Equirectangular Point Reconstruction for Domain Adaptive Multimodal 3D Object Detection in Adverse Weather Conditions

Jae Hyun Yoon*, **Jong Won Jung***, **Seok Bong Yoo**†

Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea
sbyoo@jnu.ac.kr

Abstract

A multimodal fusion technique using LiDAR-camera has been developed for precise 3D object detection in autonomous driving and provides acceptable detection performance in ideal conditions with clear weather. However, the existing multimodal methods are still vulnerable to adverse weather conditions, such as snow, rain, and fog. These factors increase the point cloud sparsity due to occlusion and attenuation of the laser signal. A point cloud becomes sparser with increased distance, posing a challenge for object detection. To address these problems, we propose a point reconstruction network using equirectangular projection for multimodal 3D object detection. This network consists of distance-constrained denoising to remove adverse weather noise and an object-centric ray generator to generate distant object points flexibly. We propose a domain adaptation method that injects feature perturbations to improve detection performance by reducing the domain gap between different datasets. Furthermore, we propose a multimodal weather noise matching method for realistic data synthesis-based training to align the adverse weather noise between synthetic point clouds and images. The experimental results on adverse weather datasets confirm that the proposed approach outperforms the existing methods.

Code — https://github.com/jhyoon964/EquiDetect

Introduction

3D object detection has received significant attention in numerous applications and has achieved substantial advancements. Particularly, multimodal 3D object detection with complementary information from multiple sources has contributed to this advancement. Fusing the light detection and ranging (LiDAR) and camera data is one promising approach. Moreover, LiDAR provides precise distance information using light signals, and the camera offers rich texture information in images. Fusing these modalities enhances the ability to identify objects and their positions accurately.

Despite these advancements, existing multimodal detection methods (Wu et al. 2022b, 2023a) using the point cloud and images have limitations. For example, existing methods are primarily optimized for ideal conditions, such as clear weather, but are vulnerable to adverse weather conditions. Fig. 1(a) illustrates the 3D average precision of an existing state-of-the-art (SOTA) model, VirConv-S (Wu et al. 2023b), in diverse weather conditions. The model was trained and evaluated on the Dense dataset (Bijelic et al. 2020) consisting of diverse weather conditions. This model performs well in clear weather conditions, but its performance degrades by $9.85\%p$, $10.62\%p$, and $12.87\%p$ in snow, rain, and fog, respectively. This degradation is attributed to brightness variations, signal attenuation, and occlusion caused by weather noise affecting the object of interest.

For LiDAR emitting lasers centered around a sensor, the laser signals are reflected by weather noise before reaching an object, generating noisy points, as illustrated in Fig. 1(b). The attenuated pulsed lasers can fail to return signals, resulting in empty spaces (hole points) where points should represent the object corresponding to the laser path. Thus, the sparsity of the object points increases, corrupting object shapes and decreasing detection performance.

Additionally, Fig. 1(c) presents our observation of detection performance variations over the number of points by distance. We observed that there is an optimal number of points ($\star$) for detection at each distance. Moreover, the change in performance along the number of points becomes more extensive as the distance increases and more gradual as the distance decreases.

Another limitation is that the amount of real-world adverse weather data is limited due to the difficulty of data collection (Hahner et al. 2022). The weather datasets synthesized by simulator (Kilic et al. 2021; Hahner et al. 2021) can be an alternative to address this issue, and the simulator provides a benefit for restoration by offering a corresponding clean data. However, no weather simulator has been designed for multimodal purposes. Dong et al. (2023) used simulators to synthesize the point cloud and images to evaluate multimodal detection performance in adverse weather conditions. However, weather-element alignment is inconsistent because the point cloud and images are synthesized independently. Fig. 1(d) compares these synthesized data with real-world data by measuring the matching accuracy as the degree of overlap between the weather noise points and masks in the projected 2D space. The inconsistently synthe-

*These authors contributed equally.
†Corresponding author.

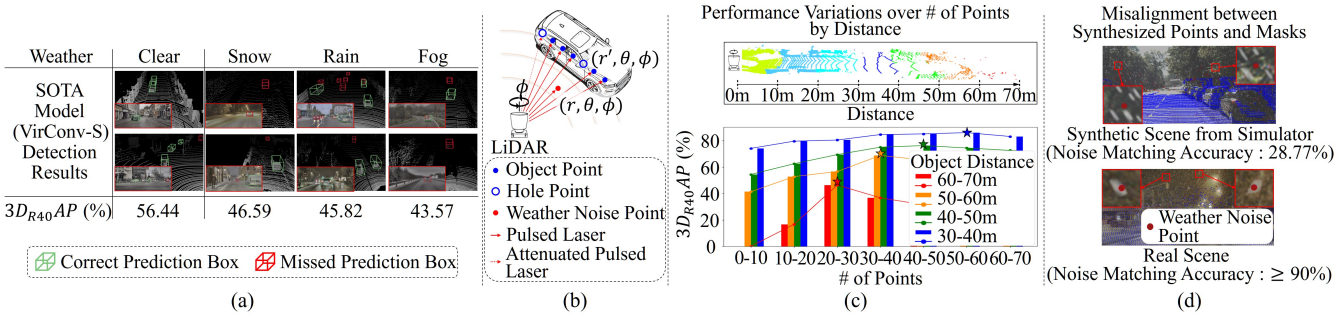

Figure 1: (a) Performance variations of the existing state-of-the-art model (VirConv-S) in adverse weather conditions on the Dense dataset. (b) Occlusion and attenuation effects of pulsed lasers due to adverse weather. (c) Performance variations over the number of points by distance on KITTI dataset. (d) Misalignment between images and point clouds synthesized using a physical-based simulator for adverse weather conditions.

sized data have a low matching accuracy of 28.77%, whereas the real data has a high matching accuracy of over 90%. This inconsistency reduces the evaluation reliability and increases the domain gap with real-world weather datasets.

