A Robust Diffusion Modeling Framework for Radar Camera 3D Object Detection

Zizhang Wu†
Fudan University
wuzizhang87@gmail.com

Yunzhe Wu†
ZongmuTech
nelson.wu@zongmutech.com

Xiaoquan Wang†
ExploAI
rocky.wang@exploai.com

Yuanzhu Gan
ZongmuTech
yuanzhu.gan@zongmutech.com

Jian Pu✉
Fudan University
jianpu@fudan.edu.cn *

Abstract

Radar-camera 3D object detection aims at interacting radar signals with camera images for identifying objects of interest and localizing their corresponding 3D bounding boxes. To overcome the severe sparsity and ambiguity of radar signals, we propose a robust framework based on probabilistic denoising diffusion modeling. We design our framework to be easily implementable on different multi-view 3D detectors without the requirement of using LiDAR point clouds during either the training or inference. In specific, we first design our framework with a denoised radar-camera encoder via developing a lightweight denoising diffusion model with semantic embedding. Secondly, we develop the query denoising training into 3D space via introducing the reconstruction training at depth measurement for the transformer detection decoder. Our framework achieves new state-of-the-art performance on the nuScenes 3D detection benchmark but with few computational cost increases compared to the baseline detectors.

1. Introduction

Aiming at identifying and estimating accurate information about objects' 3D bounding boxes, recent 3D object detectors [14, 41, 53, 54, 61] focus on exploiting methods equipped with different types of sensors to take advantage of their complimentary characteristics. Considering that camera provides rich 2D appearance features, it is usually paired with 3D sensor data, e.g., LiDAR or radar, for providing accurate 3D measurements. In this work, we focus on performing object detection with radar and camera due to its advantages of being **widely implemented**, and its capability in velocity estimation [15, 18, 41], as compared to the expensive LiDARs.

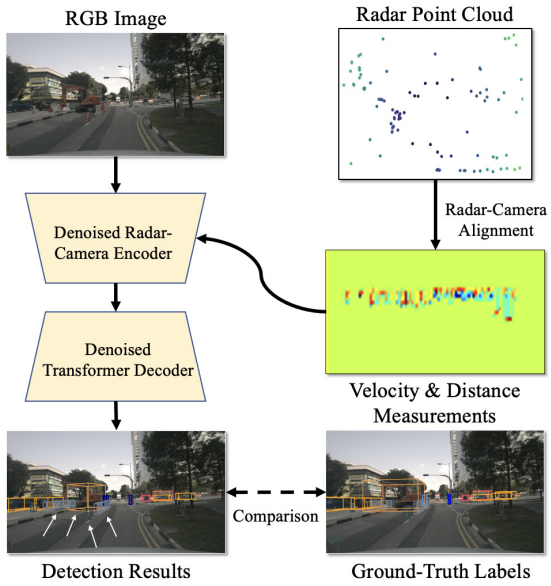

Figure 1. Our denoising diffusion modeling framework proceeds fully differentiable radar feature association with semantic embedding and takes the query denoising training for the transformer decoder at the 3D level. Compared with ground-truth labels, our model predicts quite similar outputs, especially on objects of small size and complex geometries (marked by white arrows).

Despite radar's popularity in the automotive industry, radar point clouds suffer from the drawbacks of severe *sparsity* and *ambiguity* [15, 35, 41], making it unfeasible to use for extracting objects' geometry information; see Figure 1 for a radar sample. Hence, due to the inherent difference between LiDAR and radar, applying existing LiDAR-camera fusion techniques to radar-camera is difficult. To address radar-camera object detection, a pioneer work [28] jointly

*the denotion † means these authors contributed equally and ✉ means the corresponding author.

transforms radar and image features into the same view and concatenates corresponding features for radar-camera fusion. However, direct camera-view transformation results in restrained performances due to the significant semantic loss. Later, [15, 17–19, 39, 41] suggested performing modality fusion in a two-stage manner, where the first-stage camera-only detection results are adopted to associate positive radar features during inference, while ground truth annotations are adopted for the association during training.

Apparently, it inevitably results in asymmetric radar quality during training and inference, and would cause significant radar information loss when the camera detection performance degrades. Thus, FUTR3D [6] proposed to query expanded radar signals with images with the transformer decoder [58]. And the most recent work, CRN [20] proposed to transform image features from perspective views into the BEV level with radar, and has achieved convincing performances in terms of detection accuracy. However, this approach relies on the auxiliary supervision of scarce LiDAR point clouds, which turns it not as easily implementable as pure radar-camera approaches.

To address this issue and further restrain the sparsity and ambiguity natures of the radar sensor, we propose a fully differentiable radar-camera framework with the technique of probabilistic denoising diffusion model in a Rao-Blackwellized fashion [10]. We develop our method as a framework so that it could be easily implemented on different multi-view 3D object detectors, regardless of their technique of obtaining BEV-level information from perspective monocular images. As shown in Figure 1, our framework aligns the radar point clouds with images by the 3D to 2D projection with calibration information. We introduce the designs of information denoising into both the feature encoder and the transformer detection decoder at the BEV level. As a result, we observe that our framework could detect objects in complex geometry, small size and far distance.

