RoboFusion: Towards Robust Multi-Modal 3D Object Detection via SAM

Ziying Song[1,2] , Guoxing Zhang[3] , Lin Liu[1,2] , Lei Yang[4] , Shaoqing Xu[5] , Caiyan Jia[1,2]* , Feiyang Jia[1,2] , Li Wang[6]

[1]School of Computer Science and Technology, Beijing Jiaotong University, China
[2] Beijing Key Lab of Traffic Data Analysis and Mining, China
[3]Hebei University of Science and Technology, China
[4]Tsinghua University, China
[5]University of Macau, China
[6]Beijing Institute of Technology, China
{songziying, cyjia, feiyangjia}@bjtu.edu.cn

Abstract

Multi-modal 3D object detectors are dedicated to exploring secure and reliable perception systems for autonomous driving (AD). Although achieving state-of-the-art (SOTA) performance on clean benchmark datasets, they tend to overlook the complexity and harsh conditions of real-world environments. With the emergence of visual foundation models (VFMs), opportunities and challenges are presented for improving the robustness and generalization of multi-modal 3D object detection in AD. Therefore, we propose **RoboFusion**, a robust framework that leverages VFMs like SAM to tackle out-of-distribution (OOD) noise scenarios. We first adapt the original SAM for AD scenarios named **SAM-AD**. To align SAM or **SAM-AD** with multi-modal methods, we then introduce **AD-FPN** for upsampling the image features extracted by SAM. We employ wavelet decomposition to denoise the depth-guided images for further noise reduction and weather interference. At last, we employ self-attention mechanisms to adaptively reweight the fused features, enhancing informative features while suppressing excess noise. In summary, RoboFusion significantly reduces noise by leveraging the generalization and robustness of VFMs, thereby enhancing the resilience of multi-modal 3D object detection. Consequently, RoboFusion achieves SOTA performance in noisy scenarios, as demonstrated by the KITTI-C and nuScenes-C benchmarks. Code is available at https://github.com/adept-thu/RoboFusion.

1 Introduction

Multi-modal 3D object detection plays a pivotal role in autonomous driving (AD) [Wang *et al.*, 2023a; Song *et al.*, 2024a]. Different modalities often provide complementary information. For instance, images contain richer semantic representations, yet lack depth information. In contrast, point clouds offer geometric and depth details, but they are sparse and lack semantic information. Therefore, effectively leveraging the advantages of multi-model while mitigating their limitations contributes to enhancing the robustness and accuracy of perception systems [Song *et al.*, 2023].

With the emergence of AD datasets [Geiger *et al.*, 2012; Caesar *et al.*, 2020; Zhang *et al.*, 2023c], state-of-the-art (SOTA) methods [Liu *et al.*, 2023; Bai *et al.*, 2022; Chen *et al.*, 2022; Huang *et al.*, 2020; Li *et al.*, 2023; Song *et al.*, 2024b] on 'clean' datasets [Geiger *et al.*, 2012; Caesar *et al.*, 2020] have achieved record-breaking performance. However, they overlook the exploration of robustness and generalization in out-of-distribution (OOD) scenarios [Dong *et al.*, 2023]. For example, the KITTI dataset [Geiger *et al.*, 2012] lacks severe weather conditions. When SOTA methods [Chen *et al.*, 2022; Li *et al.*, 2023; Liu *et al.*, 2023] learn from these sunny weather datasets, can they truly generalize and maintain robustness in severe weather conditions like snow and fog?

The answer is 'No', as shown in Fig. 1 and verified in Table 3. People often utilize domain adaptation (DA) techniques to address these challenges [Wang *et al.*, 2023b; Tsai *et al.*, 2023; Peng *et al.*, 2023; Hu *et al.*, 2023]. Although DA techniques improve the robustness of 3D object detection and reduce the need for annotated data, they have some profound drawbacks, including domain shift limitations, label shift issues, and overfitting risks [Oza *et al.*, 2023]. For instance, DA techniques may be constrained if the differences between two domains are significant, leading to performance degradation on the target domain.

Recently, both Natural Language Processing (NLP) and Computer Vision (CV) have witnessed the appearance and the power of a series of foundation models [Kirillov *et al.*, 2023; OpenAI, 2023; Zhao *et al.*, 2023; Zhang *et al.*, 2023a], resulting in the emergence of new paradigms in deep learning. For example, a series of novel visual foundation models (VFMs) [Kirillov *et al.*, 2023; Zhao *et al.*, 2023; Zhang *et al.*, 2023a] have been developed. Thanks to their extensive training on huge datasets, these models exhibit powerful generalization capabilities. These developments have inspired new

