Robust Single-Stage Fully Sparse 3D Object Detection via Detachable Latent Diffusion

Wentao Qu[1], Guofeng Mei[2], Jing Wang[3], Yujiao Wu[4], Xiaoshui Huang[5*], Liang Xiao[1*]

[1]Nanjing University of Science and Technology, China
[2]Fondazione Bruno Kessler, Italy
[3]HiDream.ai, China
[4]Commonwealth Scientific and Industrial Research Organisation, Australia
[5]Shanghai Jiao Tong University, China
quwentao@njust.edu.cn, huangxiaoshui@163.com, xiaoliang@mail.njust.edu.cn

Abstract

Denoising Diffusion Probabilistic Models (DDPMs) have shown success in robust 3D object detection tasks. Existing methods often rely on the score matching from 3D boxes or pre-trained diffusion priors. However, they typically require multi-step iterations in inference, which limits efficiency. To address this, we propose a **R**obust single-stage fully **S**parse 3D object **D**etection **Net**work with a Detachable Latent Framework (DLF) of DDPMs, named RSDNet. Specifically, RSDNet learns the denoising process in latent feature spaces through lightweight denoising networks like multi-level denoising autoencoders (DAEs). This enables RSDNet to effectively understand scene distributions under multi-level perturbations, achieving robust and reliable detection. Meanwhile, we reformulate the noising and denoising mechanisms of DDPMs, enabling DLF to construct multi-type and multi-level noise samples and targets, enhancing RSDNet robustness to multiple perturbations. Furthermore, a semantic-geometric conditional guidance is introduced to perceive the object boundaries and shapes, alleviating the center feature missing problem in sparse representations, enabling RSDNet to perform in a fully sparse detection pipeline. Moreover, the detachable denoising network design of DLF enables RSDNet to perform single-step detection in inference, further enhancing detection efficiency. Extensive experiments on public benchmarks show that RSDNet can outperform existing methods, achieving state-of-the-art detection.

1 Introduction

Advances in 3D hardware and data synthesis technologies have made large-scale scene point clouds increasingly accessible. Reliably locating and recognizing targets in large-scale scenes, especially under various real-world noise, is crucial for real-time downstream tasks, such as autonomous driving (Lian et al. 2023), AR/VR (Li et al. 2025), and robotics (Liu et al. 2024). Therefore, ***robust and efficient 3D object detection*** has attracted increasing attention.

In recent years, single-stage fully sparse 3D detection pipelines, built on hybrid architectures with 3D and 2D backbones, have become mainstream, exhibiting the impressive detection results (Fan et al. 2024; Zhang et al. 2024, 2025; Liu et al. 2025). However, due to environmental or sensor factors, raw point clouds from 3D devices are often contaminated by multiple perturbations, such as point-level random noise and global geometric distortions (coordinate offset, scaling and rotation) (Rakotosaona et al. 2020; Ding et al. 2024; Lian et al. 2025). While focused on efficiency and accuracy, these methods often lack robustness to perturbations, making them less reliable for stable performance in real-world scenarios (Xiang, Qi, and Li 2019).

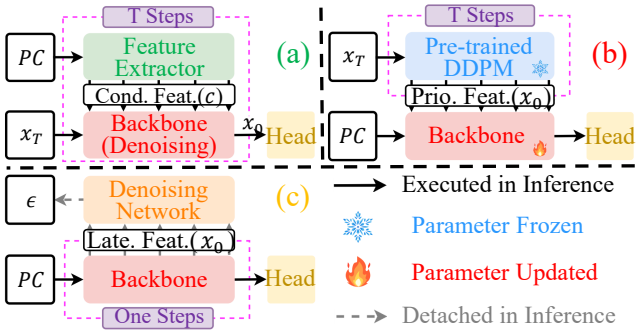

Figure 1: Existing DDPM paradigms in 3D object detection: (a) performs denoising to estimate the box scores, then refines them via a detection head. (b) introduces a pre-trained diffusion prior to enhance detection accuracy. (c) (DLF) conducts the denoising learning via a denoising network. (a) and (b) are multi-step, while (c) demand single-step inference.

Along a research line, DDPMs (Ho, Jain, and Abbeel 2020), with a noise-robust architecture, have shown significant potential in robust object detection tasks (Pellicer, Li, and Angelov 2024; Chu et al. 2025). These methods typically estimate the scores from bounding boxes, refining them via a detection head (see Fig. 1(a)) (Chen et al. 2023a; Ho et al. 2023; Chen et al. 2024a; Ranasinghe, Hegde, and Patel 2024). Alternatively, they introduce the pre-trained diffusion priors into the pipeline, improving the detection accuracy (see Fig. 1(b)) (Xu et al. 2024; Zheng et al. 2024).

However, to obtain high-quality bounding boxes and feature priors, they inevitably require multi-step iterations to accurately match the scores in inference (Song, Meng, and Ermon 2020). This requirement limits their practicality in

* Corresponding author.

downstream tasks with strict real-time constraints, such as autonomous driving. Moreover, DDPMs that traditionally model Gaussian distributions struggle to be robust to other types of perturbations (Qu et al. 2024, 2025).

These observations raise a central question: ***Can DDPMs overcome multi-step inference while modeling multiple perturbations?*** To answer this, we reveal the robustness source and rethink noising and denoising rules, providing new insights into DDPMs for 3D Object Detection (3DOD).

Inspired by these core insights, we design a Detachable Latent Framework (DLF, see Fig. 1(c)) of DDPMs, effectively overcoming multi-step inference and preserving multiple noise robustness. Unlike Fig. 1(a) and Fig. 1(b), DLF guides the model to learn the denoising process via the denoising network in the latent feature space, but detached in inference. In this manner, DLF still inherits the DDPM training pattern, thus preserving noise robustness. Meanwhile, the detachable design for the denoising network avoids the multi-step inference. This also relaxes the score matching requirement, as the task result lies in the model backbone rather than the denoising network. Furthermore, we reformulate the noising and denoising mechanisms, enabling DLF to construct multi-type and multi-level noise samples and targets in training, thereby making the model robust to multiple perturbations. Moreover, DLF, supported by conditional guidance, can inject the task-specific knowledge priors for the model, enhancing the understanding for the task.

Furthermore, we propose a **R**obust single-step fully **S**parse 3D object **D**etection **Net**work based on DLF, called RSDNet. Specifically, RSDNet treats the denoising networks as multi-level denoising autoencoders (DAEs) (Xiang et al. 2023; Chen et al. 2024b). This uses two lightweight denoising U-Nets (<6M), 3D Denoising U-Net (3DDU) and 2D Denoising U-Net (2DDU), guiding the 3D and 2D backbones to perform the denoising learning in a supervised manner. In this way, the backbones can understand the scene context in the multi-type and multi-level denoising learning, generating robust and generalized object-aware features for the detection head. Meanwhile, a semantic-geometric conditional guidance is injected to effectively mitigate the ***center feature missing*** problem caused by downsampling or sparse convolution (Zhang et al. 2024), enabling RSDNet to operate within an efficient fully sparse detection pipeline (3D and 2D sparse backbones). Moreover, thanks to the detachable design of DLF, RSDNet can achieve one-step inference.