In specific, within our denoised radar-camera encoder, we introduce Denoising Diffusion Model (DDM) on aligned radar features followed by the querying of high-level semantic features for feature association. Particularly, we develop the DDM to be aware of the guidance of foreground recognition via introducing it with the embedding of semantic features. We hence point-wisely add the associated positive radar features and image features, and send the output into the transformer decoder. In addition, we introduce the transformer decoder with query denoising at both the 2D and depth levels to further explore the potential of radar-camera association. We conduct extensive experiments to evaluate the robustness and effectiveness of our framework on the large-scale nuScenes [1] benchmark of 3D object detection. We successfully implemented our approach on three representative multi-view 3D detectors and have achieved new state-of-the-art performances. We summarize our contributions as follows:

- We develop an end-to-end differentiable framework for the robust learning of radar-camera 3D object detection based on probabilistic denoising diffusion. Our framework takes no need for LiDAR point clouds for either the training or the inference process.
- We propose to mitigate the ambiguous nature of radar signals via developing a denoising diffusion model with the embedding of semantic features. We also develop the idea of query denoising into 3D space via introducing the reconstruction training at the depth measurement for the transformer decoder [2].
- We successfully implement our framework on three representative multi-view 3D detectors, which take different techniques for the BEV decoding, with few extra costs in terms of computational complexity. We have achieved the new state-of-the-art performance on the nuScenes 3D object detection benchmark [1].

2. Related Work

Camera 3D Object Detection. The main challenge of monocular 3D object detection lies in solving the 2D-3D projection ambiguity caused by the lack of accurate 3D measurements [24, 26, 36]. Based on a single image input, recent approaches adopted geometric constraint regularization [3, 24, 30, 34, 40, 63] and depth estimation interaction [37, 38, 47] to assist 3D object detection. On the other hand, multi-view 3D object detection targets at predicting the objects' 3D bounding boxes and categories by taking multiview images as inputs. We roughly divide current methods into one scheme that lifts 2D to 3D, and the other that queries 2D from 3D.

Inspired by the success of LSS [45], BEVDet [13] and BEVDepth [26] performed the task of 3D detection by lifting multi-view 2D image features into a frustum with depth encoding and creating a unified bird's-eye-view (BEV) feature via flattening the height dimension. Moreover, BEVDet4D [12] incorporated the multi-frame images, and developed a detection pipeline in spatial-temporal 4D working space. Considering further taking advantage of temporal geometric constraints, STS [59], BEVStereo [25] and SOLOFusion [44] were proposed with convincing performances. The other scheme of multiview 3D object detection could be referred to as querying 2D from 3D. Following end-to-end 2D detection pipeline of DETR [2], DETR3D [58] predicted learnable queries in 3D space and queried the corresponding 2D image features via applying the 3D to 2D projection. PETR [31, 32] further generated the queries with 3D positional embedding. BEVFormer [27] explicitly constructed the BEV grid samples in 3D space via leveraging the spatial-temporal deformable attention on BEV features,

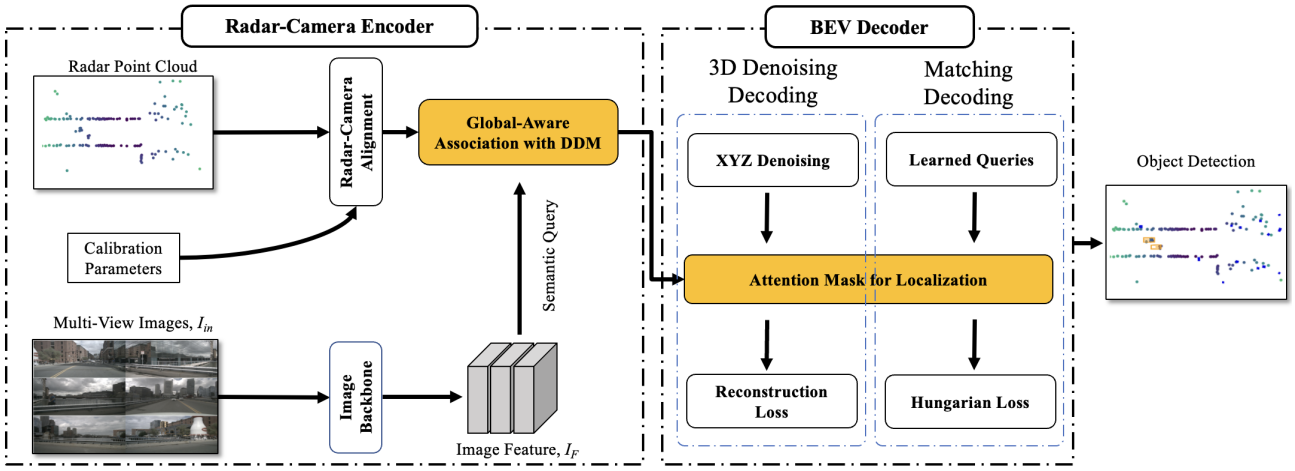

Figure 2. Framework overview. Firstly, within the feature encoder, we align radar features with the images based on calibration information. We develop a Global-Aware Association with the DDM module for the association of positive radar features by the guidance of semantic querying and embedding. We hence send the addition of image and associated radar features into the BEV decoder, where we introduce the querying denoising at the 3D level.

which aggregated multi-view image features by querying for the 2D space.

Radar-Camera 3D Object Detection. Since radar sensors have the advantages of velocity detection, long-range depth measurement, weather robustness, and low-cost implementation, recent works [7, 17, 18, 28, 39, 41] have interacted radar signals with camera sensors on the challenging 3D object detection tasks. Lim et al [28] firstly proposed to jointly transform radar and image features into the same view based on the Cartesian representation of radar signals and the inverse projection mapping of images. However, direct view transformation causes significant semantic loss, thus resulting in restrained performance.