*Corresponding author

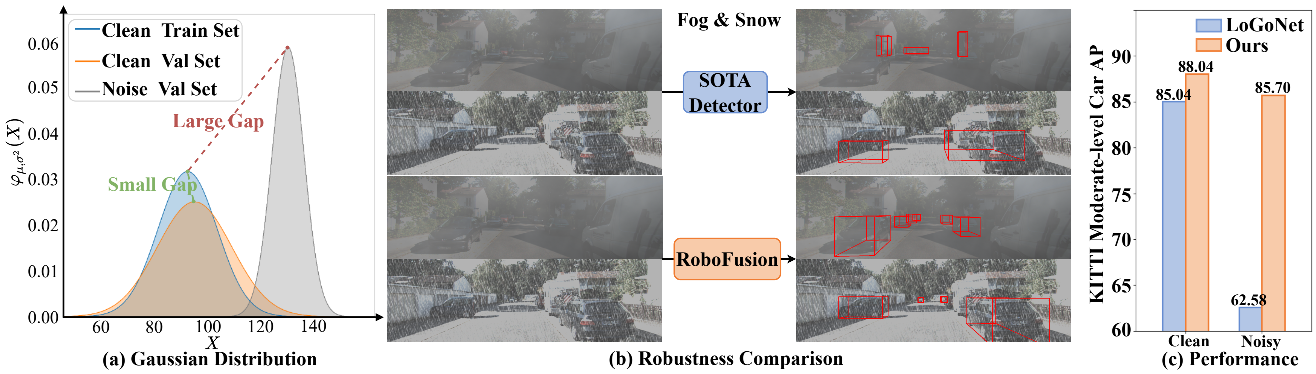

Figure 1: **(a)** We employ Gaussian distributions to represent the distributional disparities among the datasets. Indeed, there exists a large gap in data distribution between an OOD noise validation set and a clean validation set. Where the X-axis represents the set of mean pixel values in a dataset, $X = \{x_i \,|\, i = 1, 2, ..., N\}$, with $x_i = \frac{1}{H\times W\times 3}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{3}(I_{ijk})$, where N is the number of the dataset, H is the height, W is the width, and I_{ijk} denotes the pixel values for each image. **(b)** Visual foundation models (VFMs) like SAM [Kirillov *et al.*, 2023], show robust performance in many noisy scenarios. Yet, the current methods are not robust enough to predict 3D tasks for autonomous driving perception. **(c)** To this end, we propose a robust framework, RoboFusion, which employs VFMs at the SOTA multi-modal 3D object detection. Empirical results reveal that our method surpasses the Top-performing LoGoNet[Li *et al.*, 2023] on the KITTI Leaderboard by a margin of 23.12% mAP (Weather) on KITTI-C [Dong *et al.*, 2023] noisy scenarios. Notably, our RoboFusion shows better performance with LoGoNet [Li *et al.*, 2023] in clean KITTI [Geiger *et al.*, 2012] dataset.

ideas, leveraging the robustness and generalization abilities of VFMs to achieve generalization in OOD noisy scenarios, much like how adults generalize knowledge when encountering new situations, without relying on DA techniques [Wang *et al.*, 2023b; Tsai *et al.*, 2023].

Inspired by the success of VFMs in CV tasks, in this work, we intend to use these models to tackle the challenges of multi-modal 3D object detectors in OOD noise scenarios. Therefore, we propose a robust framework, RoboFusion, which leverages VFMs like SAM to adapt a 3D multi-modal object detector from clean scenarios to OOD noise scenarios. In particular, the adaptation strategies for SAM are as follows. 1) We utilize features extracted from SAM rather than inference segmentation results. 2) We propose **SAM-AD**, which is a pre-trained SAM for AD scenarios. 3) We introduce a novel **AD-FPN** to address the issue of feature upsampling for aligning VFMs with multi-modal 3D object detector. 4) To further reduce noise interference and retain essential signal features, we design a **Depth-Guided Wavelet Attention (DGWA)** module that effectively attenuates both high-frequency and low-frequency noises. 5) After fusing point cloud features and image features, we propose **Adaptive Fusion** to further enhance feature robustness and noise resistance through self-attention to re-weight the fused features adaptively. We validate RoboFusion's robustness against OOD noise scenarios in KITTI-C and nuScenes-C datasets [Dong *et al.*, 2023], achieving SOTA performance amid noise, as shown in Fig. 1.

2 Related Work

2.1 Multi-Modal 3D Object Detection

Currently, multi-modal 3D object detection has received considerable attention on popular datasets [Geiger *et al.*, 2012; Caesar *et al.*, 2020]. BEVFusion [Liu *et al.*, 2023] fuse multi-modal representations in a unified 3D or BEV space. TransFusion [Bai *et al.*, 2022] builds a two-stage pipeline where proposals are generated based on LiDAR features and further refined using query image features. DeepInteraction [Yang *et al.*, 2022] and SparseFusion [Xie *et al.*, 2023] further optimize the camera branch on top of TransFusion. Previous methods are highly optimized to achieve the best performance on clean datasets. However, they ignore common factors in the real world (*e.g.,* bad weather and sensor noise). In this work, we consider a real-world robustness perspective and design a robust multi-modal 3D perception framework, RoboFusion.

2.2 Visual Foundation Models for 3D Object Detection

Motivated by the success of Large Language Models (LLMs) [OpenAI, 2023], VFMs start to be explored in CV community. SAM [Kirillov *et al.*, 2023] leverages ViT [Dosovitskiy *et al.*, 2020] to train on the huge SA-1B dataset, containing 11 million samples, which enables SAM to be generalized to many scenes. Currently, there have been a few research endeavors aiming at integrating 3D object detectors with SAM. For instance, SAM3D [Zhang *et al.*, 2023b], as a LiDAR-only method, solely transforms LiDAR's 3D perspective into a BEV (Bird's Eye View) 2D space to harness the generalization capabilities of SAM, yielding sub-optimal performance on 'clean' datasets. Another in progress work, 3D-Box-Segment-Anything [1], tries to utilize SAM for 3D object detection. This indicates the highly attention of SAM like foundation models in 3D scenes in the literature. Our RoboFusion, as a multi-modal method, gives clear strategies to leverage the generalization capabilities of VFMs to address the OOD noise challenges inherent in existing 3D multi-modal object detection methods.