To address these limitations, we propose EquiDetect. The EquiDetect removes noise points and fills in hole points in the range image, which is transformed via equirectangular projection by applying a distance-constrained denoising. Moreover, EquiDetect generates object points using a ray generator with object location information from the images. Further, we propose feature perturbations as an unsupervised domain adaptation (UDA) that can be applied in situations with limited data and labels, reducing the domain gap between real-world and synthetic weather datasets. We also introduce a multimodal noise matching method for alignment between synthetic point clouds and images, given the laser directionality. The contributions are summarized below:

- We propose an equirectangular point reconstructor using a distance-constrained denoising to remove adverse weather noise and an object-centric ray generator with an auxiliary 2D network to densify distant object points.
- We propose UDA modules injecting iterative feature perturbations to achieve fine-grained distribution alignment between datasets, ensuring optimal performance.
- We propose a multimodal weather noise matching method based on radial manipulation for a realistic alignment between the synthetic point cloud and images in adverse weather conditions.

Related Work

3D Object Detection

Single modal 3D object detection methods (Xu, Zhong, and Neumann 2022; Huang et al. 2024b; Jin et al. 2024) have advanced. In addition, 3D object detection is categorized into voxel-based methods (Wu et al. 2022a; Chen et al. 2023), converting point clouds into 3D voxels, and point-based methods (Shi and Rajkumar 2020; Li, Wang, and Wang 2021), detecting objects directly from the raw point cloud.

In contrast, multimodal 3D object detection methods (Chen et al. 2022; Yang et al. 2022a; Wu et al. 2022b, 2023b) use 2D images and 3D point clouds, addressing the distinct characteristics of the two modalities. These models display remarkable performance but are limited to clear weather conditions. In adverse weather conditions, methods using diverse sensors have been explored. In more detail, Huang et al. (2024a) enhanced detector performance in rain conditions with LiDAR, while Mai et al. (2022) addressed data distortion in fog conditions with stereo images and LiDAR. In contrast, Chae, Kim, and Yoon (2024), Palladin et al. (2024) and Chae et al. (2024) addressed various weathers using additional 4D radar. However, these methods do not consider the sensor characteristics. To address this problem, we propose a robust multimodal 3D object detection model for weather conditions using LiDAR characteristics.

Point Cloud and Image Restoration

Point cloud upsampling methods (Zeng et al. 2021; Akhtar et al. 2022) have significantly advanced in generating high-resolution point clouds. Moreover, Li et al. (2022) has explored transforming point cloud data on challenging weather conditions into data suitable for clear weather.

Moreover, image restoration methods (Cui et al. 2023; Valanarasu, Yasarla, and Patel 2022) focus on converting images with noise or adverse weather conditions into clean images, facilitating diverse applications. For instance, Restormer (Zamir et al. 2022) tailored to multi-scale local-global representation learning to enhance image restoration ability. Although these models perform well, they still have high complexity. Therefore, we propose an auxiliary structure that shares features and an object-centric method to efficiently operate with multimodal 3D object detection.

Unsupervised Domain Adaptation

The UDA method addresses performance drops between labeled source and unlabeled target data. Recent UDA methods in 2D object detection include adversarial training (Saito et al. 2019) and distribution alignment (Lu et al. 2024).

For 3D object detection, UDA approaches have focused on such strategies as enhancing model robustness via closing the domain gap (Tsai et al. 2023), feature alignment (Chang et al. 2024), and pseudo-labeling (Yang et al. 2022b). They

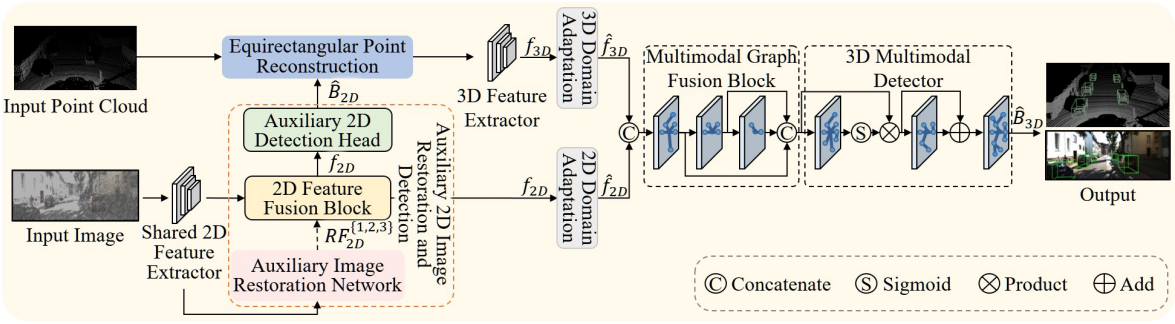

Figure 2: Overall architecture of the EquiDetect network.

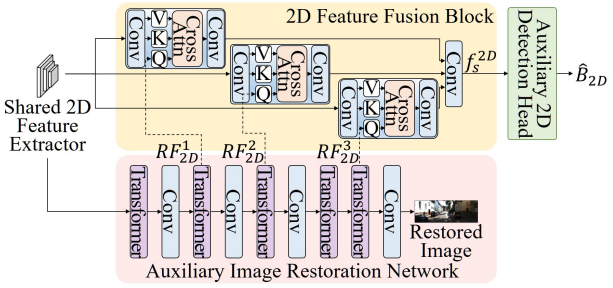

Figure 3: Auxiliary 2D image restoration and detection networks.

minimized the distance between the mean distributions of various domains, but large domain gaps resulted in the dispersion of gradient distribution in fine-grained parts. Therefore, we propose the feature perturbation method, which precisely aligns feature gradients between domains.