Our key contributions can be summarized as:

- We systematically reveal the robustness source and analyze the noising and denoising mechanisms, offering new knowledge for application of DDPMs in 3D tasks.
- We design a detachable latent framework of DDPMs, which can overcome multiple iterations in inference while preserving multiple noise robustness in training.
- We propose a robust single-step fully sparse detection network based on DLF, enabling one-step detection with strong robustness to multiple perturbations.
- Comprehensive experiments demonstrate that RSDNet can achieve robust and significant detection performance.

2 Related Works

Learnable 3D Object Detection. Benefiting from deep learning techniques, a lot of learnable 3D object detection methods have achieved success. Early methods typically employ dense voxels and multi-view 2D projections to establish ordered 3D representations within the detection pipeline (Chen et al. 2017; Zhou and Tuzel 2018). However, these methods either incur significant computational overhead due to numerous empty voxels or suffer from the loss of geometric details caused by object occlusion. PIXOR (Yang, Luo, and Urtasun 2018) stands as the pioneer to convert point clouds into the 2D Bird's-Eye View (BEV) representations, enabling 2D dense convolution effectively transferring to 3D object detection. Subsequently, some researchers propose coarse-to-fine detection pipelines based on BEV, further improving detection accuracy (Shi, Wang, and Li 2019; Yang et al. 2019). However, two-stage detection introduces significant cost in inference. For efficient detection pipelines, some methods implement single-stage detection via hybrid 3D sparse and 2D dense backbones (Bai et al. 2022; Zhang et al. 2023). Recently, fully sparse pipelines show efficiency and effectiveness, further advancing 3D detection development (Zhang et al. 2024; Liu et al. 2025).

Although existing methods have demonstrated excellent detection performance, they overlook the fact that raw point clouds from 3D sensors are often perturbed. This makes them typically sensitive to noise, limiting their practical application. To address this issue, we introduce DDPMs into 3D object detection through a detachable latent framework. This guides the backbone to generate noise-robust features through the learning multi-level denoising process, while the detachability avoids extra computational cost in inference.

DDPMs for Object Detection. Some methods have explored DDPMs in robust detection tasks, exhibiting reliable and stable performance. DiffusionDet (Chen et al. 2023a) marks the first integration of DDPMs as the fundamental mechanism for 2D object detection, displacing conventional query- and anchor-based paradigms. This generates initial 2D bounding boxes via an iterative denoising process, followed by refinement of object locations and semantics using a detection head. Building upon this paradigm, subsequent methods have achieved notable improvements in the robustness and accuracy of 2D object detection (Chen, Sun, and Lin 2024; Wang, Jia, and Dai 2024). Inspired by the success in 2D detection, 3D object detection methods have adapted DDPMs to learn the scores from 3D bounding boxes, achieving robust and impressive performance (Ho et al. 2023; Chen et al. 2024a; Ranasinghe, Hegde, and Patel 2024). Furthermore, recent studies have explored leveraging DDPMs as pre-trained models for 3D object detection, yielding significant performance (Xu et al. 2024; Zheng et al. 2024).

However, to generate high-quality 3D bounding boxes and feature priors, these methods inevitably perform multi-step inference to accurately match the scores, limiting the real-world applicability. Moreover, global coordinate distortions also often exist in raw point clouds, resulting in the unreliable detection. Thus, we propose a robust single-stage fully sparse detection network base on DLF, conducting one-step inference, achieving multiple perturbation robustness.

3 Denoising Diffusion Probabilistic Models

In this section, we first introduce the background. Then, we explain the rationale behind multi-step inference and noise robustness. Subsequently, the noising and denoising mechanisms are reformulated to model multi-type perturbations.

3.1 Background

Given an observed sample $\boldsymbol{c} \sim P_{sample}$, a fitting target $\boldsymbol{x_0} \sim P_{target}$, and a latent variable $\boldsymbol{x_T} \sim P_{noise}$, DDPMs achieve ***the distribution transformation process between*** P_{target} ***and*** P_{noise} via: a predefined diffusion process q that gradually adds noise to $\boldsymbol{x_0}$ until $\boldsymbol{x_0}$ degrades into $\boldsymbol{x_T}$, and a trainable generation process p_θ that slowly removes perturbation from $\boldsymbol{x_T}$ until $\boldsymbol{x_T}$ recovers $\boldsymbol{x_0}$ conditioned on $\boldsymbol{c}$. Following this framework, DDPMs have been successful in various 3D tasks (Qu et al. 2025). In 3DOD, Fig. 1(a) ($\boldsymbol{c}$ →point cloud, $\boldsymbol{x_0}$→3D box) and Fig. 1(b) ($\boldsymbol{c}$→task knowledge, $\boldsymbol{x_0}$→point cloud) are commonly used paradigms.

Noising and Training Objective. Given the strong performance observed in prior works (Ho, Jain, and Abbeel 2020; Qu et al. 2025), we adopt noise $\boldsymbol{\epsilon}$ as the fitting target:

$$\begin{aligned} \boldsymbol{x_t} &= \sqrt{\overline{\alpha}_t}\boldsymbol{x_0} + \sqrt{1-\overline{\alpha}_t}\cdot\boldsymbol{\epsilon_{t-1}}, \\ L(\theta) &= \mathbb{E}_{\boldsymbol{\epsilon_{t-1}}\sim\mathcal{N}(0,I)}||\boldsymbol{\epsilon_{t-1}} - \boldsymbol{\epsilon_\theta}(\boldsymbol{x_t}, C, t)||_2^2, \end{aligned} \tag{1}$$

where $C = \{c_i | i = 1, \cdots, S\}$ is an optional condition set ($C = \emptyset$ allowed) and $t \sim U[T]$ (T=1000 in this paper). The means that Eq. 1 serves as a general objective for unconditional and conditional DDPMs (Qu et al. 2024).

The Gradient of The Data Distribution. We can describe the objective using stochastic differential equations (SDEs):

$$\alpha\boldsymbol{\epsilon_\theta}(\boldsymbol{x_t}, C, t) = s_\theta(\boldsymbol{x_t}, C, t) \approx \nabla_{\boldsymbol{x_t}} \log P_t(\boldsymbol{x_t}). \tag{2}$$

Therefore, this noise objective is equivalent to the score (*i.e.*, the gradient of the data distribution) (Song et al. 2021), up to a constant factor $\alpha = -1/\sqrt{1-\overline{\alpha}_t}$.