[1] https://github.com/dvlab-research/3D-Box-Segment-Anything

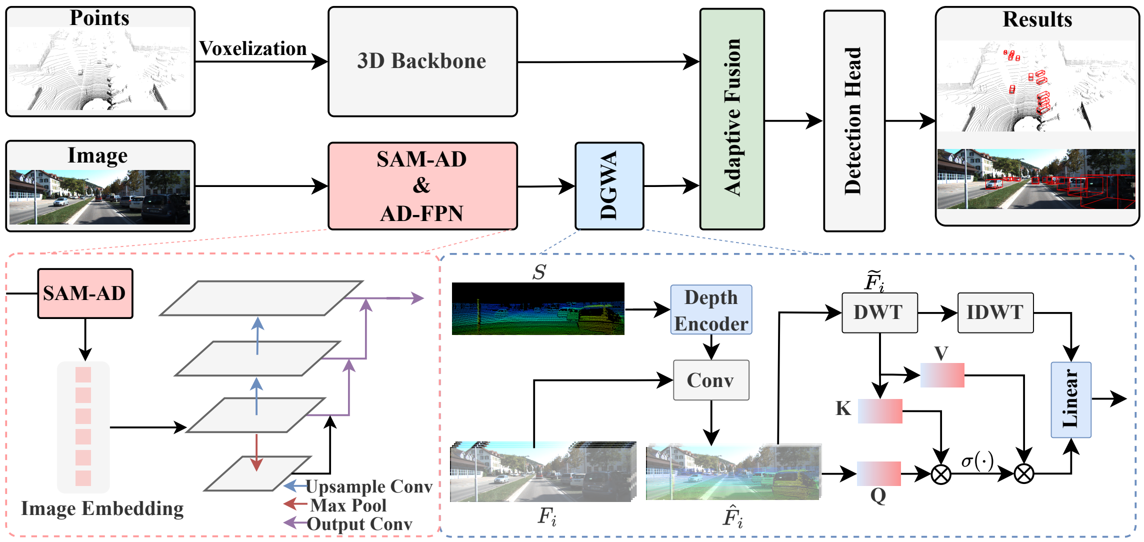

Figure 2: The framework of RoboFusion. The LiDAR branch follows the baselines [Chen *et al.*, 2022; Bai *et al.*, 2022] to generate LiDAR features. In the camera branch, first, we extract robust image features using a highly optimized SAM-AD and acquire multi-scale features using AD-FPN. Second, the sparse depth map S is generated by the raw points and fed into a depth encoder to obtain depth features and fused with multi-scale image features F_i to obtain depth-guided image features $\hat{F}_i$. Then wave attention is used to remove the mutation noise. Finally, adaptive Fusion integrates point cloud features with robust image features with depth information via self-attention mechanism.

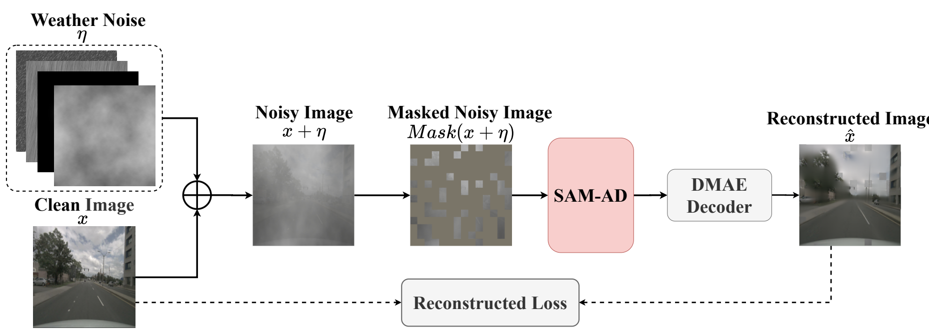

Figure 3: An illustration of the pre-training framework. We corrupt a clean image x by η which contains multiple weather noises and then randomly masking several patches on a noisy image $x+\eta$ to obtain a masked noisy image $Mask(x+\eta)$. The SAM-AD and DMAE decoder are trained to reconstruct the clean image $\hat{x}$ from $Mask(x+\eta)$.

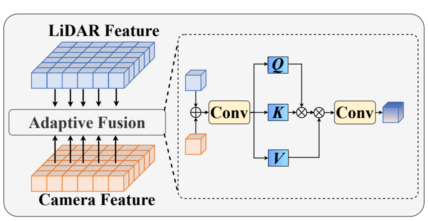

Figure 4: The architecture of **Adaptive Fusion**, which involves adaptively re-weighting the fused features using self-attention.

3 RoboFusion

In this section, we present RoboFusion, a framework that harnesses the robustness and generalization capabilities of VFMs such as SAM [Kirillov *et al.*, 2023] for multi-modal 3D object detection. The overall architecture is depicted in Fig. 2 and comprises the following components: 1) **SAM-AD & AD-FPN** module which obtains robust multi-scale image features, 2) **Depth-Guided Wavelet Attention (DGWA)** module which employs wavelet decomposition to denoise depth-guided image features, 3) **Adaptive Fusion** module which adaptively fuses point cloud features with image features.

3.1 SAM-AD & AD-FPN

Preliminaries

SAM [Kirillov *et al.*, 2023], a VFM, achieves generalization across diverse scenes due to its extensive training on the large-scale SA-1B dataset—with over 11 million samples and 1 billion high-quality masks. Currently, SAM family [Kirillov *et al.*, 2023; Zhao *et al.*, 2023; Zhang *et al.*, 2023a] primarily support 2D tasks. However, directly extending VFMs like SAM to 3D tasks presents a gap. To address this, we combine SAM with multi-modal 3D models, merging 2D robust feature representations with 3D point cloud features to achieve robust fused features.

SAM-AD

To further adapt SAM with AD (autonomous driving) scenarios, we perform pre-training on SAM to obtain SAM-AD. Specifically, we curate an extensive collection of image samples from well-established datasets (*i.e.,* KITTI [Geiger *et al.*, 2012] and nuScenes [Caesar *et al.*, 2020]), forming the foundational AD dataset. Following DMAE [Wu *et al.*, 2023], we perform pre-training on SAM to obtain SAM-AD in AD scenarios, as shown in Fig. 3. We denote x as a clean image from the AD dataset (*i.e.* KITTI [Geiger *et al.*, 2012] and nuScenes [Caesar *et al.*, 2020]) and η as a set of noise images generated by [Dong *et al.*, 2023] based on x. And the noise type and the severity are randomly chosen from the four weather (i.e., rain, snow, fog, and strong sunlight) and the five severities from 1 to 5, respectively. We employ the image encoder of SAM [Kirillov *et al.*, 2023] , MobileSAM [Zhang *et al.*, 2023a] as our encoder while the decoder and the reconstruction loss are the same as DMAE [Wu *et al.*, 2023]. For FastSAM [Zhao *et al.*, 2023], we adopt YOLOv8 [2] to pre-train FastSAM on the AD dataset. To avoid overfitting, we use random resizing and cropping as data augmentation. We also set the mask ratio as 0.75 and have trained 400 epochs on 8 NVIDIA A100 GPUs.

AD-FPN

As a promptable segmentation model, SAM has three components: image encoder, prompt encoder and mask decoder. Generally, the image encoder can provide high-quality and highly robust image embedding for downstream models, while the mask decoder is only designed to provide decoding services for semantic segmentation. Furthermore, what we require are robust image features rather than the processing of prompting information by the prompt encoder. Therefore, we employ SAM's image encoder to extract robust image features. However, SAM utilizes the ViT series [Dosovitskiy *et al.*, 2020] as its image encoder, which excludes multi-scale features and provides only high-dimensional low-resolution features. To generate the multi-scale features required for object detection, inspired by [Li *et al.*, 2022a], we design an AD-FPN that offers ViT-based multi-scale features. Specifically, leveraging height-dimensional image embedding with stride 16 (scale=$1/16$) provided by SAM, we produce a series of multi-scale features F_{ms} with stride of $\{32, 16, 8, 4\}$. Sequentially, we acquire multi-scale feature $F_i \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_i}$ by integrate F_{ms} in a bottom-up manner similar to FPN [Lin *et al.*, 2017].