Method

Overview

As depicted in Fig. 2, we propose EquiDetect, a multimodal 3D object detection method robust to adverse weather conditions. EquiDetect uses 3D point clouds and 2D images. For a image, EquiDetect restores image features degraded by adverse weather via an auxiliary image restoration network. In the 2D feature fusion block, EquiDetect fuses the restored feature with a feature from a shared 2D feature extractor, yielding a fused feature, f_{2D}. For the point cloud, a range image transformed via equirectangular projection is employed for point denoising and generation in an equirectangular point reconstruction network. This network applies the ray properties of LiDAR and object bounding boxes $\hat{B}_{2D}$ estimated from the auxiliary 2D detection head. The point cloud reconstructed from this network is input into a 3D feature extractor, leading to an extracted feature f_{3D}. The f_{2D} and f_{3D} are subjected to domain adaptation (DA) to reduce the domain gap between the datasets, resulting in $\hat{f}_{2D}$ and $\hat{f}_{3D}$. These features are integrated into a graph representation via a multimodal graph fusion block. Lastly, a 3D multimodal detector predicts the 3D bounding boxes $\hat{B}_{3D}$.

Auxiliary 2D Image Restoration and Detection

As presented in Fig. 3, the auxiliary image restoration and detection network addresses the adverse weather problem in images and supports the point reconstructor. The DLA-34 (Yu et al. 2018) is a shared 2D feature extractor, and the auxiliary image restoration network was inspired by Restormer (Zamir et al. 2022), known for its excellent image restoration performance. The features from the shared 2D feature extractor are input into the auxiliary image restoration network with a hierarchical structure of transformers and convolutional layers. The output from this network is a quarter of the size of the original image and is compared with the clean image, as follows:

$$\mathcal{L}_{aux} = \mathcal{L}_p + \mathcal{L}_{L1}, \tag{1}$$

where $\mathcal{L}_p$ denotes the perceptual loss (Johnson, Alahi, and Fei-Fei 2016) to compare features between the clean and restored images. Moreover, $\mathcal{L}_{L1}$ measures the mean absolute error between each element in two images.

Additionally, the restored features $RF_{2D}^{\{1,2,3\}}$ are combined with features from the shared 2D feature extractor in a 2D feature fusion block via cross-attention as follows:

$$CrossAttn(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V\,, \tag{2}$$

where Q represents the query matrix produced by the auxiliary image restoration network. Further, K and V represent the key and value matrices, respectively, from the shared 2D feature extractor, and D indicates the channel dimension. The restored features aid in the object localization via the cross-attention. Further, this auxiliary structure operates efficiently by using these features during the decoding process.

The hierarchically fused features pass through a final convolutional layer to obtain the fused feature f_{2D}. Additionally, f_{2D} is input into a 2D detection head inspired by CenterNet (Zhou, Wang, and Krähenbühl 2019) to predict $\hat{B}_{2D}$. Through 2D detection, objects that are challenging to detect in the highly sparse point cloud can be more effectively identified, and this information is employed for point cloud generation. The 2D detection is trained using the $\mathcal{L}_{2Ddet}$ loss, and Appendix provides the details.

Equirectangular Point Reconstruction

Distance-constrained Denoising LiDAR is highly effective in accurately detecting objects using pulsed lasers. How-

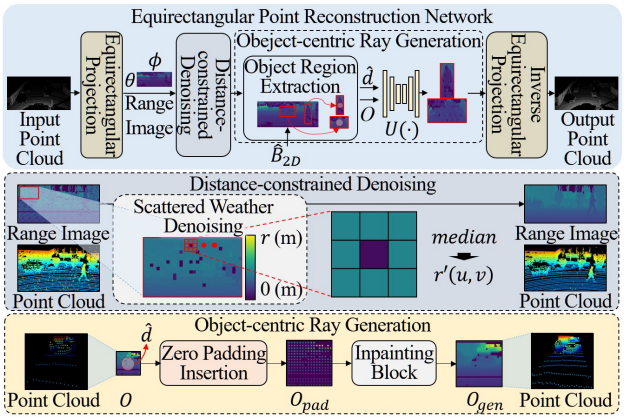

Figure 4: Equirectangular point reconstruction network.

ever, LiDAR has a significant limitation: its vulnerability to adverse weather conditions due to occlusion and signal attenuation. The characteristics of pulsed lasers and the weather noise scattered as small particles throughout the space can result in a shortened distance or lost points, similar to pepper noise, as observed in the middle row in Fig. 4.

Figure 4 depicts the proposed distance-constrained denoising method with a dense 2D representation to remove the scattered noise. LiDAR operates with vertical channels and generates points horizontally as it rotates with pulsed lasers. Based on this, the point cloud can be converted from Cartesian coordinates (x, y, z) to spherical coordinates (r, θ, ϕ). Equirectangular projection (Nakashima and Kurazume 2023) can transform the point cloud representation into a range image with dimensions of $2\times H\times W$ expressing θ and ϕ as H and W, respectively. The distance r and the intensity are encoded in the pixel values of this 2D grid. This method transforms a sparse point cloud into a dense format.

In this representation, the distance-constrained denoising method is applied to remove the weather noise, which has the pepper noise property, as follows:

$$r'(u,v) = median\{r(u+i, v+j) \mid (i,j) \in R\}\ , \quad (3)$$

where $r'(u,v)$ represents the distance initially filtered (Tukey 1974) at the pixel coordinates (u,v) in the range image. Moreover, $(i,j) \in R$ denotes the relative coordinates, with $i,j \in \{-1,0,1\}$. The occlusion or missing points due to weather noise tend to reduce the r values compared to those of the points in a clear weather. Based on this property, $\hat{r}$ is constrained to no smaller than r as follows:

$$\hat{r}(u,v) = \begin{cases} r'(u,v) & r'(u,v) > r(u,v) \\ r(u,v) & r'(u,v) \leq r(u,v) \end{cases}. \quad (4)$$

Object-centric Ray Generator Existing methods yield the decent performance for the dense close points, whereas it is difficult for these methods to deal with the sparse distant points optimally (Lai et al. 2023). Although point generation methods (Nakashima and Kurazume 2023; Zyrianov, Zhu, and Wang 2022) using range images with diffusion models have demonstrated excellent performance, they tend to have high complexity and are unsuitable for real-time applications. To address this issue, we propose an object-centric ray generation method with a range image at the object-level.