Denoising and Inference Sampling. When approximating $\boldsymbol{\epsilon_\theta}(\boldsymbol{x_t}, C, t)$ by $\boldsymbol{\epsilon_{t-1}}$, (*i.e.*, $\boldsymbol{\epsilon_\theta}(\boldsymbol{x_t}, C, t) \approx \boldsymbol{\epsilon_{t-1}}$), the trained $\boldsymbol{\epsilon_\theta}$ can then be used for iterative inference sampling as:

$$\begin{aligned} \boldsymbol{x_{t-1}} = &\frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x_t} - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\epsilon_{t-1}}) \\ &+\sqrt{\frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}(1-\alpha_t)}\cdot\boldsymbol{\epsilon}. \end{aligned} \tag{3}$$

Eq. 1 and Eq. 3 indicate DDPMs adopt the sampling formulation $\boldsymbol{x_t} = \boldsymbol{\mu}_t + \sigma_t\boldsymbol{\epsilon}$ during both training and inference (***noising via*** $q(\boldsymbol{x_t}|\boldsymbol{x_0})$***, denoising via*** $q(\boldsymbol{x_{t-1}}|\boldsymbol{x_t}, \boldsymbol{x_0})$).

3.2 Robustness Stems from Training Stage

We first explain why DDPMs require multi-step inference, and then analyze how their robustness to noise stems from the training phase rather than the inference procedure itself.

Multi-Step Inference. DDPMs actually construct richer samples and targets from the modeled distribution in training than non-DDPMs. We provide a proof for this claim. For 3DOD, given a detection pair (point cloud $\boldsymbol{c}$, 3D box $\boldsymbol{x_0}$), a non-DDPM network f_θ can directly fit the input $\boldsymbol{c}$ to the target $\boldsymbol{x_0}$. However, the DDPM denoising network $\boldsymbol{\epsilon_\theta}$ takes $\boldsymbol{c}$ and $\{\boldsymbol{x_t}|t = 1, \cdots, T\}$ as the inputs to fit the targets $\{\boldsymbol{\epsilon_{t-1}}|t = 1, \cdots, T\}$ (assuming using the MSE loss):

$$\begin{aligned} L_f(\theta) &= ||\boldsymbol{x_0} - f_\theta(\boldsymbol{c})||_2^2, \\ L_\epsilon(\theta) &= \frac{1}{T}\sum_{t=1}^{T}||\boldsymbol{\epsilon_{t-1}} - \boldsymbol{\epsilon_\theta}(\boldsymbol{x_t}, \boldsymbol{c}, t)||_2^2. \end{aligned} \tag{4}$$

As described in Eq. 4, under the same network architecture, DDPMs require T times longer to fit the targets than non-DDPMs. ***This is because transitioning between two distributions (*** P_{target} ***and*** P_{noise}***) with a large difference in one-step inference will lead to the significant error for the score matching*** (Song et al. 2021).

Robustness Source. This noise construction pattern grants DDPMs adaptability to related distribution noise, as DDPMs can perceive more noise samples and targets in training than non-DDPMs. This also means that the DDPM robustness actually stems from ***the noise samples and targets constructed in training***. Meanwhile, this training prior is independent of the inference approach. That is, DDPMs can ***retain the noise robustness without following multi-step inference***.

3.3 DDPMs Can Model Multiple Perturbations

Besides point-level random noise, raw point clouds may also be affected by global geometric distortions, such as coordinate offset, scaling, rotation and other perturbations. In this paper, we rethink noising and denoising mechanisms, offering a way to model multiple perturbations in DDPMs.

Distribution Matching. In fact, $\boldsymbol{\epsilon_{t-1}}$ acts as a transition bridge between two distributions in Eq. 1 (noising): $q(\boldsymbol{x_t}|\boldsymbol{x_{t-1}}) \overset{\boldsymbol{\epsilon_{t-1}}}{\longleftarrow} q(\boldsymbol{x_{t-1}}|\boldsymbol{x_{t-2}})$, and between two distributions in Eq. 3 (denoising): $q(\boldsymbol{x_t}|\boldsymbol{x_{t+1}}, \boldsymbol{x_0}) \overset{\boldsymbol{\epsilon_{t-1}}}{\longrightarrow} q(\boldsymbol{x_{t-1}}|\boldsymbol{x_t}, \boldsymbol{x_0})$. ***When*** $\boldsymbol{\epsilon_\theta}$ ***accurately estimates*** $\boldsymbol{\epsilon_{t-1}}$***, DDPMs actually achieve the distribution matching***, *i.e.*, $p_\theta(\boldsymbol{x_{t-1}}|\boldsymbol{x_t}) \approx q(\boldsymbol{x_{t-1}}|\boldsymbol{x_t}, \boldsymbol{x_0})$ (see Fig. 2(a)). This actually estimates the distribution of denoising at each step.

Sample Fitting. However, the distribution matching rule requires deriving $q(\boldsymbol{x_{t-1}}|\boldsymbol{x_t}, \boldsymbol{x_0})$, involving a complex formula chain. This poses challenges when we remodel other distributions. Actually, $\boldsymbol{x_t}$ in training and inference are computed via $\boldsymbol{\mu_t} + \sigma_t\boldsymbol{\epsilon}$. Meanwhile, $\boldsymbol{\mu_t}$, σ_t, and $\boldsymbol{\epsilon}$ are all known in training. This means that the complex distribution matching task can actually be reinterpreted to a simple sample fitting problem: ***When*** $\boldsymbol{\epsilon_\theta}$ ***accurately estimates*** $\boldsymbol{\epsilon_{t-1}}$***, DDPMs actually achieve the sample fitting***, *i.e.*, $\boldsymbol{x'_{t-1}} \approx \boldsymbol{x_{t-1}}$ (see Fig. 2(b)). That is, we no longer sample $\boldsymbol{x_{t-1}}$ from $q(\boldsymbol{x_{t-1}}|\boldsymbol{x_0})$ but estimate $\boldsymbol{x_{t-1}}$ directly. In fact, $\boldsymbol{\epsilon_\theta}$ interacts only with noise samples and targets ***without realizing the distribution concept*** in training. ***The distribution transformation is manually introduced at inference according to Eq. 3.*** Dropping the distribution concept, we can redefine the denoising using $q(\boldsymbol{x_{t-1}}|\boldsymbol{x_0})$ instead of $q(\boldsymbol{x_{t-1}}|\boldsymbol{x_t}, \boldsymbol{x_0})$, decoupling the complex formula chain.

