2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MonoEM: Object-level Monocular 3D Object Detection Based on Equirectangular Map under Inclement Weather

Jae Hyun Yoon†, Yeon Woo Cho† and Seok Bong Yoo*

***Abstract*— Monocular 3D object detection has received growing recognition in contemporary research due to its reduced hardware complexity and lower deployment cost compared to multi-sensor-based approaches. Prior research has primarily addressed ideal environmental settings, neglecting the influence of diverse weather scenarios, including rain, snow, and fog, that significantly hinder detection reliability. To enhance robustness under inclement weather conditions, we introduce MonoEM, a monocular 3D object detection framework that leverages object-level image representations and equirectangular maps. Starting from 2D detection results, MonoEM derives equirectangular maps through an equirectangular object-level reconstruction. Furthermore, MonoEM suppresses inclement weather noise in object-level images through image restoration. Subsequently, MonoEM fuses the reconstructed equirectangular map with the restored image and performs 3D bounding box prediction using a visual-range fusion detector. The integration of 2D-3D box alignment loss between 2D and 3D bounding boxes improves the geometric alignment and 3D object detection accuracy. Experimental results across various inclement weather conditions validate the notable accuracy and robustness of MonoEM compared to existing monocular 3D baselines. The source code is provided at https://github.com/yeonwoo29/MonoEM.**

I. Introduction

Monocular 3D object detection has emerged as a practical solution for applications like autonomous driving. Unlike approaches that rely on LiDAR or stereo vision [1]–[5], it operates using a single RGB camera, which simplifies hardware requirements and significantly reduces cost. These benefits have led to increased research interest in monocular approaches. However, most existing studies focus on clear-weather scenarios, making detection methods less robust under challenging weather conditions commonly encountered in real-world environments.

Previous approaches have utilized distance-based cues, such as depth maps or pseudo-LiDAR, to infer 3D information. However, these techniques generally assume clear weather conditions and overlook the impact of weather-induced particles on distance measurements. As shown in Fig. 1(a), a point cloud can be expressed in spherical coordinates (r, θ, ϕ) and subsequently converted into an equirectangular map [6]. Under inclement weather, noise

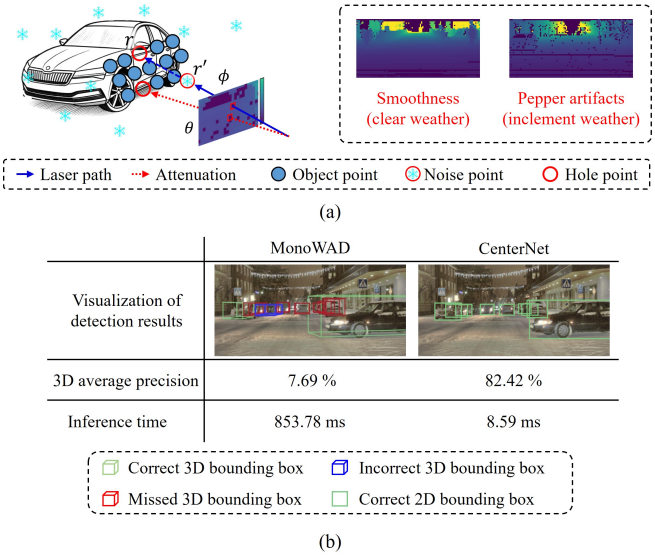

Fig. 1: (a) Equirectangular maps under inclement weather exhibit pepper noise artifacts caused by occlusion and laser signal attenuation. (b) Performance comparison between 3D and 2D object detectors on the Dense dataset.

can interfere with laser signals, either by reflecting them prematurely or attenuating them, resulting in a measured range r' that is shorter than the actual distance r. This weather-induced noise appears as pepper-like artifacts in the equirectangular map, as illustrated in Fig. 1(a). Using the equirectangular map, which captures such noise patterns, enables more effective identification and correction of noise from inclement weather.

Another important insight is that monocular 3D detectors often exhibit higher miss or false detection rates compared to their 2D counterparts. As visualized in Fig. 1(b), whereas the state-of-the-art (SOTA) monocular 3D detector MonoWAD yields some missed and inaccurate predictions with a high inference time of 854ms, the 2D detector CenterNet [7] successfully identifies most objects within about 9ms, even under challenging weather. This highlights the difficulty of directly estimating 3D information from monocular inputs, which increases detection uncertainty. In contrast, 2D detection offers low computational cost and more reliable region proposals [8], which can be effectively leveraged to facilitate accurate and efficient 3D localization. While 2D detectors maintain high recall under inclement weather, their localization precision often degrades due to blur and partial occlusions. Subtle boundary inaccuracies, though minor in

This work 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) (RS-2024-00437718, RS-2023-00256629, RS-2022-00156287).

