera representation, as described in Sec. 3.1 and are used to sample camera features φ_G^*. Then, camera features are assigned to the corresponding LiDAR pillars and the feature positions in the LiDAR BEV grid are determined through average pooling, resulting in a BEV camera feature map $\varphi_{G,BEV}$. Features $\varphi_{G,BEV}$ and φ_{LR} are fused in an additive manner to obtain a distance-encoded weighted feature map φ_{fuse} dependent on three modalities by conditioning the ranging sensor feature maps with corresponding gated camera features. Further, we apply class-dependent convolution layers onto φ_{fuse} to extract object proposal centers based on maximum intensity values and obtain the initial object proposals Q_{MM}. Q_{MM} sets the starting point for the decoder refinement process through Multi-Modal-Predictive-Interaction layers obtained from Yang et al. [83].

3.3 Training

The SAMFusion architecture, designed as a transformer network, follows the learning methodology of Carion et al. [11] and Bai et al. [1]. It first matches labels to predictions using Hungarian loss [33], then minimizes a loss composed of a weighted sum for classification (Cross-Entropy), regression, and IoU. Detailed loss formulations are provided in the supplemental material.

3.4 Implementation

We implement SAMFusion in PyTorch [52] and the open-source library MMDetection3D [17]. We initialize the camera branch with a ResNet-50 [24] backbone and pretrained Cascade Mask R-CNN [10] weights. The original RGB and gated camera images are scaled with center-based cropping to [800,400] to reduce computational cost. We define the voxels to be 0.075 m deep, 0.075 m wide and 0.2 m high. We restrict the LiDAR and radar point clouds to (0 m, 100 m) in range and to (-40 m, 40 m) in width. The height range is set to (-3 m, 1 m) and (-0.2 m, 0.4 m) for LiDAR and radar respectively. We implement four stacked transformer decoder layers, guided by RGB, gated camera, and LiDAR modalities with 200 initial multi-modal proposals. We train all models for 12 epochs in an end-to-end manner with a batch size of 4 on NVIDIA V100 GPUs. Refer to the supplemental material for hyperparameter and training settings on the SeeingThroughFog dataset [3] as well as a full latency comparison against multi-modal sensor fusion methods, proving the real-time capabilities of our approach.

Table 1: Evaluation of SAMFusions detection performance measured in AP and compared to State-of-the-Art mono- and multi-modal methods based on the car and pedestrian classes on the SeeingThoughFog [3] test set.

Average Precision for *Pedestrian* class

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

Average Precision for *Car* class

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

4 Experiments

In this section, we present experiments validating the design choices of SAMFusion. Subsection 4.1 introduces the metrics and datasets, Subsection 4.2 presents ablations of the individual contributions and Subsection 4.3 showcases comparisons against existing state-of-the-art uni- and multi-modal 3D detection methods on day, night, foggy and snowy scenarios.

4.1 Dataset And Evaluation Metrics

This section describes the evaluation of SAMFusion on the SeeingThroughFog dataset [3], consisting of 12,997 annotated samples in adverse weather conditions, covering night, fog, and snowy scenarios in Northern Europe. Following [31], we divide the dataset into 10,046 samples for training, 1,000 for validation, and 1,941 for testing. The test split is further divided into 1,046 daytime and 895 nighttime samples, with respective weather splits. Additionally, we provide results for our evaluations on the NuScenes dataset [8] in the supplemental material.

Evaluation Metrics. Object detection performance is evaluated according to the metrics specified in the KITTI evaluation framework [21], including 3D-AP and BEV-AP for the passenger car and pedestrian class. We incorporate 40 recall positions [61] for the AP calculation. To match the predictions and ground truth we apply intersection over union (IoU) [12] with an IoU of 0.2 for passenger cars and 0.1 for pedestrians. Further, we follow [81] and report results according to respective distance bins.