3.2 Depth-Guided Wavelet Attention

Although SAM-AD or SAM has the capability to extract robust image features, the gap between 2D and 3D domains still persists and cameras lacking geometric information in a corrupted environment often amplify noise and give rise to negative migration issues. To mitigate this problem, we propose the Depth-Guided Wavelet Attention (DGWA) module, which can be split into two steps. 1) A depth-guided network is designed, that adds geometry prior to image features by combining image features and depth features from a point cloud. 2) The features of an image are decomposed into four wavelet subbands using the Haar wavelet transform [Liu *et al.*, 2020a], then attention mechanism allows to denoise informative features in the subbands.

Formally, given image features $F_i \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_i}$ and raw points $P \in \mathbb{R}^{N,C_p}$ as input. We project P onto the image plane to acquire a sparse depth map $S \in \mathbb{R}^{H\times W\times 2}$. Next, we feed S into the depth encoder $DE(\cdot)$, which consists of several convolution and max pooling blocks, to acquire depth features $F_d \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_i}$. Afterward, we leverage convolution encode (F_i, F_d) to acquire depth-guided image features $\hat{F}_i \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 16}$, given by

$$\hat{F}_i = Conv(Concat(F_i, DE(S))). \tag{1}$$

Subsequently, we employ discrete wavelet transform (DWT), a reversible operator, to partition the input $\hat{F}_i$ into four subbands. Specifically, we encode the rows and columns of the input separately into one low-frequency band $\widetilde{f}_i^{LL} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 4}$ and three high-frequency bands $(\widetilde{f}_i^{LH}, \widetilde{f}_i^{HL}, \widetilde{f}_i^{HH}) \in \mathbb{R}^{\frac{H}{8},\frac{W}{8},4}$, with the low-filter $f_L = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and the high-filter $f_H = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$. In this state, the low-frequency band retains coarse-grained information while the high-frequency band retains fine-grained information. In other words, it is easier to capture the mutation signal, so as to filter the noise information. We concatenate the four-subband features along channel dimension to acquire wavelet features $\widetilde{F}_i = [\hat{f}_i^{LL}, \hat{f}_i^{LH}, \hat{f}_i^{HL}, \hat{f}_i^{HH}] \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 16}$. Next, we perform wave-attention Att_ω to query informative features in the wavelet features. Concretely, we employ $\hat{F}_i$ as a Query and $\widetilde{F}_i$ as a Key/Value given by

$$F_{att} = Att_\omega(\hat{F}_i, \widetilde{F}_i) = \sigma(\frac{\hat{F}_i W^q (\widetilde{F}_i W^k)^T}{\sqrt{C_i}})\widetilde{F}_i W^v. \tag{2}$$

Finally, we leverage the IDWT (inverse DWT) to convert $\widetilde{F}_i$ back to $\hat{F}_i$ and integrate this converted $\hat{F}_i$ and F_{att} to obtain denoise features $F_{out} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 16}$ by

$$F_{out} = MLP(Concat(F_{att}, \hat{F}_i)), \tag{3}$$

where F_{out} preserves informative features and restrains redundant mutation noise in the frequency domain.

3.3 Adaptive Fusion

Following the incorporation of image depth features within the DGWA module, we propose the **Adaptive Fusion** technique to combine point cloud attributes with robust image features enriched with depth information. Specifically, different types of noise affect LiDAR and images to different degrees, which raises a corruption imbalance problem. Therefore, considering the distinct influences of various noises on LiDAR and camera, we employ self-attention to re-weight the fused features adaptively as shown in Fig. 4. The corruption degree of modality-specificity is dynamic, and self-attention mechanism allows adaptive re-weighting features to enhance informative features and suppress redundant noise.

[2]https://github.com/ultralytics/ultralytics

Method	AP_{3D}(%) (*validation set*)				AP_{3D}(%) (*test set*)			
	mAP	Easy	Mod.	Hard	mAP	Easy	Mod.	Hard
Voxel R-CNN	86.84	92.38	85.29	82.86	83.19	90.90	81.62	77.06
VFF	86.91	92.31	85.51	82.92	83.62	89.50	82.09	79.29
CAT-Det	83.58	90.12	81.46	79.15	82.62	89.87	81.32	76.68
LoGoNet	87.13	92.04	85.04	84.31	85.87	91.80	**85.06**	**80.74**
Focals Conv-F	-	-	-	-	83.47	90.55	82.28	77.59
Baseline*	86.75	92.05	85.51	82.70	-	-	-	-
RoboFusion-L	**88.87**	**93.30**	**88.04**	**85.27**	**85.58**	91.75	84.08	80.71
RoboFusion-B	88.45	93.22	87.87	84.27	85.32	**91.98**	83.76	80.23
RoboFusion-T	88.08	93.28	87.60	83.36	85.09	91.68	83.70	79.89

Table 1: Comparison with SOTA methods on **KITTI validation and test** sets for car class with AP of R_{40}. $*$ denotes our reproduced results based on the officially released codes.

Method	LiDAR	Camera	*validation set*		*test set*	
			NDS	mAP	NDS	mAP
FUTR3D	VoxelNet	ResNet-101	68.3	64.5	-	-
BEVFusion-mit	VoxelNet	Swin-T	71.4	68.5	72.9	70.2
DeepInteraction	VoxelNet	ResNet-50	72.6	69.9	73.4	70.8
CMT	VoxelNet	ResNet-50	72.9	70.3	74.1	72.0
SparseFusion	VoxelNet	ResNet-50	72.8	70.4	73.8	72.0
TransFusion	VoxelNet	ResNet-50	71.3	67.5	71.6	68.9
Baseline*	VoxelNet	ResNet-50	70.8	67.3	-	-
RoboFusion-L	VoxelNet	SAM	72.1	69.9	72.0	69.9
RoboFusion-B	VoxelNet	FastSAM	71.9	69.4	71.8	69.4
RoboFusion-T	VoxelNet	MobileSAM	71.3	69.1	71.5	69.1

Table 2: Comparison with SOTA methods on **nuScenes validation and test** sets. $*$ denotes our reproduced results based on the officially released codes.