Modeling Multiple Perturbations. Under the sample fitting rule, we can reformulate noising $q(\boldsymbol{x_t}|\boldsymbol{x_0})$ and denoising $q(\boldsymbol{x_{t-1}}|\boldsymbol{x_0})$ to apply ***invertible affine transformations***:

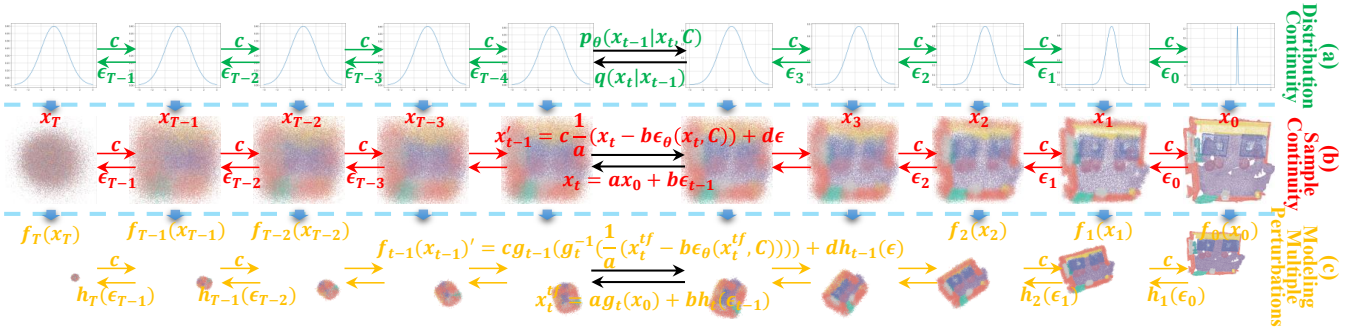

Figure 2: Visualization of distribution matching, sample fitting, and modeling multiple perturbations. (a) The distribution matching requires each step to follow continuous and correlated distribution transformations, forming a complex formula chain (Ho, Jain, and Abbeel 2020). (b) The sample fitting focuses on effectively estimating the next-step sample, simplifying the construction conception (Bansal et al. 2023). (c) The multi-type and multi-level noise samples and targets can be constructed using affine transformations. For example, $S_t(\boldsymbol{x_t}) = \sqrt{\overline{\alpha}_t} S_t(\boldsymbol{x_0}) + \sqrt{1-\overline{\alpha}_t} \cdot S_t(\boldsymbol{\epsilon_{t-1}})$, this implements an S_t-fold scaling of $\boldsymbol{x_t}$.

$$\begin{aligned} \boldsymbol{x_t^{tf}} &= a g_t(\boldsymbol{x_0}) + b \cdot h_t(\boldsymbol{\epsilon_{t-1}}), \\ \boldsymbol{x_{t-1}^{tf}} &= c g_{t-1}(g_t^{-1}(\frac{1}{a}(\boldsymbol{x_t^{tf}} - b h_t(\boldsymbol{\epsilon_{t-1}})))) + d \cdot h_{t-1}(\boldsymbol{\epsilon}) \end{aligned} \tag{5}$$

where a, b, c, and d are constant coefficients. $f_t(\boldsymbol{x_t}) = \boldsymbol{x_t^{tf}}$, $f_t(\cdot)$, $g_t(\cdot)$, $h_t(\cdot)$ denote affine transformation functions based on t (derivations in Supplementary Material (SM)).

In fact, Eq. 5 provides offers a general and flexible formulation to construct multi-type noise in DDPMs. $\boldsymbol{\epsilon}$ in Eq. 5 can transcend the limitation of the Gaussian distribution (Austin et al. 2021), even modeling the deterministic diffusion process (Bansal et al. 2023), such as snowing and masking. This enables the construction of any types of noise samples and targets without the distribution limitation, achieving robustness to multiple perturbations. Some works have actually adopted the sample fitting rule (***the noising follows*** $q(\boldsymbol{x_t}|\boldsymbol{x_0})$***, the denoising follows*** $q(\boldsymbol{x_{t-1}}|\boldsymbol{x_0})$) (Bansal et al. 2023; Naval Marimont et al. 2024), but without the unified formulation. This way avoids complex denoising derivations. Moreover, the generation diversity still comes from $\boldsymbol{\epsilon}$. ***We also construct more types of noise samples and targets (see Fig. 2(c), more implementations and visuals in SM).***

4 Methodology

4.1 Detachable Latent Framework

Based on the insights from Sec. 3.2 and Sec. 3.3, we propose a Detachable Latent Framework of DDPMs (DLF, see Fig. 1(c)) to overcome multi-step inference and achieve multiple perturbation robustness. Unlike Fig.1(a) and Fig.1(b), DLF treats the denoising network $\boldsymbol{\epsilon_\theta}$ as an auxiliary branch, guiding the backbone f_ψ to learn the multi-type and multi-level denoising process in latent feature spaces.

Specifically, to learn the noise-robust denoising process, the latent feature $\boldsymbol{x_0^{lat}}$ from f_ψ is perturbed to construct multi-type and multi-level noise samples and targets:

$$f_t^*(\boldsymbol{x_t}) = \sqrt{\overline{\alpha}_t} g_t^*(\boldsymbol{x_0^{lat}}) + \sqrt{1-\overline{\alpha}_t} \cdot h_t^*(\boldsymbol{\epsilon_{t-1}}), \tag{6}$$

where $f_t^*(\cdot)$, $g_t^*(\cdot)$ and $h_t^*(\cdot)$ denote composite affine functions with intensity varying base on t, such as translation, scaling, rotation or other transformations (see Fig. 2(c)).

Subsequently, the auxiliary denoising network $\boldsymbol{\epsilon_\theta}$ guides the backbone f_ψ to learn the task-relevant information in multi-type and multi-level denoising learning:

$$\begin{aligned} L(\theta) = \mathbb{E}_{\boldsymbol{\epsilon_{t-1}} \sim \mathcal{N}(0,I)} || h_t^*(\boldsymbol{\epsilon_{t-1}}) \\ - \boldsymbol{\epsilon_\theta}(f_t^*(\boldsymbol{x_t}), C_{task}, t) ||_2^2, \end{aligned} \tag{7}$$

where C_{task} is task-specific conditions (knowledge priors).

Next, f_ψ determines the task result in training and inference, while the denoise network $\boldsymbol{\epsilon_\theta}$ is detached in inference:

$$\begin{aligned} L(\psi) = l_{task}(h_{task}(f_\psi(I_{task})), GT_{task}), \\ P_{task} = h_{task}(f_\psi(I_{task})), \end{aligned} \tag{8}$$

where I_{task}, GT_{task}, $l_{task}(\cdot)$, $h_{task}(\cdot)$ and P_{task} represent the task-related input, Ground Truth, loss function, task head and prediction, respectively.