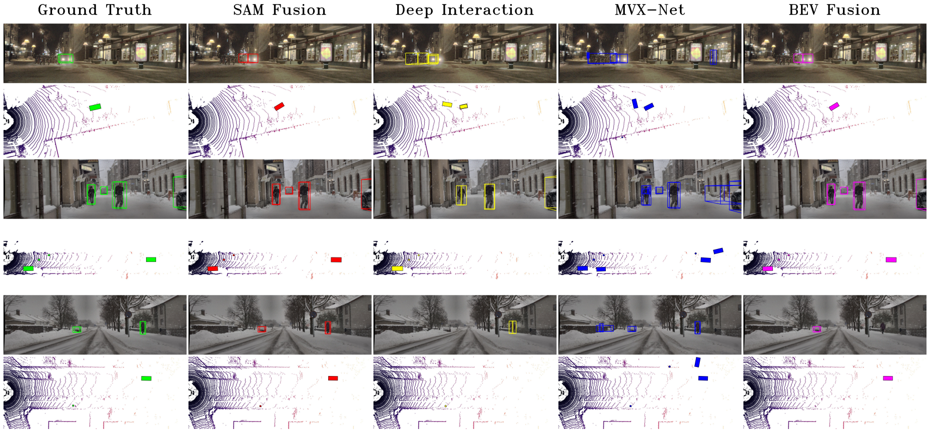


Fig. 3: Qualitative results on 3D Object detection in adverse weather compared to state-of-the-art multi-modal sensor fusion methods and ground truth (GT). While all methods perform well in the daytime setting, SAMFusion outperforms other reference methods in adverse and low light conditions (rain, snow, fog, twilight, night). In rainy and snowy settings, other methods show missing (BEVFusion) or spurious (MVXNet, DeepInteraction) detections, especially for the pedestrian class. In twilight and night, the effects are more pronounced, with missing and erroneous detections in most objects. Moreover, we see SAMFusion excel with far-away objects and pedestrian detection.

4.2 Ablation Experiments

In this subsection, we validate our methodological contributions shown in Table 2a and Table 2b.

Table 2a explores ablations with varying numbers of input modalities using the SAMFusion architecture. Configurations include single camera-LiDAR (CL), gated-LiDAR (GL), camera-LiDAR-radar (CLR), gated-LiDAR-radar (GLR), and camera-gated-LiDAR-radar (CLGR) inputs. These methods utilize queries based on LiDAR and radar data with learned distance weightings. We focus our results on the pedestrian class at extended distances, where detection is most challenging due to sparse LiDAR points. The outcomes underscore the benefits of integrating additional modalities, particularly noticeable during both day and night conditions.

Performance comparisons between single camera modalities with passive RGB and active gated imaging (GL and CL) show distinct advantages under different lighting conditions. In daylight, the inclusion of RGB color information in CL provides a performance boost of 2.85 AP-points within the 50 m to 80 m range. Conversely, at night, the superior SNR of active illumination in GL enhances detection, yielding improvements of +1.08 AP in mid-range and +3.45 AP in long-range distances. Integrating both camera technologies in the CGL configuration leverages the strengths of each, delivering enhanced performance across day and