Rai	Clean	*Weather*					*Sensor*		
		mAP	Snow	Rain	Fog	S.L.	D.	C.O.	C.T.
SECOND†	81.59	64.33	52.34	52.55	74.10	78.32	80.18	73.59	80.24
PointPillars†	78.41	49.80	36.47	36.18	64.28	62.28	76.49	70.28	70.85
PointRCNN†	80.57	59.14	50.36	51.27	72.14	62.78	80.35	73.94	71.53
PV-RCNN†	84.39	65.83	52.35	51.58	79.47	79.91	82.79	76.09	82.34
SMOKE†	7.09	4.51	2.47	3.94	5.63	6.00	-	-	-
ImVoxelNet†	11.49	3.22	0.22	1.24	1.34	10.08	-	-	-
EPNet†	82.72	46.21	34.58	36.27	44.35	69.65	82.09	76.10	82.10
Focals Conv-F†	85.88	50.40	34.77	41.30	44.55	80.97	84.95	78.06	**85.82**
LoGoNet*	85.04	62.58	51.45	55.80	67.53	75.54	83.68	77.17	82.00
RoboFusion-L	**88.04**	**85.70**	**85.29**	**86.48**	**85.53**	**85.50**	**85.71**	**83.17**	84.12
RoboFusion-B	87.87	84.70	84.11	85.54	84.00	85.15	84.34	81.30	82.45
RoboFusion-T	87.60	84.60	84.67	84.79	84.17	84.75	84.11	81.21	83.07

Table 3: Comparison with SOTA methods on **KITTI-C validation** set. The results are evaluated based on the car class with AP of R_{40} at moderate difficulty. 'S.L.', 'D.', 'C.O.', and 'C.T.' denotes Strong Sunlight, Density, Cutout, and Crosstalk, respectively. †: Results from Ref. [Dong *et al.*, 2023]. $*$ denotes re-implement result.

4 Experiments

4.1 Datasets

We perform experiments on both the clean public benchmarks (KITTI [Geiger *et al.*, 2012] and nuScenes [Caesar *et al.*, 2020]) and the noisy public benchmarks (KITTI-C[Dong *et al.*, 2023] and nuScenes-C [Dong *et al.*, 2023]).

Clean Datasets

The KITTI dataset provides synchronized LiDAR point clouds and front-view camera images, consists of 3,712 training samples, 3,769 validation samples and 7,518 test samples. The standard evaluation metric for object detection is the mean Average Precision (mAP), computed using recall at 40 positions (R40).

The nuScenes dataset is a large-scale 3D detection benchmark consisting of 700 training scenes, 150 validation scenes, and 150 testing scenes. The data are collected using six multi-view cameras and a 32-channel LiDAR sensor. It includes 360-degree object annotations for 10 object classes. To evaluate the detection performance, the primary metrics used are the mean Average Precision (mAP) and the nuScenes detection score (NDS).

Noisy Datasets

In terms of data robustness, [Dong *et al.*, 2023] has designed 27 types of common corruptions for both LiDAR and camera, with the aim of benchmarking the corruption robustness of existing 3D object detectors. [Dong *et al.*, 2023] has established corruption robustness benchmarks [3], including **KITTI-C** and **nuScenes-C**, by synthesizing corruptions on public datasets. Specifically, we utilize **KITTI-C** and **nuScenes-C** in our work. It is worth noting that Ref. [Dong *et al.*, 2023] has only added noise to the validation dataset and kept the train and test datasets clear.

4.2 Experimental Settings

Network Architecture

Our RoboFusion consists of three variants: RoboFusion-L, RoboFusion-B, and RoboFusion-T, which utilize the models SAM-B [Kirillov *et al.*, 2023], FastSAM [Zhao *et al.*, 2023], and MobileSAM [Zhang *et al.*, 2023a], respectively. It is noteworthy that due to the convolutional operations of FastSAM in RoboFusion-B which is capable of generating multi-scale features, the AD-FPN module is not employed. Since KITTI and nuScenes are distinct datasets with varying evaluation metrics and characteristics, we provide a detailed description of our RoboFusion settings for each dataset.

We validate our RoboFusion on the KITTI dataset using Focals Conv [Chen *et al.*, 2022] as the baseline. The input voxel size is set to (0.05m, 0.05m, 0.1m), with anchor sizes for cars at [3.9, 1.6, 1.56] and anchor rotations at [0, 1.57]. We adopt the same data augmentation solution as Focals Conv-F.

We validate our RoboFusion on the nuScenes dataset using TransFusion [Bai *et al.*, 2022] as the baseline. The detection range for the X and Y axis is set at [-54m, 54m] and [-5m, 3m] for the Z axis. The input voxel size is set at (0.075m, 0.075m, 0.2m), and the maximum number of point clouds contained in each voxel is set to 10. It is noteworthy that the Adaptive Fusion module is applied exclusively to Focals Conv rather than TransFusion, while TransFusion uses its own fusion module.

Training and Testing Details

Our RoboFusion is meticulously trained from scratch using the Adam optimizer and incorporates several foundation

[3] https://github.com/thu-ml/3D_Corruptions_AD

Method	Clean	Weather					Sensor		
		mAP	Snow	Rain	Fog	S.L.	D.	C.O.	C.T.
PointPillars†	27.69	25.87	27.57	27.71	24.49	23.71	27.27	24.14	25.92
SSN†	46.65	43.70	46.38	46.50	41.64	40.28	46.14	40.95	44.08
CenterPoint†	59.28	52.49	55.90	56.08	43.78	54.20	58.60	56.28	56.64
FCOS3D†	23.86	11.44	2.01	13.00	13.53	17.20	-	-	-
PGD†	23.19	12.85	2.30	13.51	12.83	22.77	-	-	-
DETR3D†	34.71	22.00	5.08	20.39	27.89	34.66	-	-	-
BEVFormer†	41.65	26.29	5.73	24.97	32.76	41.68	-	-	
FUTR3D†	64.17	55.50	52.73	58.40	53.19	57.70	63.72	62.25	62.66
TransFusion†	66.38	58.87	63.30	63.35	53.67	55.14	65.77	63.66	64.67
BEVFusion†	68.45	61.87	62.84	66.13	54.10	64.42	67.79	66.18	67.32
DeepInteraction*	69.90	62.14	62.36	66.48	54.79	64.93	68.15	66.23	68.12
CMT*	**70.28**	63.46	62.56	61.44	66.26	63.59	**69.65**	68.70	68.26
RoboFusion-L	69.91	**67.24**	**67.12**	**67.58**	**67.01**	**67.24**	69.48	**69.18**	**68.68**
RoboFusion-B	69.40	66.33	66.07	67.01	65.54	66.71	69.02	69.01	68.04
RoboFusion-T	69.09	65.82	65.96	66.45	64.34	66.54	68.58	68.20	68.17

Table 4: Comparison with SOTA methods on **nuScenes-C validation** set with mAP. 'S.L.', 'D.', 'C.O.', and 'C.T.' denotes Strong Sunlight, Density, Cutout, and Crosstalk, respectively. †: Results from Ref. [Dong *et al.*, 2023]. ∗ denotes re-implement result.