Afterwards, works [17, 18, 39, 41] associated radar information based on 3D region proposals that were first obtained from camera-based detection networks. CenterFusion [41] suggested lifting region proposals into 3D frustums for more precise associations. CramNet [15] proposed to adopt global radar signals with the cross-attention mechanism [55] to assist the depth estimation of foreground objects, but restricted its application scope in circumstances only when radio frequency images are available [49]. CRAFT [19] mitigated the system discrepancy and measurement ambiguities of coordinates by developing a proposal-level early fusion framework with the soft polar association and spatial-contextual fusion transformer. However, their IoU-based association strategy inevitably results in sensitive models as the ground-truth bounding boxes were used for training association but predicted bounding boxes were used for inference association. Afterwards, FUTR3D [6] directly fused expanded radar signals with images based on the transformer decoder [2]. CRN [20] generated the BEV feature maps by transforming image features in perspective view into BEV with radar measurements. To further restrain the sparsity and ambiguity natures of the radar sensor [15, 35], we proposed a radar-camera framework with denoising diffusion modeling. In specific, we embedded the semantic clues into the forward diffusion process to guide the association of positive radar features in a fully differentiable manner.

Denoising Diffusion Model. Recent researches [10, 42, 50] presented the powerful generative Denoising Diffusion Models (DDMs) by learning the gradients of the log data distribution. They leveraged the Langevin dynamics sampling [60] to generate novel samples in a sequence manner, and performed the information denoising starting from a random sample of a standard Gaussian distribution. Following this design, researchers successfully adopted the Denoising Diffusion Models in the learning and sampling of images [10,51], video [11], speech [4,21], etc. Additionally, for the task of text-to-image synthesis, Denoising Diffusion Models have shown significant robustness and generalization abilities, such as the works DALL-E 2 [46] and Imagen [48]. For the task of object detection, DiffusionDet [5] developed the noise-to-box detection paradigm which decouples the training and evaluation stage for dynamic boxes and progressive refinement, while we use DDMs to regularize the sparsity and ambiguity natures of radar features for better incorporation with the image features.

3. Methodology

We develop a fully differentiable denoising framework for the robust learning of 3D object detection, as shown in Figure 2. In particular, we combine radar point clouds

with multi-view images as inputs, so that our network is aware of global 3D circumstances [13, 58]. Following the inference process, our pipeline contains two main components: (i) a radar-camera feature encoder that firstly aligns radar and camera inputs with the calibration information, and performs the fully-differentiable radar-camera association by developing the global-aware attention, and the denoising diffusion model with semantic embedding. (ii) a BEV decoder that jointly decodes the multi-modality fusion features at the BEV level, which adopts the denoising of objects' localization information via introducing extra query groups. We elaborate on the details in the following subsections.

3.1. Radar-Camera Alignment

It is usual to take radar-camera detectors [15, 19, 20] with advanced designs for the encoding of features in each modality for performance boosting. However, a complex architecture is not flexible, and results in cost raises in real industry implementation, which limits the generalization of our radar-camera modality association framework. Thus, we use easily the accessible image backbones [9, 22, 62] for the information encoding of our framework only.

In specific, we provide our framework with N-view input RGB images $I_{in} \in \mathbb{R}^{N\times 3\times H_{in}\times W_{in}}$ with the resolution of $H_{in} \times W_{in}$. We hence extract their image features $I_F \in \mathbb{R}^{N\times C\times H\times W}$, where C is the number of feature channels and $H = \frac{H_{in}}{16}, W = \frac{W_{in}}{16}$. Considering radar signal processing, to alleviate the drawbacks of radar sparsity and its missing of height measurements, we expand radar points along the z-direction following the pillar expansion technique [41]. Particularly, we merge all extended radar point clouds from all radar sensors, which is five on nuScense. We leverage distance and velocity measurements as inputs and aggregate multi-sweep radar points with the ego-motion information for time alignment. Afterwards, we project radar signals onto image planes based on sensor calibrations following the 3D to 2D projection function.

Functionally, we denote the number of radar sweeps as D_r, radar inputs as R_N, intrinsic parameters as $I_{3\times 3}$, extrinsic transformation from radar to LiDAR as T_R^L, LiDAR to camera as T_L^C, and the ego-motion transformation as $T_{t_1}^{t_2}$. High-level radar features R_F is formulated as:

$$R_F = I_{3\times 3} \cdot T_L^C \cdot T_{t_1}^{t_2} \cdot T_R^L \cdot \phi_{exp}(R_{D_r}), \qquad (1)$$

where R_{D_r} indicates the merged radar signals from all five radar sensors, and ϕ_{exp} indicates the pillar expansion operation.

3.2. Global-Aware Association with DDM

We propose a denoising diffusion model with semantic embedding to allocate positive radar features and calculate the long-range dependency between features of different modalities to generate 3D-representative fusion features. Before delving into the details of each module, we first briefly review the process of the attention operation [55]. We denote this operation as ψ_{att}, and its query, key and value projection by convolution and tensor flattening as $P_{q/k/v}(\cdot)$:

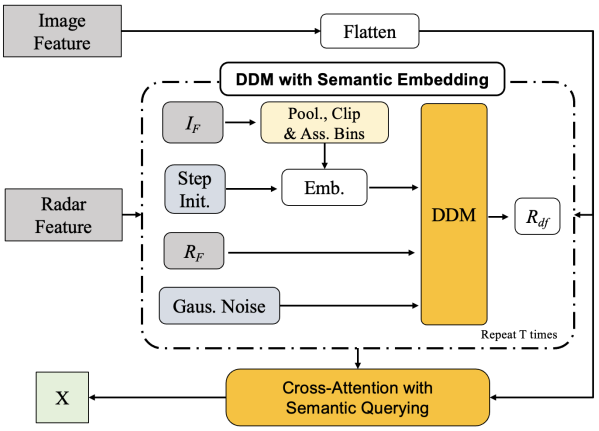

Figure 3. The main architecture of our **Global-Aware Association with DDM**. It first applied the denoising diffusion model with semantic embedding as input to alleviate the ambiguity nature of radar in a fully differentiable manner. Secondly, we apply the cross attention with semantic querying to associate positive radar features.

$$\psi_{att}(q,k,v) = \phi_{dim}(\phi_{soft}(\frac{qk^T}{\sqrt{c}})v), \qquad (2)$$

where ϕ_{soft} indicates the softmax calculation, c indicates the length of the flattened query and key, and ϕ_{dim} reshapes the vector in tensor form.