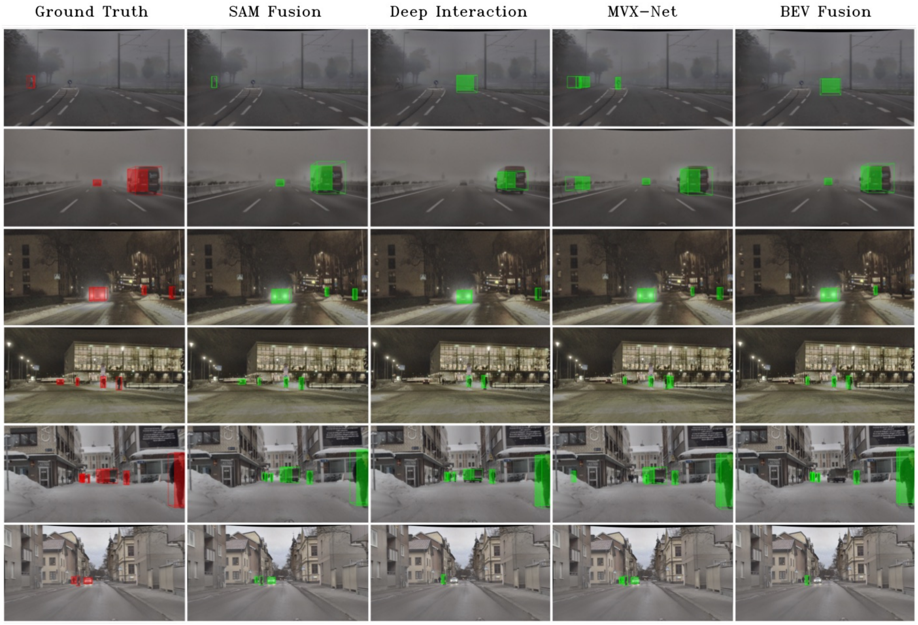


Fig. 4: We show qualitative results on different sequences (rows) of the proposed method and reference approaches (columns). On the left the ground truth is illustrated with red bounding boxes, followed by the proposed SAMFusion approach, BEVFusion [42], MVXNet [62] and DeepInteraction [83].

night settings. The addition of radar data further amplifies overall performance, although the absence of the gated camera slightly diminishes night-time efficacy.

The optimal results manifest when all four modalities (CGLR) are used, capitalizing on the unique strengths of each sensor to bolster the architecture's resilience across diverse lighting and adverse weather conditions. This configuration also benefits from leveraging proposals generated from all involved modalities.

Further, in Table 2b, we extend our validation to assess the impact of our fusion technique beyond mere modality integration. We investigate the efficacy of depth-based transformations, weighted BEV maps, and various modal proposal strategies. The incremental inclusion of these methodological enhancements correlates with notable performance improvements, indicating that simply stacking modalities is insufficient for maximizing results. For instance, incorporating multi-modal proposals elevates night-time pedestrian detection by 15.2% over solely point cloud-based proposals. Additionally, our distance-aware weighting mechanism, Γ_{MLP}, further boosts detection capabilities by up to 20.7%. Notably, proposals utilizing gated imaging data yield a larger improvement margin than those based on color data, due to their inherent distance encoding, which facilitates superior geometrical localization.

Table 2: We measure the individual method contributions on the most difficult pedestrian class. Table 2a ablates the addition of new modalities as input and in the proposal generation. We observe that adding beneficial sensor modalities improves pedestrian detection reliability, especially in low light conditions. Fusing both cameras in the adaptive blending module boosts overall detection quality of small objects due to detailed camera specific feature maps with significant information content in far distances. Table 2b ablates the proposal modality configurations and the depth-based transformations in the encoder and the learned Γ-weighting for LiDAR-radar-fusion. Object detection results are evaluated based on the 3D AP metric explicitly for the pedestrian class and the most relevant far distance from 50-80m.

(a) Ablation of Input Modality configurations.

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

(b) Ablation of SAMFusion components.

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

Table 3: Detection performance of SAMFusion measured in AP compared to multimodal methods in challenging weather conditions, evaluated on the car and pedestrian classes of weather test splits from [3]. We achieve significant performance increases shown in the last row of each Table.

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

4.3 Assessment

We compare SAMFusion against nine state-of-the-art methods, including one monocular camera 3D object detection method [6], two gated camera methods [31, 48], one stereo camera approach [36], one LiDAR approach [80], and four LiDAR-RGB fusion methods [42,62,77,83]. The results are summarized in Table 1 and further qualitative assessments are presented in Figure 3 and 4, with reported detections in both BEV and perspective view.