Method	Model Size	FPS (A100)	mAP (Weather)	mAP (Clean)	RCE (%)
RoboFusion-L	97.54M	3.1	67.24	69.91	0.04
RoboFusion-B	81.01M	3.5	66.33	69.40	0.04
RoboFusion-T	13.94M	6.0	65.82	69.09	0.05
DeepInteraction	57.82M	4.9	62.14	69.90	0.10
TransFusion	36.96M	6.2	58.37	66.38	0.12

Table 5: Performance of different VFMs on RoboFusion. 'RCE' denotes Relative Corruption Error [Dong *et al.*, 2023]. 'mAP (Weather)' denotes the average value across four types of weather corruptions, Snow, Rain, Fog, and Strong Sunlight.

models as image encoders including SAM, FastSAM and MobileSAM. To enable effective training on the KITTI and nuScenes datasets, we utilize 8 NVIDIA A100 GPUs for network training. Additionally, the runtime is evaluated on an NVIDIA A100 GPU. Specifically, for KITTI, our RoboFusion based on Focals Conv[Chen *et al.*, 2022] involves training for 80 epochs. For nuScenes, our RoboFusion based on TransFusion [Bai *et al.*, 2022] has 20 epochs of training. During the model inference stage, we employ a non-maximal suppression (NMS) operation in the Region Proposal Network (RPN) with an IoU threshold of 0.7. We select the top 100 region proposals to serve as inputs for the detection head. After refinement, we apply NMS again with an IoU threshold of 0.1 to eliminate redundant predictions. For additional details regarding our method, please refer to OpenPCDet [4] .

4.3 Comparing with State-of-the-art

We conduct evaluations on the clean datasets KITTI and nuScenes, as well as the noisy datasets KITTI-C and nuScenes-C. While SOTA methods are primarily focused on achieving high accuracy, we place greater emphasis on the robustness and generalization of the methods. These factors are crucial for the practical deployment of 3D object detection in AD scenarios, making the evaluation on the noisy datasets more important in our perspective.

Results on the Clean Benchmark

As shown in Table 1, we compare our RoboFusion with SOTA methods, including Voxel R-CNN [Deng *et al.*, 2021], VFF [Li *et al.*, 2022b], CAT-Det[Zhang *et al.*, 2022] , Focals Conv-F [Chen *et al.*, 2022], and LoGoNet [Li *et al.*, 2023] on the KITTI validation and test sets. As shown in Table 2, we also compare our RoboFusion with SOTA methods, including FUTR3D [Chen *et al.*, 2023], TransFusion [Bai *et al.*, 2022], BEVFusion [Liu *et al.*, 2023], DeeepInteraction [Yang *et al.*, 2022], CMT [Yan *et al.*, 2023] and SparseFusion [Xie *et al.*, 2023], on the nuScenes test and validation sets. Our RoboFusion has achieved SOTA performance on the clean benchmarks (KITTI and nuScenes).

Results on the Noisy Benchmark

In the real-world AD scenarios, the distribution of data often differs from that of training or testing data, as shown in Fig. 1 (a). Specifically, Ref. [Dong *et al.*, 2023] provides a novel noisy benchmark that includes KITTI-C and nuScenes-C, which we primarily use to evaluate the weather and sensor noise corruptions, including rain, snow, fog, and strong sunlight, density, cutout, and so on. In addition, comparisons of our RoboFusion with SOTA methods in other settings are presented in the Appendix [5].

As shown in Table 3, SOTA methods, including SECOND [Yan *et al.*, 2018], PointPillars [Lang *et al.*, 2019], PointRCNN [Shi *et al.*, 2019], PV-RCNN [Shi *et al.*, 2020], SMOKE [Liu *et al.*, 2020b], ImVoxelNet [Rukhovich *et al.*, 2022], EpNet[Huang *et al.*, 2020], Focals Conv-F[Chen *et al.*, 2022], and LoGoNet[Li *et al.*, 2023], experience a significant decrease in performance on the noisy scenarios, particularly for weather conditions such as snow and rain. It can be attributed to the fact that the 'clean' KITTI dataset does not include examples in snowy or rainy weather. On the other hand, VFMs like SAM-AD have been trained on a diverse range of data and exhibit robustness and generalization to OOD scenarios, leading to higher performance on our RoboFusion metric. Furthermore, multi-modal methods like LoGoNet, and Focals Conv-F demonstrate better robustness and generalization in sensor noise scenarios, while LiDAR-only methods like PV-RCNN [Shi *et al.*, 2020] are more robust in weather noise scenarios. This observation motivates our research on adaptive fusion schemes for point cloud and image features. Overall, in the KITTI-C [Dong *et al.*, 2023] dataset, our RoboFusion's performance is nearly on par with the clean scene, indicating high level of robustness and generalization.

As shown in Table 4, SOTA methods including PointPillars [Lang *et al.*, 2019], SSN [Zhu *et al.*, 2020], CenterPoint [Yin *et al.*, 2021], FCOS3D [Wang *et al.*, 2021], PGD [Wang *et al.*, 2022a], DETR3D [Wang *et al.*, 2022b], BEVFormer [Li *et al.*, 2022c], FUTR3D [Chen *et al.*, 2023], TransFusion [Bai *et al.*, 2022], BEVFusion[Liu *et al.*, 2023], DeepInteraction [Yang *et al.*, 2022] and CMT [Yan *et al.*, 2023] in nuScenes-C show relatively higher robustness than in KITTI-C when faced with weather noise. However, BEVFusion performs well in the presence of snow, rain, and strong sunlight noise but experiences a significant performance drop in foggy scenarios. In contrast, our method exhibits strong robustness and

[4]https://github.com/open-mmlab/OpenPCDet

[5]https://arxiv.org/abs/2401.03907

Solution	$AP_{3D}(\%)$				$AP_{Weather}(\%)$			
	mAP	Easy	Mod.	Hard	Snow	Rain	Fog	S.L.
Offline	80.41	88.76	77.38	75.11	-	-	-	-
No optim	86.45	91.86	84.80	82.71	45.11	47.77	63.10	79.21
Optim	88.00	92.41	86.77	84.81	57.43	54.27	68.81	82.07

Table 6: Impacts of different SAM usages on **KITTI** and **KITTI-C validation** sets for car class with AP of R_{40}. 'S.L.' denotes Strong Sunlight.