The authors are with Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju 61186, South Korea.
{jhyoon@gmail.com, cyw2628@jnu.ac.kr, sbyoo@jnu.ac.kr}
† These authors contributed equally to this work.
* Corresponding author: Seok Bong Yoo

2D, can be significantly amplified in monocular 3D detection where depth and orientation depend on precise spatial cues.

Motivated by these insights, we introduce MonoEM, a monocular 3D object detection framework designed to be robust under various weather conditions. MonoEM preserves the baseline 3D detector and instead restores geometric fidelity within detected object regions. By refining object-level representations before 3D regression, MonoEM mitigates error amplification caused by imperfect 2D localization. From the detected 2D regions, object-level images are converted into equirectangular maps, where weather-induced noise is subsequently mitigated. Given that such maps often suffer from resolution loss at long ranges, we incorporate an upsampling module to refine the structural details of distant objects. To accurately infer 3D bounding boxes, we further design a visual-range fusion detector that aligns and integrates features from the equirectangular domain and the image space. This module is trained with a 2D-3D box alignment loss to improve localization accuracy under monocular constraints. Our contributions are summarized below:

- We propose an equirectangular object-level reconstruction method comprising an equirectangular map translator to generate an equirectangular map from an image, range-aware dynamic denoising to remove weather noise, and an equirectangular map upsampler to improve the structural details of distant objects.
- We propose a visual-range fusion detector that employs coordinate alignment to bridge the equirectangular map and image spaces.
- We propose a 2D-3D box alignment loss to enhance detection consistency between the estimated 2D and 3D bounding boxes.

II. Related Work

A. Monocular 3D Object Detection

Monocular 3D object detection approaches are generally grouped into two types: those relying solely on image data and those utilizing additional external cues. Methods relying solely on monocular images [9]–[18] typically utilize either explicit geometric constraints or deep learning architectures to obtain 3D bounding boxes. For instance, Zhang et al. [19] introduced a depth-informed transformer that extracts spatial structure solely from monocular input, without requiring additional sensor data.

Given the limited 3D information inherent in monocular images, several studies [20]–[23] have explored incorporating supplementary cues to improve detection accuracy. Huang et al. [23], for example, proposed MonoDTR, which fuses depth features with contextual information. While earlier methods were primarily designed for clear weather scenarios, recent work has begun to address the challenges posed by adverse environments. Lei et al. [24] and Li et al. [25] proposed weather-specific data augmentation techniques; however, their models lack explicit mechanisms for resolving uncertainties in regions occluded by weather-induced noise, an area where restoration-based methods offer greater potential. More recently, Oh et al. [26] presented MonoWAD, a diffusion-based framework that leverages priors from clear-weather data to improve detection under foggy conditions. Although it delivers improved performance, the model suffers from slow inference speeds. To address these shortcomings, we propose an object-level monocular 3D detection framework designed to ensure reliable performance under a wide range of inclement weather conditions.

B. Image Restoration for Inclement Weather

Earlier works have aimed to recover weather-degraded images to their clean counterparts, facilitating downstream visual tasks. Several prior works [27]–[36] have focused on reconstructing images synthesized through synthesis techniques. Restormer, proposed by Zamir et al. [37], is a transformer architecture tailored to exploit multi-scale feature representations for improved image restoration performance. Valanarasu et al. [38] likewise introduced TransWeather, a unified transformer-based framework for end-to-end restoration under inclement weather.

III. Methodology

A. Overall Architecture

As visualized in Fig. 2, we introduce MonoEM, a monocular 3D object detection framework tailored for inclement weather scenarios. From a single input image, MonoEM extracts object-level regions using a 2D detector and processes them through two parallel branches: equirectangular object-level reconstruction and image restoration. In the first branch, the object image $\hat{I}_o$ is translated into an equirectangular map $\hat{E}_o$ via an equirectangular map translator. To mitigate weather-induced noise, range-aware dynamic denoising is applied, followed by an equirectangular map upsampler that refines structural details yielding the enhanced map $\hat{E}_{gen}$. In parallel, the second branch restores $\hat{I}_o$ using a CNN-transformer network. Finally, the restored image $\hat{I}_{res}$ and $\hat{E}_{gen}$ are fused via a visual-range fusion detector to predict the final 3D bounding box $\hat{B}_{3D}$.