SAMFusion outperforms all state-of-the-art multi-modal methods in pedestrian detection under adverse weather and varying lighting conditions. Particularly in the far distance range of $50\,m$ to $80\,m$, SAMFusion achieves margins of up to 34.85% during the day and 17.03% during the night for 3D pedestrian detection. Additionally, pedestrian detection performance increases in mid-range distances by 10.6%. These improvements can be attributed to the enhanced visibility at night arising from additional active sensors, but also to their effective incorporation through a multi-modal distance-based weighting scheme.

Car detection improves slightly. This is due to labeling bias in the car category for 3D annotations, which prioritize precision over completeness. Objects with fewer than five LiDAR points were marked as "don't care", making it difficult to measure improvements in such challenging cases. For pedestrians, a different strategy focusing on completeness was employed, thereby providing a greater amount of challenging ground truth labels not available for the car category.

Adverse Weather Evaluation. Table 3 validates the proposed method in adverse weather, like snow and fog. State of the art LiDAR-RGB methods struggle with reduced visibility and back-scatter in adverse weather, causing such fusion approaches to perform significantly worse than in clear conditions, despite the relatively simple scene configurations. Relative to these baselines, SAMFusion achieves improvements of up to 13.6 AP (20.4% relative) for pedestrians at mid-range and 15.62 AP (60.51% relative) at long-range compared to the second-best (LiDAR and RGB) method in snowy scenes. In foggy scenes SAMFusion achieves high margins of up to 17.2 AP (101.2% relative) for pedestrians. For the car class in foggy conditions, it achieves improvements of up to 4.6 AP (5.2% relative).

Detection performance in adverse weather correlates with scene difficulty. The relative improvement in performance compared to Table 1 can be explained by the reduced number of road users in these weather splits simplifying the general task at hand as less people participate in road traffic.

5 Conclusion

We propose SAMFusion, a multi-modal adaptive sensor fusion method for robust 3D object detection in adverse weather for autonomous driving. Our approach enhances the conventional camera-LiDAR perception stack with gated camera and radar sensors, significantly improving performance in low-light and adverse weather scenarios, particularly for detecting narrow-profiled and vulnerable road users. SAMFusion employs depth-based adaptive blending of sensing modalities in conjunction with a learned multi-modal, distance-weighted decoder-query mechanism that leverages sensor-specific visibility over distance. We validate our method on the challenging SeeingThroughFog dataset [3], achieving an improvement of 17.2 AP points for pedestrians in dense fog and 15.62 AP points in heavy snow at long range. Future work will incorporate additional tasks such as planning and propagating uncertainty in adverse weather for improved decision-making and trajectory planning, further enhancing the robustness and effectiveness of autonomous driving systems in challenging conditions.

Acknowledgments This work was supported by the AI-SEE project with funding from the FFG, BMBF, and NRC-IRA. Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award, and an Amazon Science Research Award. Further the authors would like to thank Stefanie Walz, Samuel Brucker, Anush Kumar, Fathemeh Azimi and David Borts.