As illustrated in Fig. 4, the object region of $\hat{B}_{2D}$ obtained from the image is used, and object patch O in the range image is extracted via the point information projected in the $\hat{B}_{2D}$ region. Then, the object distance is estimated using the average $\hat{r}$ value at the center of O as follows:

$$\hat{d} = \frac{1}{9}\sum_{i=-1}^{1}\sum_{j=-1}^{1} O(u_c+i, v_c+j), \quad (5)$$

where (u_c, v_c) denotes the center coordinate of the object patch O, and $\hat{d}$ denotes the estimated distance to the object. Using the obtained value of $\hat{d}$, the points are generated based on the optimal number of points ($\star$) observed in Fig. 1(c). For point generation in 2D representation, the U-Net and inpainting approach are applied as in the work by Nakashima and Kurazume (2023). Zero padding is inserted between the pixels of the object patch. The padded object patch O_{pad} is processed with a zero-padding mask M to generate an inpainted object patch O_{gen} in the range image as follows:

$$O_{gen} = O_{pad} \odot M + U(O_{pad}) \odot (1 - M), \quad (6)$$

where $U(\cdot)$ denotes the ray generator that fills in the zero-padding space, and $\odot$ represents the element-wise production. This equation preserves the existing points, whereas new points are generated in the masked regions.

For ray generator training, a 3D loss function calculated with O_{gen} is employed as follows:

$$\mathcal{L}_{3D} = \mathcal{L}_G + \mathcal{L}_p + \mathcal{L}_{L1}, \quad (7)$$

$$\mathcal{L}_G = ||O_{gt}G_x - O_{gen}G_x||_1 + ||O_{gt}G_y - O_{gen}G_y||_1, \quad (8)$$

where G_x and G_y denote the gradient operators along the horizontal and vertical directions. Further, O_{gt} represents the ground truth (GT) object patch in the clean range image. This pretrained ray generator is used for detection.

Domain Adaptation

The existing DA methods reduce the mean distribution between the source and target domains. However, when features of two domains are distributed differently, their gradients in non-overlapping regions may disperse in distinct parts of the highly complicated decision boundary. This large domain shift leads to a large gradient distribution discrepancy (Gao et al. 2021). To address this problem, we propose a perturbation method that adjusts the gradients of the target domain features to the align fine-grained distribution between the source and target domains.

As depicted in Fig. 2, the 3D and 2D DA modules take f_{3D} and f_{2D} as input, respectively, and both modules follow the same process, except for the dimension of the module layers. Additionally, DA is applied during the training and inference stages. As illustrated in the top row in Fig. 5, the mean distribution discrepancy between the two domains is minimized in the training stage, as in the work by Kobayashi et al. (2017). Thus, the pretrained feature extractor is employed using source data. The source feature f^s and

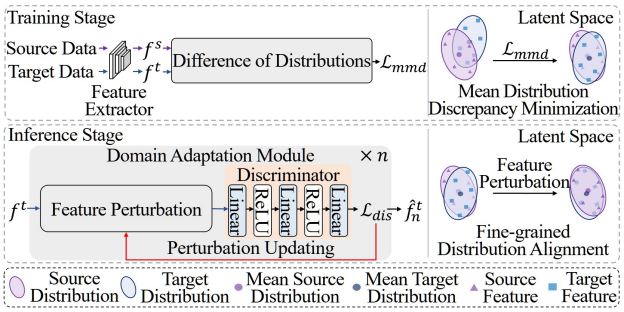

Figure 5: Domain adaptation module.

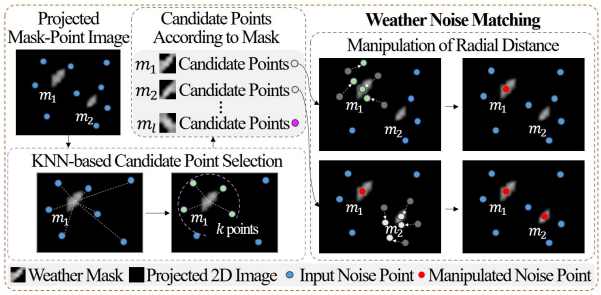

Figure 6: LiDAR-camera weather noise matching.

target feature f^t are extracted using the pretrained feature extractor. Then, the extractor is fine-tuned using the maximum mean discrepancy (MMD) loss as follows:

$$\mathcal{L}_{mmd} = \left\| \frac{1}{n^s} \sum_{p=1}^{n^s} f_p^s - \frac{1}{n^t} \sum_{q=1}^{n^t} f_q^t \right\|_2, \tag{9}$$

where n^s and n^t denote the number of samples in the source and target domains. The fine-tuned extractor minimizes the mean distribution discrepancy between the two domains, improving the detection performance in the target domain.

This MMD loss can reduce the domain shift, but aligning a fine distribution has limitations. To address these limitations, in the inference stage, perturbation is injected into the f^t to align the fine-grained distribution of the two domains gradually. As shown in the bottom row in Fig. 5, the DA module consists of feature perturbation and a discriminator. The discriminator is an anomaly detector comprising of linear layers and ReLU activation functions and is pretrained with f^s. The discriminator loss $\mathcal{L}_{dis}$ uses the one-class deep SVDD (Ruff et al. 2018) to learn to distinguish between the two domain features. The deep SVDD is defined as follows:

$$\mathcal{L}_{dis} = \frac{1}{n^s} \sum_{p=1}^{n^s} \|\Psi(f_p^s; \mathcal{W}) - c\|_2, \tag{10}$$

where $\mathcal{W}$ represents the weights of the discriminator, and $\Psi(f_p^s; \mathcal{W})$ denotes the feature representation of f_p^s extracted using discriminator $\Psi(\cdot\,; \mathcal{W})$. Furthermore, c represents the hypersphere center. This discriminator applies feature perturbations inspired by Madry et al. (2017). The fine-grained feature perturbations are applied to f^t to align the distributions between the two domains by reducing $\mathcal{L}_{dis}$ as follows:

$$f_{h+1}^t = \Pi_{f^t,\epsilon}(f_h^t - \alpha \cdot sign(\nabla_{f^t} \mathcal{L}_{dis})), \tag{11}$$

where $\Pi_{f^t,\epsilon}$ denotes the projection operator that maps the updated value in the ϵ (set to 0.03) range centered at f^t, and f_h^t represents the perturbed target feature at the h-th iteration. Furthermore, α (set to 0.01) indicates the step size for each perturbation update, and $sign(\nabla_{f^t} \mathcal{L}_{dis})$ signifies the gradient sign of $\mathcal{L}_{dis}$ for input f^t. The DA module performs g iterations to output the iteratively perturbed target feature $\hat{f}^t$. Thus, the constrained perturbations allow the fine-grained alignment of the anomaly target features with the source domain distributions via the discriminator.