B. Object Region Extraction

The existing SOTA detection approach [26], which incorporates image restoration modules, processes the entire image affected by inclement weather. However, this global processing is computationally intensive, posing limitations for real-time deployment. To overcome this issue, we introduce an object-level framework that uses 2D detection to localize and process only relevant regions.

As depicted in Fig. 2, our method employs CenterNet [7] with a DLA-34 architecture [39] for the 2D detection. This convolutional architecture offers a strong efficiency-accuracy balance, thus enabling real-time deployment. From a single input image, the detector predicts 2D bounding boxes $\hat{B}_{2D}$, which are then expanded to $2\times$ the original width (W) and height (H) to include contextual background information. The cropped object $\hat{I}_o$ is subsequently processed through two parallel branches: equirectangular object-level reconstruction and image restoration.

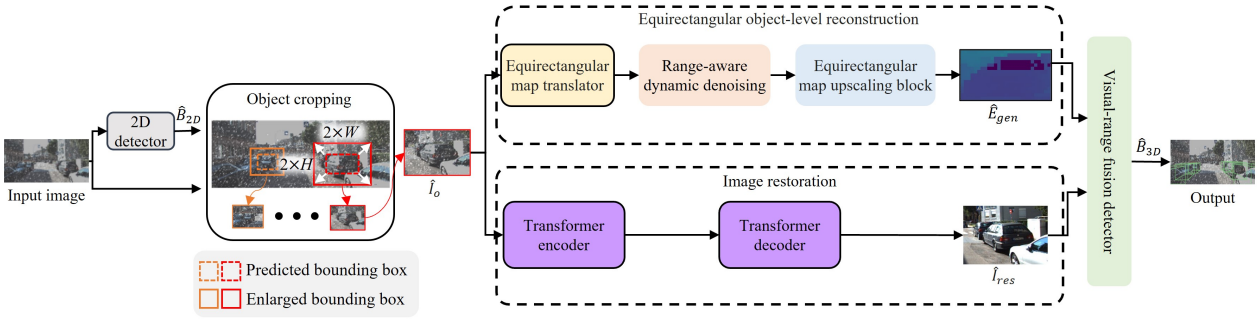

Fig. 2: Overview of the MonoEM framework.

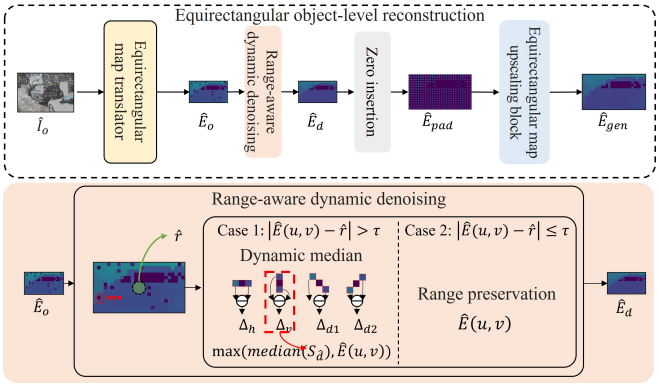

Fig. 3: Equirectangular object-level reconstruction.

C. Equirectangular Object-level Reconstruction

Monocular 3D detectors typically rely on depth maps or pseudo-LiDAR for spatial cues. However, depth maps are defined in the perspective image plane, where weather-induced artifacts are entangled with projection distortions, making noise suppression challenging. In contrast, equirectangular maps represent range in an angular domain ($W \times H$ as $\phi \times \theta$), where weather artifacts appear as angularly inconsistent range spikes. Our denoising suppresses these inconsistencies while preserving object-consistent geometric gradients, leading to more stable 3D regression under adverse conditions. Hence, we adopt equirectangular maps as an alternative 3D representation that captures weather-specific noise while reducing complexity.

Figure 3 illustrates the proposed equirectangular object-level reconstruction framework, which extracts clean and informative 3D cues from the object image $\hat{I}_o$. The $\hat{I}_o$ is converted into its corresponding equirectangular representation $\hat{E}_o$ using a translation network based on image-to-image translation [40]. To enhance edge fidelity and suppress irregular noise artifacts, we incorporate an additional gradient consistency loss $\mathcal{L}_{grad}$, defined as:

$$\mathcal{L}_{grad} = \sum_{d\in\{h,v\}} ||\nabla_d(\hat{E}_o) - \nabla_d(E_o)||_1 \quad (1)$$

where $\hat{E}_o$ denotes the translated output, and E_o is the ground-truth (GT) equirectangular map. The operators ∇_h and ∇_v compute the horizontal and vertical gradients.