This simple and effective DDPM paradigm:

- **Following Noise Construction Pattern.** Eq. 6 indicates that DLF constructs multi-type and multi-level noise samples and targets in training, aligning with Eq. 5 (noising), preserving the multi-type noise robustness.
- **Aligning with Training objective.** Eq. 7 shows that DLF follows the original training objective (the score, $h_t^*(\boldsymbol{\epsilon_{t-1}}) \approx \nabla_{\boldsymbol{x_t}} \log P_t(\boldsymbol{x_t})$, the noise target $h_t^*(\boldsymbol{\epsilon_{t-1}})$ performs better than $\boldsymbol{\epsilon_{t-1}}$, $\boldsymbol{x_0^{lat}}$ and $g_t^*(\boldsymbol{x_0^{lat}})$ in the additional ablation study of SM), aligning with Eq. 2, achieving the distribution matching continuity/the sample fitting continuity (see Sec. 3.3).
- **Relaxing Score Matching Requirement.** Eq. 8 presents that the task result relies on f_ψ rather than $\boldsymbol{\epsilon_\theta}$, reducing the score learning difficulty in training.
- **With Minimal Cost.** Eq. 6, Eq. 7 and Eq. 8 mean that DLF introduces only limited cost in training by operating in latent feature spaces, while also avoiding extra training data and inference cost.

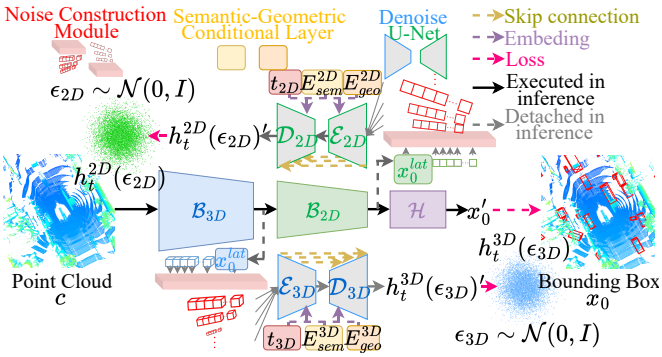

Figure 3: The overall framework of RSDNet. DUNet takes multi-type and multi-level noise samples from NCM and semantic-geometric conditional embeddings from SGCL to guide FSP in denoising, while the DUNet detachable design enables FSP to perform one-step detection in inference.

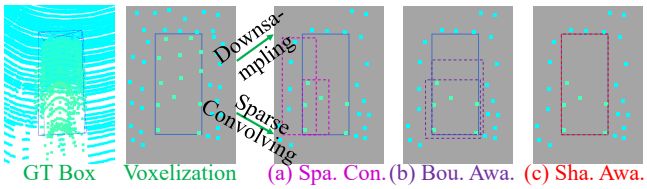

Figure 4: The sparse features generated by downsampling or sparse convolution may lead to the center feature missing problem. (a) Missing center features may cause imprecise and unstable bounding box predictions. In such cases, background points may be misclassified as foreground due to the feature confusion, leading to false positives or incomplete detections. (b) The boundary aware guidance enhances the bounding box prediction for differentiating foreground points and background points. (c) The shape aware guidance further improves the alignment between the predicted bounding boxes and the ground truth bounding boxes.

4.2 Network Architecture

In this section, we introduce the overall framework of RSDNet. This consists of four main components: the Fully Sparse Pipeline (FSP), the Noise Construction Module (NCM), the Semantic-Geometric Conditional Layer (SGCL) and the Denoising U-Net (DUNet), as illustrated in Fig. 3 (the implementation and parameter details in SM).

Fully Sparse Pipeline. FSP follows traditional hybrid sparse detection pipelines (Zhang et al. 2024; Liu et al. 2025), focusing on the detection results. The 3D sparse backbone first progressively (downsampling) extracts 3D sparse features via a voxel feature encoder (VFE) and five 3D sparse convolution layers. Then, two-stage linear self-attention block and a max-pooling layer are used to enhance the sparse representation perceive field. Subsequently, the 2D sparse BEV representations from the compressed 3D sparse features further perceive contextual cues in the BEV space through a feature diffusion module and four 2D sparse convolution layers. Next, they are fed into the sparse detection head (Zhang et al. 2024) for the final prediction. We use only point clouds as inputs, ensuring FSP concentrates on efficient and accurate object localization in a pure manner.

Noise Construction Module. NCM perturbs the latent feature $\boldsymbol{x_0^{lat}}$ from the backbone (2D or 3D), constructing multi-type and multi-level noise samples and targets. To ensure the lightweight design, $\boldsymbol{x_0^{lat}}$ is first progressively projected to a low-dimensional space through a three-layer sparse convolution. Then, as described in Eq. 6, NCM applies composite affine transformations to $\boldsymbol{x_0^{lat}}$ and $\boldsymbol{\epsilon_{t-1}}$:

$$\begin{aligned} g_t^*(\boldsymbol{x_0^{lat}}) &= R_t \cdot S_t(\boldsymbol{x_0^{lat}} - \boldsymbol{T_t}), \\ h_t^*(\boldsymbol{\epsilon_{t-1}}) &= R_t \cdot S_t(\boldsymbol{\epsilon_{t-1}} - \boldsymbol{T_t}), \end{aligned} \tag{9}$$

where $\boldsymbol{T_t}$, $S_t(\cdot)$ and R_t denote the offset vector, the scaling function, and the rotation matrix with intensity varying based on t, respectively (the implementation details in SM).

Next, NCM performs the Gaussian noising process of DDPMs for $g_t^*(\boldsymbol{x_0^{lat}})$ and $h_t^*(\boldsymbol{\epsilon_{t-1}})$ to construct $f_t^*(\boldsymbol{x_t})$.

Semantic-Geometric Conditional Layer. SGCL embeds semantic and geometric priors from the ground-truth 3D bounding boxes B_{all}, enhancing the backbone awareness for object boundaries and shapes in the denoising learning. This first selects the most relevant 3D bounding box B_s for each point feature F based on $\mathcal{L}_2$ distance between the box center B_c and the voxel center V_c. Then, SGCL checks whether the voxel falls within the corresponding box:

$$\begin{aligned} B_s &= argmin(\mathcal{L}_2(B_c, V_c), B_{all}), \\ I_{mask} &= \mathcal{I}(B_s, V_c), \end{aligned} \tag{10}$$

where $\mathcal{I}(\cdot)$ denotes an indicator function: the outputs 1 if the voxel falls within the corresponding box (foreground point), and 0 otherwise (background point). Meanwhile, SGCL uses 0 to fill in the box information for background voxels.

Subsequently, SGCL embeds the semantic embedding E_{sem} from I_{mask} and the geometric embedding E_{geo} from B_s, and fuses them with the point feature F:

$$\begin{aligned} E_{sem} = embeding(I_{mask}), E_{geo} = mlp(B_f), \\ F' = F + mlp(cat(E_{sem}, E_{geo})). \end{aligned} \tag{11}$$

The class label is excluded from E_{geo} due to the absence of shape information about the object. Guided by the semantic and geometric priors, this can effectively mitigate the ***center feature missing*** problem (Zhang et al. 2024) caused by downsampling or sparse convolution (see Fig. 4).