Multimodal Graph Fusion and Detector

Figure 2 depicts the multimodal graph fusion block that receives $\hat{f}_{3D}$ and $\hat{f}_{2D}$ from the 3D and 2D DA modules as input. These inputs are processed through hierarchical layers that extract features at multiple scales and consist of graph layers with three hidden dimensions (64, 32, and 16). The structure sequentially processes the inputs, from the largest kernel size to the smallest. Each layer output is concatenated to create a comprehensive feature representation of the nodes. Therefore, these layers capture multi-scale spatial relationships and patterns between 3D and 2D features. Subsequently, the concatenated features are input into the 3D multimodal detector, which is a graph neural network inspired by Yang et al. (2022a), to predict the $\hat{B}_{3D}$.

Overall Training Process

Weather Noise Matching via Radial Manipulation Some existing simulators operate on a single modality (LiDAR or camera), and each independent simulated dataset does not align between two domains. To address this issue, we propose a multimodal noise matching method that aligns the simulated point cloud and image.

Weather noise points are generated for LiDAR using physical-based simulators (Hahner et al. 2022; Kilic et al. 2021) to obtain multimodal synthetic data. Moreover, weather particles are added for images using the process by Hendrycks and Dietterich (2019). Figure 6 depicts the employment of the projected mask-point image, representing the weather particles $m_{o\in\{1,2,\cdots,l\}}$ and noise points.

The weather noise points arise when the straight laser is reflected by weather particles. The noise points are manipulated in the radial direction based on these physical characteristics to align with the particle in the projected mask-point image. Furthermore, the k-nearest neighbor-based candidate point selection is used to extract the k-nearest candidate points for each weather particle for efficient mask-point matching. This process selects candidate points $candi_{e\in\{1,2,\cdots,k\}}$ for all particles, which are manipulated in the radial direction as follows:

$$candi'_e = (r_{candi_e} + \Delta\hat{\beta}_e, \theta_{candi_e}, \phi_{candi_e}), \tag{12}$$

$$\Delta\hat{\beta}_e = \underset{\Delta\beta_e}{\operatorname{argmin}} \|(u_{m_o}, v_{m_o}) - proj(r_{candi_e} + \Delta\beta_e, \theta_{candi_e}, \phi_{candi_e})\|_2, \tag{13}$$

where r_{candi_e}, θ_{candi_e}, and ϕ_{candi_e} denote r, θ, and ϕ of $candi_e$, respectively. Further, $candi'_e$ is the manipulated candidate noise point. Moreover, $\Delta\beta_e$ denotes the manipulation required to minimize the distance to the particle in the mask-point image, and $\Delta\hat{\beta}_e$ represents the optimal value. Furthermore, $proj(\cdot)$ denotes a function that projects the spherical coordinates onto the image coordinates. Additionally, (u_{m_o}, v_{m_o}) denotes the central coordinate of the o-th particle, and $r_{candi_e} + \Delta\beta_e$ are limited to the minimum and maximum range that the LiDAR can measure. Among $candi'_{e\in 1,2,\cdots,k}$, the matched point that overlaps with the particle and is closest to its center is selected.

Total Objective During EquiDetect training, regression loss $\mathcal{L}_{reg}$ is utilized to measure the difference between the predicted and GT 3D bounding boxes. Additionally, cross-entropy loss $\mathcal{L}_{cls}$ is used for object classification. The detector loss function is a combination of these losses as follows:

$$\mathcal{L}_{3Ddet} = \mathcal{L}_{reg} + \mathcal{L}_{cls} \ . \tag{14}$$

Thus, the total objective $\mathcal{L}_{total}$ is defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{3Ddet} + \mathcal{L}_{2Ddet} + \lambda_{aux}\mathcal{L}_{aux} \ , \tag{15}$$

where λ_{aux} denotes a weight of $\mathcal{L}_{aux}$. This objective is designed to guide the network by merging the detection and auxiliary losses.

Experiments

Dataset

The KITTI dataset (Geiger et al. 2013) consists of 7,481 LiDAR and image frames, divided into a training set of 3,712 and a validation set of 3,769. Performance of the models was evaluated on the 3D average precision under 40 recall thresholds ($3D_{R40}AP$). This evaluation metric has an intersection over union (IoU) threshold of 0.7. For adverse weather conditions (snow, rain, and fog), the synthetic point cloud was created by using simulators (Hahner et al. 2022; Kilic et al. 2021; Hahner et al. 2021) in LiDAR. The Python library (Hendrycks and Dietterich 2019) was employed to create synthetic weather images. Moreover, the proposed multimodal weather noise matching method aligns synthetic point clouds and images in snow and rain weather conditions. Using this method, matching accuracies of 93.25% and 94.51% were achieved under snow and rain conditions, respectively, compared to 28.77% and 31.49% by existing simulators (Hahner et al. 2022; Kilic et al. 2021; Hendrycks and Dietterich 2019). Based on this effective matching method, the realistic multimodal synthetic dataset S-KITTI was generated for the experiment.

The CADC (Pitropov et al. 2021) dataset is for a 3D object detection in real-world snow weather condition. The dataset comprises 4,225 LiDAR and images for training and 1,238 for validation. This dataset was converted into the KITTI format and evaluated with an IOU threshold of 0.5.

The Dense (Bijelic et al. 2020) dataset for 3D object detection contains real-world weather conditions (clear, snow, rain, and fog). For training, 3,526, 3,786, 225, and 1,551 LiDAR and image frames were used for clear, snow, rain, and fog conditions, respectively. For validation, 808, 947, 57, and 388 frames were used for the same conditions. An IoU threshold of 0.5 was used for evaluation.