VFM	*Weather*				*Sensor*		
	Snow	Rain	Fog	S.L.	D.	C.O.	C.T.
SAM	57.43	54.27	68.81	82.07	84.21	83.04	84.06
SAM-AD	80.68	81.68	81.67	83.48	84.71	84.17	84.12

Table 7: Influence of pre-training on SAM at **KITTI-C validation** set for car class with AP of R_{40} at moderate difficulty. 'S.L.', 'D.', 'C.O.', and 'C.T.' denotes Strong Sunlight, Density, Cutout, and Crosstalk, respectively.

generalization in both weather and sensor noise scenarios in nuScenes-C.

4.4 Ablation Study

Performance of Different VFMs on RoboFusion

In order to analyze the noise robustness and FPS performance of different-sized VFMs, SAM, FastSAM and MobileSAM, we conduct comparative experiments of RoboFusion-L, RoboFusion-B and RoboFusion-T with SOTA methods, DeepInteraction [Yang *et al.*, 2022] and TransFusion [Bai *et al.*, 2022], on the nuScenes-C [Dong *et al.*, 2023] validation set, as shown in Table 5. Specifically, our RoboFusion exhibits remarkable robustness to weather noise scenarios. Furthermore, our RoboFusion-T has a similar FPS to TransFusion [Bai *et al.*, 2022]. Overall, we have presented a viable application of SAM in 3D object detection tasks.

Impacts of Different SAM Usages

As shown in Table 6, our RoboFusion-L is experimented upon. Specifically, the first row is the offline usage, which involves loading pre-saved image features during training. It implies that certain online data augmentation cannot be utilized. The second (No optim) and the third (Optim) rows are online usages, where the former omits fine-tuning and keeps the model parameters fixed, the latter follows fine-tuning and updating. Therefore, offline usage perform worse than online usages. Additionally, fine-tuning the weights of SAM has demonstrated superior performance, resulting in a performance improvement in the presence of snow, rain, and fog noise scenarios.

Influence of Pre-training on SAM

As shown in Table 7, to investigate the scientific value of pre-trained VFMs like SAM, FastSAM, and MobileSAM in AD scenarios, we conduct our RoboFusion-L with SAM evaluation on SAM and SAM-AD. Through pre-training, SAM-AD has gained a better understanding of AD scenarios than the original SAM. The pre-training strategy effectively improves the performance of our RoboFusion, demonstrating a significant improvement in the snow, rain, and fog noise scenarios.

Method	SAM-AD	AD-FPN	DGWA	A.F.	Snow	Rain	Fog	S.L.	FPS(A100)
a)					34.77	41.30	44.55	80.97	10.8
b)	✓				80.68	81.68	81.67	83.48	4.0
c)	✓	✓			82.32	83.60	82.39	83.98	3.6
d)	✓	✓	✓		83.99	85.63	84.01	84.81	3.4
e)	✓	✓	✓	✓	85.29	86.48	85.53	85.50	3.1

Table 8: Roles of SAM3DFusion-L modules on **KITTI-C validation** set for car class with AP of R_{40} at moderate difficulty. 'A.F.' denotes **Adaptive Fusion** module. 'S.L.' denotes strong sunlight.

Roles of Different Modules in RoboFusion

As shown in Table 8, we present ablation experiments for different modules of our RoboFusion-L, built upon SAM-AD, including AD-FPN, DGWA, and Adaptive Fusion. Leveraging the strong capabilities of SAM-AD in AD scenarios, SAM-AD has a significant improvement from baseline Focals Conv [Chen *et al.*, 2022] (34.77%, 41.30%, 44.55%, 80.97%) to (80.68%, 81.68%, 81.67%, 83.48%). Subsequently, AD-FPN, DGWA, and Adaptive Fusion achieve even higher performance on the foundation of SAM-AD. This further highlights the substantial contributions of diverse modules within our RoboFusion framework in addressing OOD noise scenarios in AD.

5 Conclusions

In this work, we propose a robust framework RoboFusion to enhance the robustness and generalization of multi-modal 3D object detectors using VFMs like SAM, FastSAM, and MobileSAM. Specifically, we pre-train SAM for AD scenarios, yielding SAM-AD. To align SAM or SAM-AD with multi-modal 3D object detectors, we introduce AD-FPN for feature upsampling. To further mitigate noise and weather interference, we apply wavelet decomposition for depth-guided image denoising. Subsequently, we utilize self-attention mechanisms to adaptively reweight fused features, enhancing informative attributes and suppressing excess noises. Extensive experiments demonstrate that our RoboFusion effectively integrates VFMs to boost feature robustness and address OOD noise challenges. We anticipate this work to lay a strong foundation for future research on building robust and dependable foundation AD models.

Limitation and Future Work

First, RoboFusion has a heavy reliance on the representation capability of VFMs. This raises the baseline models' generalization ability, but increases their complexities. Second, the inference speed of RoboFusion-L and RoboFusion-B is relatively slow due to the limitations of SAM and FastSAM. However, the inference speed of RoboFusion-T is competitive with some SOTA methods (e.g. TransFusion) without VFMs. In the future, for improving the real-time application ability of VFMs, we will attempt to incorporate SAM only in the training phase to guide a fast-speed student model, meanwhile explore more noise scenarios.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (2023YJS019), the National Key R&D Program of China (2018AAA0100302).

References

[Bai *et al.*, 2022] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2022.

[Caesar *et al.*, 2020] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[Chen *et al.*, 2022] Yukang Chen, Yanwei Li, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Focal sparse convolutional networks for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5428–5437, 2022.

[Chen *et al.*, 2023] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. FUTR3D: A unified sensor fusion framework for 3D detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 172–181, 2023.