References

1. Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.L.: Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1090–1099 (2022)
2. Baumann, N., Baumgartner, M., Ghignone, E., Kühne, J., Fischer, T., Yang, Y.H., Pollefeys, M., Magno, M.: Cr3dt: Camera-radar fusion for 3d detection and tracking. arXiv preprint arXiv:2403.15313 (2024)
3. Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., Heide, F.: Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11682–11692 (2020)
4. Bijelic, M., Gruber, T., Ritter, W.: A benchmark for lidar sensors in fog: Is detection breaking down? In: 2018 IEEE Intelligent Vehicles Symposium (IV). pp. 760–767. IEEE (2018)
5. Bijelic, M., Gruber, T., Ritter, W.: Benchmarking image sensors under adverse weather conditions for autonomous driving. In: 2018 IEEE Intelligent Vehicles Symposium (IV). pp. 1773–1779. IEEE (2018)
6. Brazil, G., Liu, X.: M3d-rpn: Monocular 3d region proposal network for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9287–9296 (2019)
7. Broedermann, T., Sakaridis, C., Dai, D., Van Gool, L.: Hrfuser: A multi-resolution sensor fusion architecture for 2d object detection. In: IEEE International Conference on Intelligent Transportation Systems (ITSC) (2023)
8. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: Nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
9. Cai, H., Zhang, Z., Zhou, Z., Li, Z., Ding, W., Zhao, J.: Bevfusion4d: Learning lidar-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation. arXiv preprint arXiv:2303.17099 (2023)
10. Cai, Z., Vasconcelos, N.: Cascade r-cnn: High quality object detection and instance segmentation. IEEE transactions on pattern analysis and machine intelligence **43**(5), 1483–1498 (2019)
11. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
12. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1907–1915 (2017)
13. Chen, X., Zhang, T., Wang, Y., Wang, Y., Zhao, H.: Futr3d: A unified sensor fusion framework for 3d detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 172–181 (2023)
14. Chen, Y., Li, Y., Zhang, X., Sun, J., Jia, J.: Focal sparse convolutional networks for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5428–5437 (2022)
15. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Largekernel3d: Scaling up kernels in 3d sparse cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13488–13498 (2023)

16. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21674–21683 (2023)
17. Contributors, M.: MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d (2020)
18. Diaz-Ruiz, C.A., Xia, Y., You, Y., Nino, J., Chen, J., Monica, J., Chen, X., Luo, K., Wang, Y., Emond, M., et al.: Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21383–21392 (2022)
19. Ge, C., Chen, J., Xie, E., Wang, Z., Hong, L., Lu, H., Li, Z., Luo, P.: Metabev: Solving sensor failures for bev detection and map segmentation. arXiv preprint arXiv:2304.09801 (2023)
20. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
21. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
22. Grauer, Y.: Active gated imaging in driver assistance system. Advanced Optical Technologies **3**(2), 151–160 (2014)
23. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
25. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., Lu, L., Jia, X., Liu, Q., Dai, J., Qiao, Y., Li, H.: Planning-oriented autonomous driving (2023)
26. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
27. Huang, L., Li, Z., Sima, C., Wang, W., Wang, J., Qiao, Y., Li, H.: Leveraging vision-centric multi-modal expertise for 3d object detection. Advances in Neural Information Processing Systems **36** (2024)
28. Hwang, J.J., Kretzschmar, H., Manela, J., Rafferty, S., Armstrong-Crews, N., Chen, T., Anguelov, D.: Cramnet: Camera-radar fusion with ray-constrained cross-attention for robust 3d object detection. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII (2022)
29. Hwang, J.J., Kretzschmar, H., Manela, J., Rafferty, S., Armstrong-Crews, N., Chen, T., Anguelov, D.: Cramnet: Camera-radar fusion with ray-constrained cross-attention for robust 3d object detection. In: European Conference on Computer Vision. pp. 388–405. Springer (2022)
30. Jiao, Y., Jie, Z., Chen, S., Chen, J., Ma, L., Jiang, Y.G.: Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21643–21652 (2023)
31. Julca-Aguilar, F., Taylor, J., Bijelic, M., Mannan, F., Tseng, E., Heide, F.: Gated3d: Monocular 3d object detection from temporal illumination cues. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2938–2948 (2021)