Metric	$3D_{R40}\,AP(\%)$			
Weather	Clear	Snow	Rain	Fog
LiDAR-based				
PV-RCNN (Shi et al. 2020)	80.66	63.99	64.01	69.18
VoxelNeXt (Chen et al. 2023)	74.50	54.56	54.78	68.01
GD-MAE (Yang et al. 2023)	78.23	56.77	56.00	67.03
HINTED (Xia et al. 2024)	83.11	68.76	67.81	73.09
Multimodal				
Focals Conv (Chen et al. 2022)	84.21	66.92	67.15	71.07
Graph-VoI (Yang et al. 2022a)	85.82	67.12	67.20	70.93
SFD (Wu et al. 2022b)	83.91	68.17	68.14	72.69
TED-M (Wu et al. 2023a)	85.89	70.21	67.58	78.54
VirConv-S (Wu et al. 2023b)	87.02	72.17	72.20	80.23
Ours	**87.89**	**77.86**	**78.22**	**82.85**

Table 1: Results of 3D object detectors on S-KITTI for weather conditions by car class at moderate difficulty.

Metric	$3D_{R40}\,AP(\%)$				
Data	Dense				CADC
Weather	Clear	Snow	Rain	Fog	Snow
LiDAR-based					
PV-RCNN (Shi et al. 2020)	35.40	33.44	29.57	27.57	29.06
VoxelNeXt (Chen et al. 2023)	33.68	30.65	24.84	25.65	26.47
GD-MAE (Yang et al. 2023)	36.54	33.07	32.57	30.89	26.54
HINTED (Xia et al. 2024)	32.75	32.53	31.07	30.66	37.39
Multimodal					
Focals Conv (Chen et al. 2022)	34.16	31.57	30.46	24.57	35.34
Graph-VoI (Yang et al. 2022a)	37.81	32.72	32.64	31.75	42.41
SFD (Wu et al. 2022b)	31.60	24.86	23.83	21.78	28.23
TED-M (Wu et al. 2023a)	33.97	30.72	29.10	25.06	42.81
VirConv-S (Wu et al. 2023b)	39.04	36.32	35.32	34.92	43.18
Ours (w/o DA)	**40.09**	**37.91**	**37.07**	**35.29**	**45.66**

Table 2: Results of 3D object detectors on the Dense and CADC datasets for weather conditions based on the car class at moderate difficulty.

Implementation Details

This study set λ_{aux} to 0.1 in total objective, $\mathcal{L}_{total}$, and k to 4 in the k-nearest neighbor algorithm. The voxel size for the 3D representation was set to $0.05\times0.05\times0.1$, whereas the image input had a resolution of 1280×384. During training, 80 epochs were conducted with a learning rate initialized to 0.0001, a momentum of 0.9, and a weight decay of 0.01 using the Adam optimizer. The EquiDetect network was trained on an RTX-3080 GPU.

Main Results

This section compares EquiDetect with existing SOTA models. The 3D detection models in Tables 1 and 2, DA models in Table 3, and restoration models in Table 4 have commonly been trained on the S-KITTI dataset for all weather conditions at once. The best scores in tables are marked in bold.

Table 1 compares the proposed model with the existing LiDAR-based and multimodal models on the S-KITTI dataset, focusing on the car class at a moderate difficulty. Although the SOTA models are degraded in adverse weather conditions, the proposed model exhibits outstanding performance across all weather conditions. These results confirm the effectiveness of the equirectangular point reconstruction and auxiliary modules.

Metric		$3D_{R40}\,AP(\%)$				
Data		S-KITTI → Dense				S-KITTI → CADC
DA	Detection	Clear	Snow	Rain	Fog	Snow
ST3D++	Focals Conv	41.90	38.95	38.15	32.81	40.09
+ NSA	Graph-VoI	44.01	40.46	41.03	37.44	45.58
AttProto	Focals Conv	42.27	38.22	38.14	33.80	42.17
+ NSA	Graph-VoI	44.08	39.81	41.33	38.10	47.38
Ours		**48.84**	**45.87**	**45.68**	**42.82**	**52.09**

Table 3: Results of 3D object detectors with DA models for weather conditions by car class at moderate difficulty.

Metric		$3D_{R40}\,AP(\%)$			
Restoration	Detection	Clear	Snow	Rain	Fog
Restormer	TED-M	81.17	72.16	70.49	79.30
	VirConv-S	85.26	74.06	73.54	79.26
FocalNet	TED-M	82.61	72.63	66.72	80.31
	VirConv-S	86.70	73.01	70.35	81.12
Ours		**87.89**	**77.86**	**78.22**	**82.85**

Table 4: Results of 3D object detectors with restoration models for weather conditions by car class at moderate difficulty.

Table 2 presents the evaluation of 3D object detection models on real-world adverse weather datasets (Dense and CADC). These models, along with EquiDetect without DA modules, assess generalization performance. Compared with the SOTA models, EquiDetect achieves higher performance across all real-world weather conditions. This results is attributed to the auxiliary learning and the equirectangular point reconstructon based on LiDAR characteristics.

Table 3 evaluates the performance of multimodal 3D object detection models integrated with 3D and 2D DA methods on real-world adverse weather datasets (Dense and CADC). This study employs ST3D++ (Yang et al. 2022b) and AttProto (Hegde and Patel 2024) for 3D DA, and NSA (Zhou et al. 2023) for 2D DA to detection models. Further, we employed Focals Conv and Graph-VoI, late fusion methods that can be integrated with DA models for multimodal 3D object detection. These models were trained on S-KITTI with labels and further trained on the Dense or CADC datasets without labels. The proposed model outperformed Focals Conv and Graph-VOI combined with SOTA 3D and 2D DA methods. Moreover, compared with the results in Table 2, the improvement rate of other combined methods in Table 3 is $7.07\%p$, whereas the proposed method has an increase of $7.86\%p$, demonstrating its effectiveness. These results indicate the superiority of the feature perturbations in alleviating fine domain discrepancies.