[Deng *et al.*, 2021] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel R-CNN: Towards high performance voxel-based 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1201–1209, 2021.

[Dong *et al.*, 2023] Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, et al. Benchmarking robustness of 3D object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1022–1032, 2023.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE conference on computer vision and pattern recognition*, pages 3354–3361, 2012.

[Hu *et al.*, 2023] Qianjiang Hu, Daizong Liu, and Wei Hu. Density-insensitive unsupervised domain adaption on 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17556–17566, 2023.

[Huang *et al.*, 2020] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. EPNet: Enhancing point features with image semantics for 3D object detection. In *European Conference on Computer Vision*, pages 35–52. Springer, 2020.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[Lang *et al.*, 2019] Alex H Lang, Sourabh Vora, et al. PointPillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019.

[Li *et al.*, 2022a] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.

[Li *et al.*, 2022b] Yanwei Li, Xiaojuan Qi, Yukang Chen, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Voxel field fusion for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1120–1129, 2022.

[Li *et al.*, 2022c] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, et al. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022.

[Li *et al.*, 2023] Xin Li, Tao Ma, Yuenan Hou, Botian Shi, et al. LoGoNet: Towards accurate 3D object detection with local-to-global cross-modal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17524–17534, 2023.

[Lin *et al.*, 2017] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[Liu *et al.*, 2020a] Lin Liu, Jianzhuang Liu, Shanxin Yuan, Gregory Slabaugh, Aleš Leonardis, Wengang Zhou, and Qi Tian. Wavelet-based dual-branch network for image demoiréing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 86–102. Springer, 2020.

[Liu *et al.*, 2020b] Zechen Liu, Zizhang Wu, and Roland Tóth. SMOKE: Single-stage monocular 3D object detection via keypoint estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 996–997, 2020.

[Liu *et al.*, 2023] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023.

[OpenAI, 2023] OpenAI. Gpt-4 technical report. *https://cdn.openai.com/papers/gpt-4.pdf*, 2023.

[Oza *et al.*, 2023] Poojan Oza, Vishwanath A Sindagi, Vibashan Vishnukumar Sharmini, and Vishal M Patel. Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Peng *et al.*, 2023] Xidong Peng, Xinge Zhu, and Yuexin Ma. CL3D: Unsupervised domain adaptation for cross-

LiDAR 3D detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2047–2055, 2023.

[Rukhovich *et al.*, 2022] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. ImVoxelNet: Image to voxels projection for monocular and multi-view general-purpose 3D object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2397–2406, 2022.

[Shi *et al.*, 2019] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3D object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019.

[Shi *et al.*, 2020] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: point-voxel feature set abstraction for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020.

[Song *et al.*, 2023] Ziying Song, Haiyue Wei, Lin Bai, et al. GraphAlign: Enhancing accurate feature alignment by graph matching for multi-modal 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3358–3369, 2023.

[Song *et al.*, 2024a] Ziying Song, Lin Liu, Feiyang Jia, et al. Robustness-aware 3D object detection in autonomous driving: A review and outlook. *arXiv preprint arXiv:2401.06542*, 2024.

[Song *et al.*, 2024b] Ziying Song, Guoxin Zhang, Jun Xie, Lin Liu, et al. Voxelnextfusion: A simple, unified and effective voxel fusion framework for multi-modal 3D object detection. *arXiv preprint arXiv:2401.02702*, 2024.

[Tsai *et al.*, 2023] Darren Tsai, Julie Stephany Berrio, Mao Shan, Eduardo Nebot, and Stewart Worrall. Viewer-centred surface completion for unsupervised domain adaptation in 3D object detection. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9346–9353, 2023.

[Wang *et al.*, 2021] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021.

[Wang *et al.*, 2022a] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022.

[Wang *et al.*, 2022b] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, et al. DERT3D: 3D object detection from multi-view images via 3D-to-2D queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022.

[Wang *et al.*, 2023a] Li Wang, Xinyu Zhang, Ziying Song, et al. Multi-modal 3d object detection in autonomous driving: A survey and taxonomy. *IEEE Transactions on Intelligent Vehicles*, 2023.

[Wang *et al.*, 2023b] Yan Wang, Junbo Yin, Wei Li, et al. SSDA3D: Semi-supervised domain adaptation for 3D object detection from point cloud. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2707–2715, 2023.

[Wu *et al.*, 2023] QuanLin Wu, Hang Ye, Yuntian Gu, et al. Denoising masked autoencoders help robust classification. In *The Eleventh International Conference on Learning Representations*, 2023.

[Xie *et al.*, 2023] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, et al. SparseFusion: Fusing multi-modal sparse representations for multi-sensor 3D object detection. *arXiv preprint arXiv:2304.14340*, 2023.

[Yan *et al.*, 2018] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[Yan *et al.*, 2023] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, et al. Cross modal transformer via coordinates encoding for 3D object dectection. *arXiv preprint arXiv:2301.01283*, 2023.

[Yang *et al.*, 2022] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, et al. DeepInteraction: 3D object detection via modality interaction. *Advances in Neural Information Processing Systems*, 35:1992–2005, 2022.

[Yin *et al.*, 2021] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[Zhang *et al.*, 2022] Yanan Zhang, Jiaxin Chen, and Di Huang. CAT-Det: Contrastively augmented transformer for multi-modal 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 908–917, 2022.

[Zhang *et al.*, 2023a] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, et al. Faster segment anything: Towards lightweight SAM for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

[Zhang *et al.*, 2023b] Dingyuan Zhang, Dingkang Liang, Hongcheng Yang, Zhikang Zou, Xiaoqing Ye, Zhe Liu, and Xiang Bai. SAM3D: Zero-shot 3D object detection via segment anything model. *arXiv preprint arXiv:2306.02245*, 2023.

[Zhang *et al.*, 2023c] Xinyu Zhang, Li Wang, Jian Chen, et al. Dual radar: A multi-modal dataset with dual 4d radar for autononous driving. *arXiv preprint arXiv:2310.07602*, 2023.

[Zhao *et al.*, 2023] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, et al. Fast segment anything, 2023.

[Zhu *et al.*, 2020] Xinge Zhu, Yuexin Ma, Tai Wang, et al. SSN: Shape signature networks for multi-class object detection from point clouds. In *European Conference on Computer Vision*, pages 581–597. Springer, 2020.