32. Ku, J., Harakeh, A., Waslander, S.L.: In defense of classical image processing: Fast depth completion on the cpu. In: 2018 15th Conference on Computer and Robot Vision (CRV). pp. 16–22. IEEE (2018)
33. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics (NRL) **52** (1955)
34. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019)
35. Li, J., Wang, P., Xiong, P., Cai, T., Yan, Z., Yang, L., Liu, J., Fan, H., Liu, S.: Practical stereo matching via cascaded recurrent network with adaptive correlation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16263–16272 (2022)
36. Li, P., Chen, X., Shen, S.: Stereo r-cnn based 3d object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7644–7652 (2019)
37. Li, Q., Wang, Y., Wang, Y., Zhao, H.: Hdmapnet: An online HD map construction and evaluation framework. CoRR **abs/2107.06307** (2021)
38. Li, Y., Chen, Y., Qi, X., Li, Z., Sun, J., Jia, J.: Unifying voxel-based representation with transformer for 3d object detection. Advances in Neural Information Processing Systems **35**, 18442–18455 (2022)
39. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., Li, Z.: Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1477–1485 (2023)
40. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers (2022)
41. Liang, M., Yang, B., Chen, Y., Hu, R., Urtasun, R.: Multi-task multi-sensor fusion for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7345–7353 (2019)
42. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z.: Bevfusion: A simple and robust lidar-camera fusion framework. Advances in Neural Information Processing Systems **35**, 10421–10434 (2022)
43. Lin, Z., Liu, Z., Xia, Z., Wang, X., Wang, Y., Qi, S., Dong, Y., Dong, N., Zhang, L., Zhu, C.: Rcbevdet: Radar-camera fusion in bird's eye view for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14928–14937 (2024)
44. Liu, X., Zheng, C., Cheng, K.B., Xue, N., Qi, G.J., Wu, T.: Monocular 3d object detection with bounding box denoising in 3d by perceiver. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6436–6446 (2023)
45. Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., Zhang, X.: Petrv2: A unified framework for 3d perception from multi-camera images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3262–3272 (2023)
46. Liu, Z., Wu, Z., Tóth, R.: Smoke: Single-stage monocular 3d object detection via keypoint estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 996–997 (2020)
47. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2774–2781. IEEE (2023)

48. Ma, X., Liu, S., Xia, Z., Zhang, H., Zeng, X., Ouyang, W.: Rethinking pseudo-lidar representation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 311–327. Springer (2020)
49. Meyer, M., Kuschk, G.: Automotive radar dataset for deep learning based 3d object detection. In: 2019 16th european radar conference (EuRAD). pp. 129–132. IEEE (2019)
50. Mirza, M.J., Buerkle, C., Jarquin, J., Opitz, M., Oboril, F., Scholl, K.U., Bischof, H.: Robustness of object detectors in degrading weather conditions. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). pp. 2719–2724. IEEE (2021)
51. Nabati, R., Qi, H.: Centerfusion: Center-based radar and camera fusion for 3d object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1527–1536 (2021)
52. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS 2017 Workshop on Autodiff (2017), https://openreview.net/forum?id=BJJsrmfCZ
53. Peng, L., Chen, Z., Fu, Z., Liang, P., Cheng, E.: Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs (2022)
54. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 918–927 (2018)
55. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
56. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
57. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
58. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10529–10538 (2020)
59. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 770–779 (2019)
60. Shi, S., Wang, Z., Shi, J., Wang, X., Li, H.: From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. IEEE transactions on pattern analysis and machine intelligence **43**(8), 2647–2664 (2020)
61. Simonelli, A., Bulo, S.R., Porzi, L., López-Antequera, M., Kontschieder, P.: Disentangling monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1991–1999 (2019)
62. Sindagi, V.A., Zhou, Y., Tuzel, O.: Mvx-net: Multimodal voxelnet for 3d object detection. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 7276–7282. IEEE (2019)
63. Sun, J., Cao, Y., Chen, Q.A., Mao, Z.M.: Towards robust {LiDAR-based} perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In: 29th USENIX Security Symposium (USENIX Security 20). pp. 877–894 (2020)