The estimated equirectangular map $\hat{E}_o$ derived from $\hat{I}_o$ exhibits weather-induced pepper noise artifacts, as presented in Fig. 1(b). To suppress these artifacts, we apply range-aware dynamic denoising, depicted in Fig. 3. Assuming the object lies at the center, we estimate its center range as follows:

$$\hat{r} = median\{\hat{E}_o(u_c+i, v_c+j) | i,j \in \{-1,0,1\}\}. \quad (2)$$

Here, (u_c, v_c) indicates the center of $\hat{E}_o$. This estimated distance $\hat{r}$ is then used to distinguish object pixels from noisy surroundings. Pixels with values close to $\hat{r}$ are retained, while others are smoothed to yield the denoised map $\hat{E}_d$ below:

$$\hat{E}_d(u,v) = \begin{cases} \mathcal{D}(\hat{E}_o(u,v)), & |\hat{E}_o(u,v)-\hat{r}| > \tau \\ \hat{E}_o(u,v), & \text{otherwise} \end{cases}. \quad (3)$$

Here, τ is a threshold that defines the acceptable deviation from $\hat{r}$ for preserving foreground pixels. When the difference exceeds τ, we apply a dynamic median $\mathcal{D}$ defined as:

$$\mathcal{D}(\hat{E}_o(u,v)) = \max(median(\mathcal{S}_{\hat{d}}), \hat{E}_o(u,v)). \quad (4)$$

Here, $\mathcal{S}^{\hat{\theta}}$ denotes the selected directional set as follows:

$$\mathcal{S}_{\hat{d}} = \{\hat{E}_o(u-\Delta^i_{\hat{d}}, v-\Delta^j_{\hat{d}}), \hat{E}_o(u,v), \hat{E}_o(u+\Delta^i_{\hat{d}}, v+\Delta^j_{\hat{d}})\}, \quad (5)$$

where the optimal direction $\hat{d}$ is determined by minimizing the absolute difference as follows:

$$\hat{d} = \underset{d\in\{h,v,d_1,d_2\}}{\operatorname{argmin}} \Delta_d, \quad (6)$$

$$\Delta_d = |\hat{E}_o(u+\Delta^i_d, v+\Delta^j_d) - \hat{E}_o(u-\Delta^i_d, v-\Delta^j_d)|. \quad (7)$$

Here, the directions are defined as $\Delta^{i,j}_h = (1,0)$, $\Delta^{i,j}_v = (0,1)$, $\Delta^{i,j}_{d_1} = (1,1)$, and $\Delta^{i,j}_{d_2} = (-1,1)$. Consequently, the range-aware dynamic denoising mitigates weather-induced noise in the equirectangular map and preserves object pixels.

Although the proposed denoising techniques are effective in suppressing noise within the equirectangular map, they fall short in preserving structural details of distant objects. To address this, we utilize the estimated object range $\hat{r}$ to adaptively upsample the equirectangular map to an optimal resolution. Figure 4 presents the correlation between 3D average precision on the KITTI dataset [41] and the number of sampled points, segmented by object range in 10-meter intervals. Based on the results with our base MonoDETR [19],

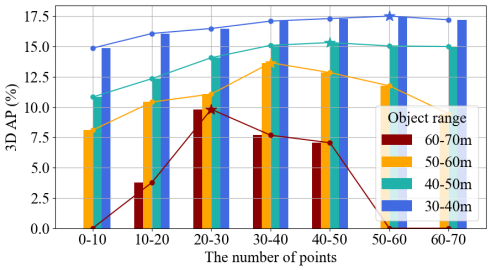

Fig. 4: The optimal number of points across object range.

the optimal number of points for each range (denoted by $\star$) is derived from $\hat{r}$, which subsequently guides the adaptive upsampling process. For upsampling, we employ a U-Net-based inpainting network inspired by [6]. The initial step involves inserting zero-valued pixels between the original values of $\hat{E}_o$, resulting in a zero-padded map $\hat{E}_{pad}$. A binary mask M is then used to inpaint the missing values:

$$\hat{E}_{gen} = \Phi(\hat{E}_{pad}) \odot M + \hat{E}_{pad} \odot (1 - M), \tag{8}$$

where Φ denotes the upsampling network and $\odot$ is element-wise multiplication. In this formulation, original pixels are preserved, and the newly inserted regions are produced.