DDM with Semantic Embedding. The details of our DDM with semantic embedding are shown in Figure 3. During the training of the radar denoising model, we first construct the diffusion process from projected extended radar features to noisy feature maps and then train the model to reverse this process. For detailed explanations for the process of adding Gaussian noises, please refer to [10]. We set the total step number as T for the reverse and diffusion process, and $\epsilon \sim \mathrm{N}(0, I)$ as the corresponding Gaussian noise. We set our variance schedule as $\{\beta_t\}_{t=1}^T$, with $\alpha_t = 1-\beta_t$, and $\bar{\alpha}_t$ refers to the cumulative product from α_1 to α_t. For limiting the increase of computational costs, we design the denoising model as two blocks of light-weight residual connections with layers of 2D convolution, ReLU activation and batch normalization, and we denote it as ϵ_θ. Within the DDM, the embeddings are added to the mapped radar feature. We hence train the diffusion process via optimizing the negative log-likelihood of the designed Markov Chain, that is equivalent to performing gradient descent on:

$$\nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}R_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, \phi_e(I_F))\|^2, \qquad (3)$$

where ϕ_e refers to the embedding of semantic features. In specific, we obtain the semantic embedding via firstly maximum pooling of the image feature, clip features to the interval $(0, 1)$, assign feature values into uniformly discretized K bins, and embed them into dictionaries.

Since that we aim at the denoising of radar features that are inherently noised and ambiguous, we do not follow conventional diffusion processes [5, 21] that start from the features sampled in Gaussian distribution. Instead, we design the inference procedure of our radar DDM as a sampling process from the original radar feature R_F and denoise it to obtain R_{df} under the guidance of semantic embedding. Functionally, we could thus obtain from step T to step 1 as the following:

$$R_{t-1} = \frac{1}{\sqrt{\alpha_t}}(R_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(R_t, t, \phi_e(I_F)) + \sigma_t \mathbf{z}), \quad (4)$$

where σ_t refers to the untrained step dependent constants, $\mathbf{z} \sim N(0, I)$ when $t > 1$, but equals zero when $t = 1$. For more details, please refer to the experiments. Our experiments reveal that the semantic embedding from the image features is important for the refinement of radar features, while direct applying the conventional denoising process will show no contributions to the 3D object detection task.

Association by Semantic Querying. Since our semantic features are learned for detecting foreground objects, their high-level representations are accomplished with enhanced foreground information and could provide meaningful guidance for the denoising of radar signals. As a result, we set projected semantic features as the query and key to allocating positive radar reflections. In function, we have:

$$X = \psi^a_{att}(P^a_q(I_F), P^a_k(I_F), P^a_v(R_{df})), \quad (5)$$

where $X \in \mathbb{R}^{N\times C\times H\times W}$ indicates our associated positive radar features. Afterwards, we implement spatial-wise attention ψ^b_{att} and channel-wise attention ψ^c_{att} on X to obtain spatial-wise feature X_S and channel-wise feature X_C, respectively. We point-wisely add X, X_S, X_C and I_F to obtain the final fusion feature X_{fu}. Specifically, spatial-wise attention operates on the $(H \times W)$ plane, i.e. $F_1 = N \times C, F_2 = H \times W$, to selectively aggregate fusion features across all possible spatial positions. In function, we have:

$$X_S = \psi^b_{att}(P^b_q(X), P^b_k(X), P^b_v(X)^T), \quad (6)$$

with $X_S \in \mathbb{R}^{N\times C\times H\times W}$. While, channel-wise attention operates on the $(N \times C)$ plane, i.e. $F_1 = H \times W, F_2 = N \times C$, to selectively emphasize interdependent fusion channels. In function, we have:

$$(X_C)^T = \psi^c_{att}(P^c_q(X)^T, P^c_k(X)^T, P^c_v(X)), \quad (7)$$

with $X_C \in \mathbb{R}^{N\times C\times H\times W}$. We thus formulate our final fusion feature as follows:

$$X_{fu} = X + \gamma_S X_S + \gamma_C X_C + I_F, \quad (8)$$

where γ_S and γ_C indicate hyper-parameters for spatial-channel balancing.

3.3. BEV Decoder with Localization Denoising

We apply the BEV decoder for the task of 3D object detection. We experimented our framework on three representative multi-view 3D detection baselines, namely BEVDet [13], PETR [31], and BEVFormer [27] which adopt different BEV decoding techniques. Particularly, as shown in the second part of Figure 2, inspired by the usage of denoising training in DN-DETR [23] in 2D object detection, for the NMS-free transformer decoder [2] in PETR, we propose the auxiliary training of query denoising for the regression of 3D bounding box centers, i.e, denoising modeling on the 'XYZ' measurement values.

In specific, despite the learnable queries which are trained to match by the Hungarian loss [52], we introduce the decoder with D denoising query groups, that are obtained from the sampling of object labels with Gaussian noises. In specific, we set N_D queries to each of the D groups, which is selected to be larger than the most possible number of objects of interest within a 3D circumstance. We experimented to find that adding noises to other features of interest shows no significant contributions to the model performance. Besides, we train the attention operations with the masking of parameters for the denoising queries following the design of DN-DETR [23], and only the regular learnable queries are used to decode for 3D bounding boxes. Functionally, we denote our initialized object queries as Q_0, initialized noised localization queries as Q^{xyz}_0, and our ith layer of the transformer decoder as ϕ_i. Hence, we formulate our transformer decoder as follows:

$$Q_{i+1} = \phi_i(\psi_{\text{enc}}(X_{fu}), Q_i, Q^{xyz}_i), \quad i = 1, ..., L \quad (9)$$

where L indicates the total number of decoder layers, and ψ_{enc} indicates the projection of fusion feature.