Table 4 evaluates the performance of SOTA 3D object detection models on the S-KITTI dataset when integrated with the image restoration models, Restormer (Zamir et al. 2022) and FocalNet (Cui et al. 2023), in adverse weather conditions. The compared detection models were trained on output from image restoration models. The proposed model outperformed the independently combined models. The sequentially combined models heavily depend on the results of the restoration model, whereas our auxiliary learning with the total loss can reduce this dependency on restoration, leading to more optimal detection performance.

Table 5 compares the complexity of EquiDetect with that of multimodal 3D object detection models regarding FLOPs, Params, and inference time on the S-KITTI dataset. Although the proposed model has slightly higher parameters than the existing multimodal methods, it achieves the second-best performance concerning inference time, crucial indicator of real-time capability. The object-centric point generation, an auxiliary structure, and simplified denoising in the range image contribute to this time efficiency.

Metric	FLOPs (G)	Params (M)	Time (ms)
Focals Conv	405.7	30.4	202.6
Graph-VoI	118.3	17.1	119.3
SFD	57.4	31.5	761.2
TED-M	90.6	14.4	139.1
VirConv-S	176.8	13.7	345.2
Ours	174.4	36.2	122.9

Table 5: Complexity of 3D object detectors on S-KITTI.

Auxiliary Image Restoration Network	Distance-constrained Denoising	Ray Generator	$3D_{R40}\,AP(\%)$
✓	✓	✓	81.71
✓	✓		80.29
✓		✓	76.89
	✓	✓	78.64
			73.12

Table 6: Ablation study for the equirectangular point reconstruction module on the S-KITTI at moderate difficulty.

Ablation Study

Table 6 analyzes the EquiDetect model's mean $3D_{R40}AP$ across all weather conditions on the S-KITTI dataset. The checkmark (✓) indicates a module was activated. The first row presents the baseline performance of the entire model, incorporating all modules. The second row reveals the performance without the ray generator. The third row lists the results of not undergoing the distance-constrained denoising process. The fourth row reveals the performance without the auxiliary image restoration network. The final row exhibits the backbone performance. When comparing the first row with the second through the final rows, the experimental results indicate that all components contributed to the performance improvements.

Conclusion

This paper addresses the multimodal 3D object detection limitations in adverse weather conditions. Given the characteristics of pulsed laser-based LiDAR, we proposed an equirectangular point reconstructor. Projecting the 3D point cloud into a 2D representation and applying a distance-constrained denoising method, we effectively removed sparse weather noise. Additionally, we proposed the object-centric ray generator, which reconstructs the points of objects in the 2D representation to aid in precise 3D detection. Moreover, we introduced feature perturbations regarding UDA, enabling effective object detection even in data- and label-limited environments. Last, we proposed a LiDAR-camera multimodal adverse weather noise matching approach to construct a more realistic multimodal synthetic dataset. EquiDetect operates effectively and achieves SOTA performance in synthetic and real-world domains.

Acknowledgements

This paper was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0020536) and the IITP grant funded by the Korea government (MSIT) (No. 2021-0-02068, RS-2022-00156287, RS-2023-00256629, RS-2024-00437718).

References

Akhtar, A.; Li, Z.; Van der Auwera, G.; Li, L.; and Chen, J. 2022. Pu-dense: Sparse tensor-based point cloud geometry upsampling. *IEEE Transactions on Image Processing*, 31: 4133–4148.

Bijelic, M.; Gruber, T.; Mannan, F.; Kraus, F.; Ritter, W.; Dietmayer, K.; and Heide, F. 2020. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11682–11692.

Chae, Y.; Kim, H.; Oh, C.; Kim, M.; and Yoon, K.-J. 2024. LiDAR-Based All-Weather 3D Object Detection via Prompting and Distilling 4D Radar. In *European Conference on Computer Vision*, 368–385. Springer.

Chae, Y.; Kim, H.; and Yoon, K.-J. 2024. Towards Robust 3D Object Detection with LiDAR and 4D Radar Fusion in Various Weather Conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15162–15172.

Chang, G.; Roh, W.; Jang, S.; Lee, D.; Ji, D.; Oh, G.; Park, J.; Kim, J.; and Kim, S. 2024. CMDA: Cross-Modal and Domain Adversarial Adaptation for LiDAR-Based 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 972–980.

Chen, Y.; Li, Y.; Zhang, X.; Sun, J.; and Jia, J. 2022. Focal sparse convolutional networks for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5428–5437.

Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; and Jia, J. 2023. 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.

Cui, Y.; Ren, W.; Cao, X.; and Knoll, A. 2023. Focal network for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, 13001–13011.

Dong, Y.; Kang, C.; Zhang, J.; Zhu, Z.; Wang, Y.; Yang, X.; Su, H.; Wei, X.; and Zhu, J. 2023. Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1022–1032.

Gao, Z.; Zhang, S.; Huang, K.; Wang, Q.; and Zhong, C. 2021. Gradient distribution alignment certificates better adversarial domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8937–8946.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237.

Hahner, M.; Sakaridis, C.; Bijelic, M.; Heide, F.; Yu, F.; Dai, D.; and Van Gool, L. 2022. Lidar snowfall simulation for robust 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16364–16374.

Hahner, M.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2021. Fog simulation on real LiDAR point clouds for 3D object detection in adverse weather. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15283–15292.

Hegde, D.; and Patel, V. M. 2024. Attentive Prototypes for Source-Free Unsupervised Domain Adaptive 3D Object Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3066–3076.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Huang, X.; Wu, H.; Li, X.; Fan, X.; Wen, C.; and Wang, C. 2024a. Sunshine to rainstorm: Cross-weather knowledge distillation for robust 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2409–2416.

Huang, Y.; Zhou, S.; Zhang, J.; Dong, J.; and Zheng, N. 2024b. Voxel or Pillar: Exploring Efficient Point Cloud Representation for 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2426–2435.