Denoise U-Net. DUNet takes $f_t^*(\boldsymbol{x_t})$ from NCM and performs the denoising learning guided by E_{sem} and E_{geo} from SGCL. For the hybrid backbone, DUNet includes two some architecture sub-networks: the 3D Denoising U-Net (3DDU) and 2D Denoising U-Net (2DDU). According to Sec. 4.1, DUNet should exhibit lightweight due to the insignificant score matching requirement. Therefore, this follows a four-layer low-channel sparse convolution framework in the encoder and decoder (Shi et al. 2020). Meanwhile, DUNet introduces the time label t to model the diffusion process.

Methods	NDS	mAP	Car	Truck	Bus	Trailer	Vehicle	Pedestrian	Motor	Bike	Cone	Barrier
CenterPoint	66.5	59.2	84.9	57.4	70.7	38.1	16.9	85.1	59.0	42.0	69.8	68.3
PillarNeXt	68.4	62.2	85.0	57.4	67.6	35.6	20.6	86.8	68.6	53.1	77.3	69.7
VoxelNeXt†	68.7	63.5	83.9	55.5	70.5	38.1	21.1	84.6	62.8	50.0	69.4	69.4
HEDNet	71.4	66.7	87.7	60.6	77.8	50.7	28.9	87.1	74.3	56.8	76.3	66.9
FSDv2†	70.4	64.7	84.4	57.3	75.9	44.1	28.5	86.9	69.5	57.4	72.9	73.6
SAFDNet†	71.0	66.3	87.6	60.8	78.0	43.5	26.6	87.8	75.5	58.0	75.0	69.7
FSHNet†	71.7	68.1	88.7	61.4	79.3	47.8	26.3	89.3	76.7	60.5	78.6	72.3
Baseline†	71.2	68.0	87.7	62.1	78.3	42.7	26.9	88.7	76.6	59.5	78.5	79.2
RSDNet†	71.9	68.9	88.4	63.1	79.0	43.2	28.2	89.2	77.7	59.9	79.4	80.4

Table 1: The detection results on nuScenes. RSDNet significantly outperforms other methods in terms of NDS and mAP.

Methods	LEVEL1	LEVEL2	LEVEL1			LEVEL2		
	mAP/mAPH	mAP/mAPH	Vehicle	Pedestrian	Cyclist	Vehicle	Pedestrian	Cyclist
CentorPoint	74.4/71.7	68.2/65.8	74.2/73.6	76.6/70.5	72.3/71.1	66.2/65.7	68.8/63.2	69.7/68.5
PillarNeXt	78.0/75.7	71.9/69.7	78.4/77.9	82.5/77.1	73.2/72.2	70.3/69.8	74.9/69.8	70.6/69.6
VoxelNeXt†	78.6/76.3	72.2/70.1	78.2/77.7	81.5/76.3	76.1/74.9	69.9/69.4	73.5/68.6	73.3/72.2
HEDNet	81.4/79.4	75.3/73.4	81.1/80.6	84.4/80.0	78.7/77.7	73.2/72.7	76.8/72.6	75.8/74.9
FSDv2†	80.3/79.5	75.6/73.5	79.8/79.3	84.8/79.7	80.7/79.6	71.4/71.0	77.4/72.5	77.9/76.8
SAFDNet†	81.8/79.8	75.7/73.9	80.6/80.1	84.7/80.4	80.0/79.0	72.7/72.3	77.3/73.1	77.2/76.2
FSHNet†	82.7/80.6	77.1/74.9	82.2/81.7	85.9/80.8	80.5/79.4	74.5/74.0	78.9/73.9	78.0/76.9
Baseline†	82.7/80.5	76.9/74.8	82.0/81.5	85.7/80.7	80.3/79.2	74.2/73.8	78.8/73.7	77.8/76.8
RSDNet†	83.7/81.4	77.8/75.6	82.8/82.3	86.7/81.5	81.6/80.5	74.9/74.5	79.8/74.7	78.7/77.7

Table 2: The detection results on Waymo Open. RSDNet demonstrates excellent detection performance in large-scale scenes.

4.3 Training and Inference

Training. As mentioned in Sec. 4.1, the training objective of RSDNet includes the diffusion loss and the task loss:

$$L_{total} = \lambda L(\theta) + L(\psi), \tag{12}$$

where $L(\theta) = L_{3D}(\theta) + L_{2D}(\theta)$. Meanwhile, $L(\psi) = L_{reg}(\psi) + L_{cls}(\psi)$, $L_{reg}(\psi)$ and $L_{cls}(\psi)$ mean the regression and classification loss functions (see Fig. 3).

Inference. Thanks to the detachable design of DLF, RSDNet can perform detection in only one-step inference:

$$\boldsymbol{x_0'} = \mathcal{H}(\mathcal{B}_{2D}(\mathcal{B}_{3D}(\boldsymbol{c}))), \tag{13}$$

where $\boldsymbol{x_0'}$ means the predicted bounding box (see Fig. 3).

5 Experiments

5.1 Experiment Setup

Dataset. We perform the main experiments on nuScenes (Caesar et al. 2020), following the official protocol to divide train/val/test with 700/150/150 scenes. We also conduct experiments on Waymo Open (Sun et al. 2020), splitting train/-val/test into 798/202/150 scenes.

Detection Methods. We compare RSDNet with current popular detection methods: CenterPoint (Yin, Zhou, and Krahenbuhl 2021), PillarNeXt (Li, Luo, and Yang 2023), VoxelNeXt†(Chen et al. 2023b), HEDNet (Zhang et al. 2023), FSDv2†(Fan et al. 2024), SAFDNet†(Zhang et al. 2024), FSHNet†(Liu et al. 2025). †means a fully sparse detector.

Baseline. To validate the effectiveness of our method, we remove 3DDU and 2DDU from RSDNet and treat the Fully Sparse Pipeline (FSP) as the baseline.

5.2 Comparison of Detection Results

Results on nuScenes. We first conduct the evaluation on the nuScenes dataset. As shown in Tab. 1, RSDNet achieves the significant detection performance, outperforming the existing state-of-the-art methods. This is because, multi-type and multi-level denoising learning improves the robustness and generalization of RSDNet in unseen scenes, especially for noise-sensitive small objects. Meanwhile, the semantic-geometric conditional guidance alleviates the center feature missing problem from downsampling and sparse convolution, further enhancing small object detection results.

Results on Waymo Open. Furthermore, we also conduct the evaluation on the longer-range Waymo Open. As shown in Tab. 2, RSDNet still achieves state-of-the-art detection performance. Benefiting from the robustness to perturbations, RSDNet exhibits strong generalization in scenes with the sparser object distribution (Qu et al. 2025).