3.4. Denoising Framework Loss Function.

The denoising query groups interact with the fusion features as the regular learnable queries, but are trained with the direction regression of bounding box localization information without Hungarian matching [52], as they are initialized based on one-to-one matching during the denoising query preparation. Thus, the 3D detection loss of BEV decoder could be formulated as the following:

$$\mathcal{L} = \gamma_1 \mathcal{L}_{DDM} + \gamma_2 \mathcal{L}_{reg} + \gamma_3 \mathcal{L}_{cls} + \gamma_4 \mathcal{L}_{xyz}, \quad (10)$$

Method	Modality	Backbone	NDS↑	mAP↑	mAVE↓	mATE↓	mASE↓	mAOE↓	mAAE↓
BEVDet [13]	C	V2-99*	48.8	42.4	95.0	52.4	24.2	37.3	14.8
BEVDet+DR	R+C	V2-99*	54.8	43.9	46.3	51.2	24.1	36.7	13.5
Improvements	-	-	**+6.0**	**+1.5**	**+43.7**	**+1.2**	**+0.1**	**+0.6**	**+ 1.3**
PETR [31]	C	V2-99*	50.4	44.1	80.8	59.3	24.9	38.3	13.2
PETR+DR	R+C	V2-99*	55.4	45.9	45.5	55.1	24.9	38.1	12.9
Improvements	-	-	**+4.0**	**+1.8**	**+35.3**	**+4.2**	**+0.0**	**+0.2**	**+0.3**
BEVFormer [27]	C	V2-99*	56.9	48.1	37.8	58.2	25.6	37.5	12.6
BEVFormer+DR	R+C	V2-99*	59.4	50.2	31.3	52.3	24.5	36.6	11.7
Improvements	-	-	**+2.5**	**+2.1**	**+6.5**	**+5.9**	**+1.1**	**+0.9**	**+0.9**

Table 1. Quantitative comparisons on the nuScense testing split. **DR** refers to our **d**enoising **r**adar-camera framework. Our improvements relative to baseline multi-view models are listed with +. * notes that VoVNet-99 [22] was pre-trained with extra data [43].

Method	Modality	Backbone	LiDAR	NDS↑	mAP↑	mAVE↓	mATE↓	mASE↓	mAOE↓	mAAE↓
BEVDepth [26]	C	V2-99*	✓	60.0	50.3	32.0	44.5	24.5	37.8	12.6
BEVStereo [25]	C	V2-99*	✓	61.0	52.5	35.7	43.1	24.6	35.8	13.8
CRN [20]	R+C	ConvNextB + PointNet	✓	62.4	57.5	36.5	41.6	26.4	45.6	13.0
FCOS3D [56]	C	R101		42.8	35.8	143.4	69.0	24.9	45.2	12.4
CenterFusion [41]	R+C	DLA34		44.9	32.6	61.4	63.1	26.1	51.6	11.5
DETR3D [58]	C	V2-99*		47.9	41.2	84.5	64.1	25.5	39.4	13.3
BEVDet [13]	C	V2-99*		48.8	42.4	95.0	52.4	24.2	37.3	14.8
PETR [31]	C	V2-99*		50.4	44.1	80.8	59.3	24.9	38.3	13.2
CRAFT [19]	R+C	DLA34 + PointNet		52.3	41.1	51.9	46.7	26.8	45.6	11.4
BEVDet4D [12]	C	Swin-B		56.9	45.1	30.1	**51.1**	**24.1**	38.6	12.1
BEVFormer [27]	C	V2-99*		56.9	48.1	37.8	58.2	25.6	37.5	12.6
PolarFormer [16]	C	V2-99*		57.2	49.3	44.0	55.6	25.6	**36.4**	12.7
FrustumFormer [57]	C	V2-99*		58.9	**51.6**	38.9	55.5	24.9	37.2	12.6
Ours	R+C	DLA34		53.8	44.2	45.6	60.7	25.0	39.4	12.7
	R+C	V2-99*		**59.4**	50.2	**31.3**	52.3	24.5	36.6	**11.7**

Table 2. Comparisons to state-of-art on the nuScenes testing set. * notes that VoVNet-99 [22] was pre-trained with extra data [43]. LiDAR refers to training depth estimation via extra modality supervision from LiDAR.

where $\mathcal{L}_{cls}$ indicates the focal loss [29] that balances the sample for classification, $\mathcal{L}_{reg}$ indicates the L1 loss that regresses the 3D bounding box information, $\mathcal{L}_{DDM}$ indicates the gradient descent optimization for our radar DDM, and $\mathcal{L}_{xyz}$ indicates the loss of the reconstruction loss [23]. We set hyper-parameter γ_4 as zero when we do not experiment with the 3D transformer decoder [31, 58].

4. Experiments and Results

4.1. Dataset and Evaluation Metrics

We evaluate our network on the challenging and large-scale nuScenes dataset [1], which contains radar, LiDAR and multi-camera data with annotated 3D bounding boxes. It considers 10 categories for metric comparison. We conduct experiments following its official data spits, which contain 700 scenes for training, 150 scenes for validation and 150 scenes for testing. For the network analysis, we experimented with the validation split. Following prior works [19,31,41,58], we adopt the nuScenes detection score (NDS) and mean average precision (mAP) as the main metrics for performance comparison. NDS is calculated as a weighted sum of mAP, and one minus each of the five other mean average errors, namely velocity (mAVE), translation (mATE), scale (mASE), orientation (mAOE) and attributes (mAAE). We assign ↑ and ↓ to metrics that are expected to be higher and lower for better performance, respectively.