64. Uricár, M., Ulicny, J., Sistu, G., Rashed, H., Krizek, P., Hurych, D., Vobecky, A., Yogamani, S.: Desoiling dataset: Restoring soiled areas on automotive fisheye cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
65. Vora, S., Lang, A.H., Helou, B., Beijbom, O.: Pointpainting: Sequential fusion for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4604–4612 (2020)
66. Walia, A., Walz, S., Bijelic, M., Mannan, F., Julca-Aguilar, F.D., Langer, M.S., Ritter, W., Heide, F.: Gated2gated: Self-supervised depth estimation from gated images. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2801–2811 (2021)
67. Walz, S., Bijelic, M., Ramazzina, A., Walia, A., Mannan, F., Heide, F.: Gated stereo: Joint depth estimation from gated and wide-baseline active stereo cues. In: Proceedings - 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 (2023)
68. Walz, S., Bijelic, M., Ramazzina, A., Walia, A., Mannan, F., Heide, F.: Gated stereo: Joint depth estimation from gated and wide-baseline active stereo cues. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13252–13262 (2023)
69. Wang, C., Ma, C., Zhu, M., Yang, X.: Pointaugmenting: Cross-modal augmentation for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11794–11803 (2021)
70. Wang, H., Tang, H., Shi, S., Li, A., Li, Z., Schiele, B., Wang, L.: Unitr: A unified and efficient multi-modal transformer for bird's-eye-view representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6792–6802 (2023)
71. Wang, J., Lan, S., Gao, M., Davis, L.S.: Infofocus: 3d object detection for autonomous driving with dynamic information modeling. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16. pp. 405–420. Springer (2020)
72. Wang, S., Caesar, H., Nan, L., Kooij, J.F.: Unibev: Multi-modal 3d object detection with uniform bev encoders for robustness against missing sensor modalities. arXiv preprint arXiv:2309.14516 (2023)
73. Wang, T., Zhu, X., Pang, J., Lin, D.: Fcos3d: Fully convolutional one-stage monocular 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 913–922 (2021)
74. Wang, Y., Fathi, A., Kundu, A., Ross, D.A., Pantofaru, C., Funkhouser, T., Solomon, J.: Pillar-based object detection for autonomous driving. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. pp. 18–34. Springer (2020)
75. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: Conference on Robot Learning. pp. 180–191. PMLR (2022)
76. Wu, P., Gu, L., Yan, X., Xie, H., Wang, F.L., Cheng, G., Wei, M.: Pv-rcnn++: semantical point-voxel feature interaction for 3d object detection. The Visual Computer **39**(6), 2425–2440 (2023)
77. Xie, Y., Xu, C., Rakotosaona, M.J., Rim, P., Tombari, F., Keutzer, K., Tomizuka, M., Zhan, W.: Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection (2023)

78. Xu, S., Zhou, D., Fang, J., Yin, J., Bin, Z., Zhang, L.: Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). pp. 3047–3054. IEEE (2021)
79. Yan, J., Liu, Y., Sun, J., Jia, F., Li, S., Wang, T., Zhang, X.: Cross modal transformer: Towards fast and robust 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18268–18278 (2023)
80. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
81. Yang, B., Luo, W., Urtasun, R.: Pixor: Real-time 3d object detection from point clouds. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7652–7660 (2018)
82. Yang, Z., Sun, Y., Liu, S., Jia, J.: 3dssd: Point-based 3d single stage object detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11040–11048 (2020)
83. Yang, Z., Chen, J., Miao, Z., Li, W., Zhu, X., Zhang, L.: Deepinteraction: 3d object detection via modality interaction. Advances in Neural Information Processing Systems **35**, 1992–2005 (2022)
84. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11784–11793 (2021)
85. Yin, T., Zhou, X., Krähenbühl, P.: Multimodal virtual point 3d detection. Advances in Neural Information Processing Systems **34**, 16494–16507 (2021)
86. Yoo, J.H., Kim, Y., Kim, J., Choi, J.W.: 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. pp. 720–736. Springer (2020)
87. Zhang, J., Singh, S.: Visual-lidar odometry and mapping: Low-drift, robust, and fast. In: 2015 IEEE International Conference on Robotics and Automation (ICRA). pp. 2174–2181. IEEE (2015)
88. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018)
89. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable {detr}: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)