5.3 Validation for Perturbation Robustness

To verify reliable and stable detection, we conduct the robustness evaluation for multiple perturbations on nuScenes.

Methods	Small τ (mAP)		Big τ (mAP)	
	τ=0.05	τ=0.08	τ=0.125	τ=0.15
HEDNet	58.5	39.7	5.7	1.0
SAFDNet†	57.4	38.7	4.4	1.0
FSHNet†	60.1	40.4	9.4	1.2
Baseline†	59.8	40.2	8.9	1.2
RSDNet†	64.0	51.2	18.4	7.5

Table 3: The results for Gaussian noise on nuScenes.

Point-level Random Noise. Raw point clouds from the sensor often contain point-wise random noise. We add the Gaussian noise $\boldsymbol{n_G} \sim \mathcal{N}(\boldsymbol{n_G}; \mathbf{0}, \tau \boldsymbol{I})$ to the normalized input of the model (Qu et al. 2025), *i.e.*, $\boldsymbol{c'} = \boldsymbol{c} + \boldsymbol{n_G}$, evaluating the robustness for point-level random noise. Tab. 3 shows that RSDNet exhibits the strong noise robustness compared to other methods. Meanwhile, the detection results drop close to 0 when τ=0.15, indicating that existing detection methods are relatively sensitive to point-level random noise.

Global Geometric Distortions. Coordinate offsets, scaling, and rotations may also often be presented in raw point clouds. Similarly, we apply translation ($\boldsymbol{c'} = \boldsymbol{c} - \boldsymbol{T}$), scaling ($\boldsymbol{c'} = S\boldsymbol{c}$), and rotation ($\boldsymbol{c'} = R \cdot \boldsymbol{c}$) perturbations to the normalized input of the model. Fig. 5 shows the results. Benefiting from multi-type noise samples and targets, RSDNet exhibits strong robustness to global geometric distortions.

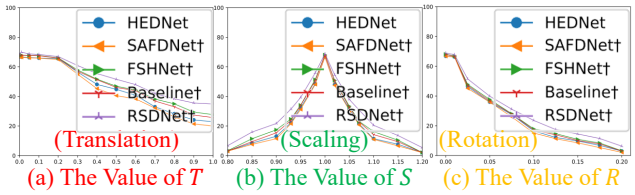

Figure 5: (a), (b), and (c) mean the results of applying translation, scale, and rotation perturbations, respectively.

5.4 Generalization for DLF

We also conduct the experiments for introducing DLF into HEDNet and SAFDNet. We simply integrate 3DDU and 2DDU into HEDNet and SAFDNet. Tab. 4 demonstrates that integrating DLF obtains the better noise robustness and detection performance for HEDNet and SAFDNet. Meanwhile, this avoids introducing the additional inference cost, as the detachable design of DLF.

Methods	#Params	mAP	τ=0.10	MIT/MM
HEDNet	15.3M	58.5	20.4	0.18s/3.8G
HEDNet+DLF	18.3M	60.8	31.1	0.18s/3.8G
SAFDNet†	15.7M	57.4	20.2	0.15s/4.7G
SAFDNet+DLF†	18.7M	60.5	30.8	0.15s/4.7G
RSDNet†	16.5M	64.0	34.2	0.16s/5.8G

Table 4: The results of multiple backbones on nuScenes. 'MIT' and 'MM' indicate the *Mean Inference Time* and the *Mean Memory* for each point cloud, respectively.

5.5 Ablation Study

The Denoising Learning. We first conduct the ablation study for the denoising learning in RSDNet. Tab. 5 shows the results on nuScenes. RSDNet* exhibits a significant drop across all evaluation metrics. Removing the conditional denoising learning under the task-specific knowledge guidance hinders learning robust and generalizable representations for RSDNet, reducing the object detection performance in unseen scenes, especially for noise-sensitive small objects.

Methods	#Params	mAP	Robustness (mAP)			
			τ=0.05	$\boldsymbol{T}$=0.5	S=0.95	R=0.05
Baseline	11.1M	68.0	59.8	46.9	34.1	37.7
RSDNet*	15.8M	67.9	59.4	45.3	33.2	36.4
RSDNet$_S$	15.8M	68.6	62.9	50.4	38.1	38.9
RSDNet	16.5M	68.9	64.0	51.6	39.3	39.9

Table 5: Ablation study of the denoising learning in RSDNet on nuScenes. The baseline (Fully Sparse Pipeline, FSP) means removing DUNet from RSDNet. Meanwhile, RSDNet* denotes removing the denoising process from RSDNet, but retaining the DUNet targeting $\boldsymbol{x_0^{lat}}$. RSDNet$_S$ indicates removing SGCL from RSDNet. This results demonstrate the importance of the denoising learning for RSDNet.

This also suggests that simply increasing model capacity cannot guarantee the better detection performance and may even lead to overfitting (RSDNet* is lower than baseline).

The Channel Dimension of Projecting $\boldsymbol{x_0^{lat}}$. As mentioned in Sec. 4.1, DUNet should expect to be lightweight. Therefore, we explore the impact of projecting $\boldsymbol{x_0^{lat}}$ into different latent spaces for RSDNet. Fig. 6(a) shows that projecting $\boldsymbol{x_0^{lat}}$ into 8 channels yields the best trade-off between performance and efficiency. Meanwhile, further increasing the channel dimensions leads to saturated or even degraded performance. We believe that the larger channel dimension may introduce more unreasonable perturbations, impairing the effectiveness of the denoising learning for RSDNet.

The Value of λ. Furthermore, we also conduct the ablation study for the value of λ, as illustrated in Fig. 6(b). We can observe that when λ becomes large, the performance of RSDNet drops significantly. This further validates that excessive perturbations can impair the denoising learning effect for RSDNet, hindering effectively understanding the scene context under the multi-type and multi-level perturbations.

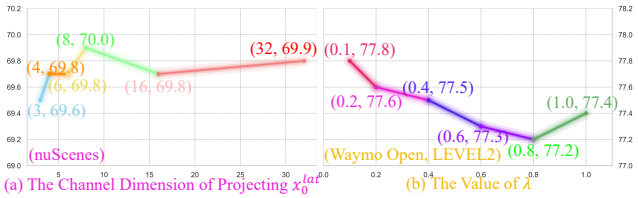

Figure 6: The y-axis represents mAP. (a) The ablation study of the channel dimension for projecting $\boldsymbol{x_0^{lat}}$ on nuScenes. (b) The ablation study for the value of λ on Waymo Open.