4.2. Implementation Details

We take three sweeps of radar signals ($D_r = 3$), and six transformer decoding layers ($L = 6$) for our denoised transformer decoder. We set $\gamma_S = 1.0$, $\gamma_C = 0.5$ for spatial-channel balancing, set $\gamma_{cls} = 2.0$, $\gamma_{reg} = 1.0$ for classification and regression balancing, and set $\gamma_2 = 2.0$, $\gamma_3 = 1.0$ for classification and regression balancing.

For our denoising designs, we experimented to find that

Method	Modality	Backbone	LiDAR	NDS↑	mAP↑	mAVE↓	mATE↓	mASE↓	mAOE↓	mAAE↓
BEVDepth [26]	C	R101	✓	53.5	41.2	33.1	56.5	26.6	35.8	19.0
STS	C	R101	✓	54.2	43.1	36.9	52.5	26.2	38.0	20.4
CRN [20]	R+C	R101 + PointNet	✓	59.2	52.5	35.2	46.0	27.3	44.3	18.0
FCOS3D [56]	C	R101†		37.2	29.5	131.5	80.6	26.8	51.1	17.0
CenterFusion [41]	R+C	DLA34		45.3	33.2	54.0	64.9	26.3	53.5	**14.2**
DETR3D [58]	C	R101†		43.4	34.9	84.2	71.6	26.8	37.9	20.0
PETR [31]	C	R101†		44.2	37.0	86.5	71.1	26.7	38.3	20.1
CRAFT [19]	R+C	DLA34 + PointNet		51.7	41.1	48.6	49.4	27.6	45.4	17.6
BEVFormer [27]	C	R101†		51.7	41.6	40.9	64.8	27.0	34.8	19.8
PolarFormer [16]	C	R101†		52.8	43.2	40.9	64.8	27.0	**34.8**	20.1
FrustumFormer [57]	C	R101†		54.6	**45.7**	38.0	62.4	26.5	36.2	19.1
Ours	R+C	DLA34		52.7	43.9	46.3	**60.8**	**25.4**	41.8	18.5
Ours	R+C	R101†		**55.0**	45.2	**34.1**	61.3	26.6	35.9	18.3

Table 3. Comparisons to state-of-art on the nuScenes validation set. † indicates backbones initialized from the training of FCOS3D [56]. LiDAR refers to training depth estimation via extra modality supervision from LiDAR.

	Denoised Encoder			Denoised Decoder		NDS↑	mAP↑	mAVE↓
	Attention	DDM	Sem. Emb.	2D-DN	Depth-DN			
(i)	—	—	—	—	—	40.3	33.9	90.7
(ii)	✓	—	—	—	—	43.1	33.7	51.8
(iii)	✓	✓	—	—	—	43.9	34.2	50.1
(iv)	✓	✓	✓	—	—	45.5	35.1	47.2
(v)	✓	✓	✓	✓	—	46.4	35.6	45.5
(vi)	✓	✓	✓	✓	✓	**47.0**	**36.2**	**45.2**

Table 4. Analysis of our denoising radar-camera framework on the nuScenes validation split. 'Attention' refers to the cross-attention with the semantic querying operation, and the joint of the channel- and special-wise self-attention on associated positive radar feature. 'DDM' refers to our denoising diffusion model without semantic embedding. 'Sem. Emb.' refers to our designed semantic embedding for DDM. '2D-DN' refers to directly applying the query denoising from DN-DETR on a 2D plane. 'Depth-DN' refers to query denoising on depth.

setting $\gamma_1 = 0.5$ and $\gamma_4 = 0.1$ generates the detector with the best performance. We set $\gamma_4 = 0$ if the baseline multi-view model does not adopt the transformer decoder [2, 58]. For our DDM on radar features, we set the total step $T = 5$, and the bin size is set to 10. For our query denoising training, we set the denoising groups as two, and set fifty queries to each of them. The total number of queries for decoding is set as 900. For experiments to compare against the SOTA works, the image resolution is set as 1600×900. While for the network analysis, the resolution is set as 1408×512. Settings for optimizers, batch sizes and numbers of GPUs used follow those for baseline models [13, 27, 31].

We train our models with the VoVNetV2(V2-99) [22], DLA34 [8] and ResNet [9] backbones to evaluate against existing state-of-the-art multi-camera and radar-camera detectors [27, 31] on the testing and validation splits. Particularly, our framework requires no extra supervision from the expensive LiDAR sensor and takes no extra backbone for the feature encoding of radar signals.

4.3. Comparisons with Baselines

To evaluate the effectiveness of our radar-camera denoising framework, we implemented it on three multi-view baselines and re-trained each multi-modality network model on the nuScense 3D object detection dataset. The detection performances are reported in Table 1. For easy observation, the percentages of improvements in overall metrics are shown in bold. Clearly, equipped with our denoising radar-camera fusion framework, the performance of all three multi-modality models is improved significantly on the important NDS and mAP metrics. Particularly, we observe clear performance boosts on the regression of objects' velocities, i.e., on mAVE. It shows that our framework can successfully associate and transfer precise depth and velocity knowledge from radar measuring to the baseline models.

4.4. Comparison with the State-of-the-Arts

We evaluate our pipeline against state-of-the-art multi-view and radar-camera methods in Table 2 on the testing set.