Jin, X.; Liu, K.; Ma, C.; Yang, R.; Hui, F.; and Wu, W. 2024. SwiftPillars: High-Efficiency Pillar Encoder for Lidar-Based 3D Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2625–2633.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 694–711. Springer.

Kilic, V.; Hegde, D.; Sindagi, V.; Cooper, A. B.; Foster, M. A.; and Patel, V. M. 2021. Lidar light scattering augmentation (lisa): Physics-based simulation of adverse weather conditions for 3d object detection. *arXiv preprint arXiv:2107.07004*.

Kobayashi, H.; Lei, C.; Wu, Y.; Mao, A.; Jiang, Y.; Guo, B.; Ozeki, Y.; and Goda, K. 2017. Label-free detection of cellular drug responses by high-throughput bright-field imaging and machine learning. *Scientific reports*, 7(1): 12454.

Lai, X.; Chen, Y.; Lu, F.; Liu, J.; and Jia, J. 2023. Spherical transformer for lidar-based 3d recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17545–17555.

Li, B.; Li, J.; Chen, G.; Wu, H.; and Huang, K. 2022. Desnowing LiDAR point clouds with intensity and spatial-temporal features. In *2022 International Conference on Robotics and Automation (ICRA)*, 2359–2365. IEEE.

Li, Z.; Wang, F.; and Wang, N. 2021. Lidar r-cnn: An efficient and universal 3d object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7546–7555.

Lu, Y.; Shen, M.; Ma, A. J.; Xie, X.; and Lai, J.-H. 2024. MLNet: Mutual Learning Network with Neighborhood Invariance for Universal Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3900–3908.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mai, N. A. M.; Duthon, P.; Salmane, P. H.; Khoudour, L.; Crouzil, A.; and Velastin, S. A. 2022. Camera and LiDAR analysis for 3D object detection in foggy weather conditions. In *2022 12th International Conference on Pattern Recognition Systems (ICPRS)*, 1–7. IEEE.

Nakashima, K.; and Kurazume, R. 2023. Lidar data synthesis with denoising diffusion probabilistic models. *arXiv preprint arXiv:2309.09256*.

Palladin, E.; Dietze, R.; Narayanan, P.; Bijelic, M.; and Heide, F. 2024. Samfusion: Sensor-adaptive multimodal fusion for 3d object detection in adverse weather. In *European Conference on Computer Vision*, 484–503. Springer.

Pitropov, M.; Garcia, D. E.; Rebello, J.; Smart, M.; Wang, C.; Czarnecki, K.; and Waslander, S. 2021. Canadian adverse driving conditions dataset. *The International Journal of Robotics Research*, 40(4-5): 681–690.

Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep one-class classification. In *International conference on machine learning*, 4393–4402. PMLR.

Saito, K.; Ushiku, Y.; Harada, T.; and Saenko, K. 2019. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6956–6965.

Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10529–10538.

Shi, W.; and Rajkumar, R. 2020. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1711–1719.

Tsai, D.; Berrio, J. S.; Shan, M.; Nebot, E.; and Worrall, S. 2023. Viewer-centred surface completion for unsupervised domain adaptation in 3D object detection. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 9346–9353. IEEE.

Tukey, J. 1974. Nonlinear (nonsuperposable) methods for smoothing data. CONGRESS RECORD (EASCO).

Valanarasu, J. M. J.; Yasarla, R.; and Patel, V. M. 2022. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2353–2363.

Wu, H.; Deng, J.; Wen, C.; Li, X.; Wang, C.; and Li, J. 2022a. CasA: A cascade attention network for 3-D object detection from LiDAR point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–11.

Wu, H.; Wen, C.; Li, W.; Li, X.; Yang, R.; and Wang, C. 2023a. Transformation-equivariant 3D object detection for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2795–2802.

Wu, H.; Wen, C.; Shi, S.; Li, X.; and Wang, C. 2023b. Virtual Sparse Convolution for Multimodal 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21653–21662.

Wu, X.; Peng, L.; Yang, H.; Xie, L.; Huang, C.; Deng, C.; Liu, H.; and Cai, D. 2022b. Sparse fuse dense: Towards high quality 3d detection with depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5418–5427.

Xia, Q.; Ye, W.; Wu, H.; Zhao, S.; Xing, L.; Huang, X.; Deng, J.; Li, X.; Wen, C.; and Wang, C. 2024. HINTED: Hard Instance Enhanced Detector with Mixed-Density Feature Fusion for Sparsely-Supervised 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15321–15330.

Xu, Q.; Zhong, Y.; and Neumann, U. 2022. Behind the curtain: Learning occluded shapes for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2893–2901.

Yang, H.; He, T.; Liu, J.; Chen, H.; Wu, B.; Lin, B.; He, X.; and Ouyang, W. 2023. GD-MAE: generative decoder for MAE pre-training on lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9403–9414.

Yang, H.; Liu, Z.; Wu, X.; Wang, W.; Qian, W.; He, X.; and Cai, D. 2022a. Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph. In *European Conference on Computer Vision*, 662–679. Springer.

Yang, J.; Shi, S.; Wang, Z.; Li, H.; and Qi, X. 2022b. St3d++: Denoised self-training for unsupervised domain adaptation on 3d object detection. *IEEE transactions on pattern analysis and machine intelligence*, 45(5): 6354–6371.

Yu, F.; Wang, D.; Shelhamer, E.; and Darrell, T. 2018. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2403–2412.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zeng, G.; Li, H.; Wang, X.; and Li, N. 2021. Point cloud up-sampling network with multi-level spatial local feature aggregation. *Computers & Electrical Engineering*, 94: 107337.

Zhou, W.; Fan, H.; Luo, T.; and Zhang, L. 2023. Unsupervised domain adaptive detection with network stability analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6986–6995.

Zhou, X.; Wang, D.; and Krähenbühl, P. 2019. Objects as points. *arXiv preprint arXiv:1904.07850*.

Zyrianov, V.; Zhu, X.; and Wang, S. 2022. Learning to generate realistic lidar point clouds. In *European Conference on Computer Vision*, 17–35. Springer.