6 Conclusion

In this paper, we revealed the source of robustness and reformulated the noising and denoising rules. Building upon these core insights, a Detachable Latent Framework of DDPMs was designed to overcome multi-step inference and model multiple perturbations. Furthermore, based on this, we proposed a robust single-stage fully sparse object detection network, exhibiting superior robustness and performance. Overall, we provided new knowledge for applying DDPMs, hoping to inspire further extensions in 3D tasks.

Acknowledgments

This work was supported in part by the Frontier Technologies R&D Program of Jiangsu under grant BF2024070, in part by the National Natural Science Foundation of China under Grant 62471235, in part by Hunan Natural Science Foundation Project (No. 2025JJ50338) and Shanghai Education Committee AI Project (No. JWAIYB-2), in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX25_0754, and in part by PNRR FAIR - Future AI Research (PE00000013).

References

Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; and Van Den Berg, R. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34: 17981–17993.

Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; and Tai, C.-L. 2022. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1090–1099.

Bansal, A.; Borgnia, E.; Chu, H.-M.; Li, J.; Kazemi, H.; Huang, F.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Cold diffusion: Inverting arbitrary image transforms without noise. *Advances in Neural Information Processing Systems*, 36: 41259–41282.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023a. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 19830–19843.

Chen, X.; Liu, Z.; Luo, K.; Datta, S.; Polavaram, A.; Wang, Y.; You, Y.; Li, B.; Pavone, M.; Chao, W.-L. H.; et al. 2024a. Diffubox: Refining 3d object detection with point diffusion. *Advances in Neural Information Processing Systems*, 37: 103681–103705.

Chen, X.; Liu, Z.; Xie, S.; and He, K. 2024b. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1907–1915.

Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; and Jia, J. 2023b. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21674–21683.

Chen, Z.; Sun, K.; and Lin, X. 2024. CamoDiffusion: Camouflaged object detection via conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1272–1280.

Chu, B.; Xu, X.; Wang, X.; Zhang, Y.; You, W.; and Zhou, L. 2025. Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12830–12839.

Ding, S.; Chen, X.; Ai, C.; Wang, J.; and Yang, H. 2024. A noise-reduction algorithm for raw 3D point cloud data of asphalt pavement surface texture. *Scientific Reports*, 14(1): 16633.

Fan, L.; Wang, F.; Wang, N.; and Zhang, Z. 2024. Fsd v2: Improving fully sparse 3d object detection with virtual voxels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ho, C.-J.; Tai, C.-H.; Lin, Y.-Y.; Yang, M.-H.; and Tsai, Y.-H. 2023. Diffusion-ss3d: Diffusion model for semi-supervised 3d object detection. *Advances in Neural Information Processing Systems*, 36: 49100–49112.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Li, J.; Luo, C.; and Yang, X. 2023. PillarNeXt: Rethinking network designs for 3D object detection in LiDAR point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17567–17576.

Li, J.; Saltori, C.; Poiesi, F.; and Sebe, N. 2025. Cross-modal and uncertainty-aware agglomeration for open-vocabulary 3d scene understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19390–19400.

Lian, J.; Du, X.; Liu, J.; Hui, L.; and Yang, J. 2025. Cross-Modal Driven Object Restoration for 3D Point Cloud Backdoor Defense. *IEEE Transactions on Information Forensics and Security*.

Lian, J.; Wang, D.-H.; Wu, Y.; and Zhu, S. 2023. Multi-branch enhanced discriminative network for vehicle re-identification. *IEEE Transactions on Intelligent Transportation Systems*, 25(2): 1263–1274.

Liu, S.; Cui, M.; Li, B.; Liang, Q.; Hong, T.; Huang, K.; and Shan, Y. 2025. FSHNet: Fully Sparse Hybrid Network for 3D Object Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8900–8909.

Liu, X.; Xiaoxu, X.; Li, J.; Zhang, Q.; Wang, X.; Sebe, N.; Lin, M.; et al. 2024. Less: Label-efficient and single-stage referring 3d instance segmentation. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. NeurIPS.

Naval Marimont, S.; Siomos, V.; Baugh, M.; Tzelepis, C.; Kainz, B.; and Tarroni, G. 2024. Ensembled cold-diffusion restorations for unsupervised anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 243–253. Springer.

Pellicer, A. L.; Li, Y.; and Angelov, P. 2024. PUDD: towards robust multi-modal prototype-based deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3809–3817.

Qu, W.; Shao, Y.; Meng, L.; Huang, X.; and Xiao, L. 2024. A Conditional Denoising Diffusion Probabilistic Model for

Point Cloud Upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20786–20795.

Qu, W.; Wang, J.; Gong, Y.; Huang, X.; and Xiao, L. 2025. An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27325–27335.

Rakotosaona, M.-J.; La Barbera, V.; Guerrero, P.; Mitra, N. J.; and Ovsjanikov, M. 2020. Pointcleannet: Learning to denoise and remove outliers from dense point clouds. In *Computer graphics forum*, volume 39, 185–203. Wiley Online Library.

Ranasinghe, Y.; Hegde, D.; and Patel, V. M. 2024. Monodiff: Monocular 3d object detection and pose estimation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10659–10670.

Shi, S.; Wang, X.; and Li, H. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 770–779.

Shi, S.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE transactions on pattern analysis and machine intelligence*, 43(8): 2647–2664.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.

Wang, L.; Jia, J.; and Dai, H. 2024. OrientedDiffDet: Diffusion model for oriented object detection in aerial images. *Applied Sciences*, 14(5): 2000.

Xiang, C.; Qi, C. R.; and Li, B. 2019. Generating 3d adversarial point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9136–9144.

Xiang, W.; Yang, H.; Huang, D.; and Wang, Y. 2023. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15802–15812.

Xu, C.; Ling, H.; Fidler, S.; and Litany, O. 2024. 3difftection: 3d object detection with geometry-aware diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10617–10627.

Yang, B.; Luo, W.; and Urtasun, R. 2018. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 7652–7660.

Yang, Z.; Sun, Y.; Liu, S.; Shen, X.; and Jia, J. 2019. Std: Sparse-to-dense 3d object detector for point cloud. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1951–1960.

Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11784–11793.

Zhang, G.; Chen, J.; Gao, G.; Li, J.; Liu, S.; and Hu, X. 2024. Safdnet: A simple and effective network for fully sparse 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14477–14486.

Zhang, G.; Junnan, C.; Gao, G.; Li, J.; and Hu, X. 2023. Hednet: A hierarchical encoder-decoder network for 3d object detection in point clouds. *Advances in Neural Information Processing Systems*, 36: 53076–53089.

Zhang, J.; Zhang, Y.; Qi, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2025. Geobev: Learning geometric bev representation for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9960–9968.

Zheng, X.; Huang, X.; Mei, G.; Hou, Y.; Lyu, Z.; Dai, B.; Ouyang, W.; and Gong, Y. 2024. Point Cloud Pre-training with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22935–22945.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4490–4499.