	NR-Init.	R-Init.	NDS↑	mAP↑	mAVE↓
(i)	—	—	44.1	34.7	48.8
(ii)	✓	—	43.7	34.3	49.2
(iii)	—	✓	**47.0**	**36.2**	**45.2**
(iv)	✓	✓	46.5	35.6	46.9

Table 5. Analysis of DDM with semantic embedding on nuScenes validation set. 'NR-Init.' refers to starting the diffusion process conventionally from noised radar features. 'R-Init.' refers to the start of the diffusion process from direct projected extended radar.

	x	v_{xy}	rcs	NDS↑	mAP↑	mAVE↓
(i)	—	—	—	40.3	33.9	90.7
(ii)	✓	—	—	41.9	36.5	86.7
(iii)	✓	✓	—	**47.0**	**36.2**	**45.2**
(iv)	✓	✓	✓	46.8	37.4	48.5

Table 6. Analysis of radar characteristics on nuScenes validation set. 'x', 'v_{xy}' and 'rcs' indicate the distance, two-direction radial velocity, and radar cross-section measurements, respectively.

Among methods that take no LiDAR measurements auxiliary supervisions, our denoising radar-camera pipeline (experimented on BEVFormer [27]) ranks first place in terms of NDS, mAVE rankings, while second place in terms of mAP ranking. Compared to the two-stage radar-camera detectors (CenterFusion [41] and CRAFT [19]), our model with DLA34 [8] surpasses them by a large margin.

While comparing to the CRN [20], our method takes no extra supervision from the LiDAR, which is crucial as radar sensors have been widely deployed to the production of vehicles, instead of the expensive LiDAR. Besides, although ConvNextB [33] is considered as a stronger backbone, our model on V2-99 [22] still shows competitive results in terms of NDS, and stronger performance in terms of mAVE against CRN [20]. We also compared the state-of-the-art on the validation set in Table 3. We experimented with the backbone of R101(initialized from the training of FCOS3D [56]), and with DLA34 to make fair comparisons against CenterFusion [41] and CRAFT [19].

4.5. Network Analysis

In this section, we experiment with the performance of our framework on PETR [31] with the backbone of ResNet50 [7] to make detailed explorations about the effectiveness of our designs on the nuScenes validation set.

Analysis of our denoising radar-camera framework. To validate the effectiveness of our denoised decoder and denoised encoder, we conducted ablation studies and the results are summarized in Table 4. The first line model refers to the performance of our baseline. The second line results are obtained from directly sending aligned radar features into the cross-attention with semantic querying, followed by the dual-attention operation on the associated positive radar features. For the third line, we introduce the network to the DDM model with the step embedding only. For the fourth line, radar DDM with our semantic embedding. For the fifth line, we introduce the transformer decoder with 2D-level query denoising training following DN-DETR. For the last line, we further extend the 2D-level query denoising training to the 3D level by developing the denoising process on depth measurements. We observe that both the radar feature DDM with semantic embedding, and the 3D-level query denoising could bring significant performance improvements,

Analysis of DDM with semantic embedding. The analysis is shown in Table 5. For the first line, we experiment by dropping the radar feature DDM with semantic embedding, while the query denoising training within the BEV decoder is included. For the last line, we experiment by sending both the noised radar and original radar into our DDM and point-wisely add their denoised results. Our experiments once more approve the ambiguous nature of radar sensors. It also shows that denoising on raw radar features with semantic embedding could benefit the multi-modality feature association process for the task of 3D object detection.

Analysis of radar characteristics. The analyais is shown in Table 6. It shows that distance (x) and velocity measurements (v_{xy}) could benefit 3D detection, while radar cross-section measurement (rcs) brings no convincing improvements to the NDS metric. The first line refers to our baseline, which takes image inputs only.

5. Conclusion

In this work, we propose a robust denoising framework for the task of radar-camera 3D objection. An end-to-end fully-differentiable framework that adopts the DDM with semantic embedding to association radar responses, and adopts the 3D-level query denoising training for the decoding of bounding boxes under the BEV view. We found that the denoising diffusion model with guidance from semantic information could effectively mitigate the ambiguous nature of radar sensors. Our framework takes no usage of LiDAR point clouds during either the inference or the training process, which is important as expensive LiDAR sensors are not as widely implemented onto vehicles as radar or camera. Our framework is flexible and could bring significant performance improvements for major multi-view 3D detectors that take different techniques for BEV-level decoding. Our framework turns multi-view detectors into robust multi-modality radar-camera detectors with significant performance gains on the nuScense [1] 3D detection benchmark and also causes a limited increase in terms of computational costs.

References

[1] 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 *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631. IEEE, 2020. 2, 6, 8

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conf. on Computer Vision (ECCV)*, pages 213–229. Springer, 2020. 2, 3, 5, 7

[3] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. MonoRUN: Monocular 3d object detection by reconstruction and uncertainty propagation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10379–10388, 2021. 2

[4] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. WaveGrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020. 3

[5] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022. 3, 5

[6] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. FUTR3D: A unified sensor fusion framework for 3d detection. *arXiv preprint arXiv:2203.10642*, 2022. 2, 3

[7] Florian Drews, Di Feng, Florian Faion, Lars Rosenbaum, Michael Ulrich, and Claudius Gläser. DeepFusion: A robust and modular 3d object detector for lidars, cameras and radars. *arXiv preprint arXiv:2209.12729*, 2022. 3, 8

[8] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2002–2011, 2018. 7, 8

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4, 7

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 2, 3, 4

[11] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 3

[12] Junjie Huang and Guan Huang. BEVDet4D: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 2, 6

[13] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. BEVDet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2, 4, 5, 6, 7

[14] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H. Hsu. MonoDTR: Monocular 3d object detection with depth-aware transformer. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[15] Jyh-Jing Hwang, Henrik Kretzschmar, Joshua Manela, Sean Rafferty, Nicholas Armstrong-Crews, Tiffany Chen, and Dragomir Anguelov. CramNet: Camera-radar fusion with ray-constrained cross-attention for robust 3d object detection. In *European Conf. on Computer Vision (ECCV)*, pages 388–405. Springer, 2022. 1, 2, 3, 4

[16] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. PolarFormer: Multi-camera 3d object detection with polar transformers. In *AAAI Conf. on Artificial Intell. (AAAI)*, 2023. 6, 7

[17] Jinhyeong Kim, Youngseok Kim, and Dongsuk Kum. Low-level sensor fusion for 3d vehicle detection using radar range-azimuth heatmap and monocular image. In *Asian Conference on Computer Vision*, pages 388–402. Springer, 2020. 2, 3

[18] Youngseok Kim, Jun Won Choi, and Dongsuk Kum. GRIF Net: Gated region of interest fusion network for robust 3d object detection from radar point cloud and monocular image. In *IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 10857–10864. IEEE, 2020. 1, 2, 3

[19] Youngseok Kim, Sanmin Kim, Jun Won Choi, and Dongsuk Kum. CRAFT: Camera-radar 3d object detection with spatio-contextual fusion transformer. *arXiv preprint arXiv:2209.06535*, 2022. 2, 3, 4, 6, 7, 8

[20] Youngseok Kim, Sanmin Kim, Juyeb Shin, Jun Won Choi, and Dongsuk Kum. CRN: Camera radar net for accurate, robust, efficient 3d perception. *arXiv preprint arXiv:2304.00670*, 2023. 2, 3, 4, 6, 7, 8

[21] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 3, 5

[22] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13906–13915, 2020. 4, 6, 7, 8

[23] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-DETR: Accelerate detr training by introducing query denoising. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13619–13627, 2022. 5, 6

[24] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. RTM3D: Real-time monocular 3d detection from object keypoints for autonomous driving. In *European Conf. on Computer Vision (ECCV)*, pages 644–660. Springer, 2020. 2

[25] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. BEVStereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. *arXiv preprint arXiv:2209.10248*, 2022. 2, 6

[26] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 2, 6, 7

[27] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BevFormer:

Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conf. on Computer Vision (ECCV)*, pages 1–18. Springer, 2022. 2, 5, 6, 7, 8

[28] Teck-Yian Lim, Amin Ansari, Bence Major, Daniel Fontijne, Michael Hamilton, Radhika Gowaikar, and Sundar Subramanian. Radar and camera early fusion for vehicle detection in advanced driver assistance systems. In *Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems*, volume 2, page 7, 2019. 1, 3

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2980–2988, 2017. 6

[30] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1810–1818, 2022. 2

[31] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3D object detection. In *European Conf. on Computer Vision (ECCV)*, pages 531–548. Springer, 2022. 2, 5, 6, 7, 8

[32] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETRv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022. 2

[33] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[34] Zongdai Liu, Dingfu Zhou, Feixiang Lu, Jin Fang, and Liangjun Zhang. Autoshape: Real-time shape-aware monocular 3d object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 15641–15650, 2021. 2

[35] Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, and Praveen Narayanan. Radar-camera pixel depth association for depth completion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12507–12516, 2021. 1, 3

[36] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3111–3121, 2021. 2

[37] Xinzhu Ma, Shinan Liu, Zhiyi Xia, Hongwen Zhang, Xingyu Zeng, and Wanli Ouyang. Rethinking pseudo-lidar representation. In *European Conf. on Computer Vision (ECCV)*, pages 311–327. Springer, 2020. 2

[38] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6851–6860, 2019. 2

[39] Michael Meyer and Georg Kuschk. Deep learning based 3d object detection for automotive radar and camera. In *2019 16th European Radar Conference (EuRAD)*, pages 133–136. IEEE, 2019. 2, 3

[40] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7074–7082, 2017. 2

[41] Ramin Nabati and Hairong Qi. CenterFusion: Center-based radar and camera fusion for 3d object detection. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 1527–1536, 2021. 1, 2, 3, 4, 6, 7, 8

[42] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Int. Conf. on Machine Learning (ICML)*, pages 8162–8171. PMLR, 2021. 3

[43] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 3142–3152, 2021. 6

[44] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *Int. Conf. on Learning Representations (ICLR)*, 2023. 2

[45] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conf. on Computer Vision (ECCV)*, pages 194–210. Springer, 2020. 2

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[47] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8555–8564, 2021. 2

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. 3

[49] Marcel Sheeny, Emanuele De Pellegrin, Saptarshi Mukherjee, Alireza Ahrabian, Sen Wang, and Andrew Wallace. Radiate: A radar dataset for automotive perception in bad weather. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2021. 3

[50] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Int. Conf. on Machine Learning (ICML)*, pages 2256–2265. PMLR, 2015. 3

[51] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Int. Conf. on Learning Representations (ICLR)*, 2021. 3

[52] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2325–2333, 2016. 5

[53] Yunlei Tang, Sebastian Dorn, and Chiragkumar Savani. Center3d: Center-based monocular 3d object detection with joint depth understanding. In *DAGM German Conference on Pattern Recognition*, pages 289–302. Springer, 2020. 1

[54] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6411–6420, 2019. 1

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3, 4

[56] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 913–922, 2021. 6, 7, 8

[57] Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. FrustumFormer: Adaptive instance-aware resampling for multi-view 3d detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6, 7

[58] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2, 4, 6, 7

[59] Zengran Wang, Chen Min, Zheng Ge, Yinhao Li, Zeming Li, Hongyu Yang, and Di Huang. STS: Surround-view temporal stereo for multi-view 3d detection. *arXiv preprint arXiv:2208.10145*, 2022. 2

[60] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Int. Conf. on Machine Learning (ICML)*, pages 681–688, 2011. 3

[61] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 11784–11793, 2021. 1

[62] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2403–2412, 2018. 4

[63] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2