To train the equirectangular map upsampler, we adopt the loss function as follows:

$$\mathcal{L}_{range} = \mathcal{L}_{L1} + \mathcal{L}_{perc} + \mathcal{L}_{grad}, \tag{9}$$

where $\mathcal{L}_{L1}$ minimizes the pixel-wise differences, and $\mathcal{L}_{perc}$ denotes the perceptual loss [42], which measures the discrepancy in VGG16 feature space. In addition, $\mathcal{L}_{grad}$ is computed as in Eq. (1) using the upsampled output $\hat{E}_{gen}$ and the GT equirectangular map E_{gt}. For training, we feed the network with equirectangular maps that are uniformly subsampled.

D. Image Restoration

The image restoration module operates in parallel with the equirectangular object-level reconstruction. It takes the $\hat{I}_o$ as input and passes it through a U-Net-shaped transformer architecture inspired by [37]. The encoder consists of a hybrid design combining convolutional layers and transformer blocks to effectively capture local textures and global contexts. The decoder progressively reconstructs the object image, producing the restored output $\hat{I}_{res}$, which is later fused with range-based information in the detection stage.

E. Visual-range Fusion Detector

Although our reconstruction methods yield clean data, fusing the equirectangular map and image remains challenging due to spatial misalignment. The equirectangular map is defined in polar coordinates (ϕ, θ), while the image uses Cartesian coordinates (u, v), hindering effective fusion and reducing detection accuracy. To address this, we propose a visual-range fusion detector that unifies feature representations via coordinate alignment and improves 3D region detection by reinforcing consistency with 2D detection results.

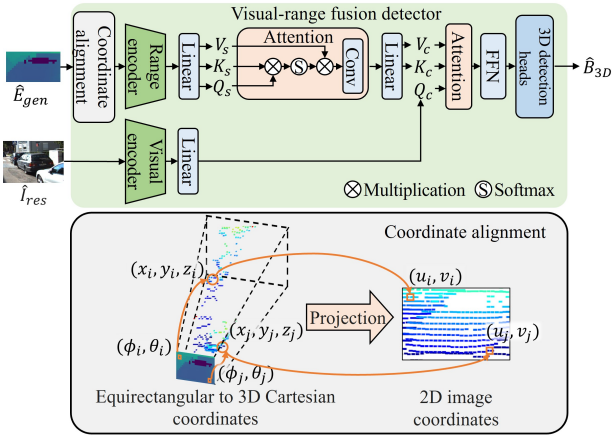

Fig. 5: Visual-range fusion detector.

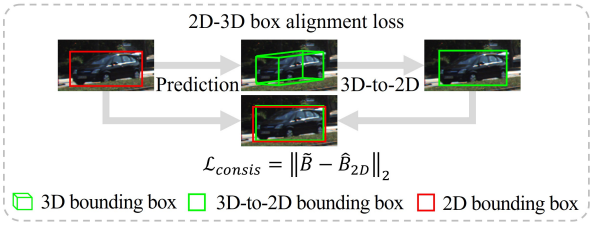

Fig. 6: 2D-3D box alignment loss.

In Fig. 5, the visual-range fusion detector takes E_{gen} and $\hat{I}_{res}$ as inputs, where $\hat{I}_{res}$ includes $\hat{B}_{2D}$. To ensure spatial alignment, E_{gen} undergoes coordinate alignment to correspond with $\hat{I}_{res}$. The equirectangular map coordinates (ϕ, θ) are first converted to 3D Cartesian coordinates (x, y, z) and then projected onto 2D image coordinates (u, v) using the camera projection pipeline as follows:

$$u = u'/s, \quad v = v'/s, \tag{10}$$

$$(u', v', s) = P \cdot R \cdot T \cdot (x, y, z, 1). \tag{11}$$

Here, P is the 3×4 projection matrix, R the rectification matrix, and T the transformation matrix (both 4×4). The s denotes the scaling factor. This step ensures spatial correspondence by mapping 3D range points to the image plane. Next, $\hat{E}_{gen}$ and $\hat{I}_{res}$ are encoded using range and visual encoders, respectively, following the architecture of [19]. The range encoder output is linearly projected to Q_s, K_s, and V_s for self-attention, enhancing spatially detailed 3D cues. The visual encoder output serves as Q_c in the cross-attention, integrating 3D cues with texture-rich visual features. The fused representation is passed through a feed-forward network (FFN) and a 3D detection head to predict $\hat{B}_{3D}$.

F. 2D-3D Box Alignment Loss

As shown in Fig. 6, we propose a 2D-3D box alignment loss to enhance spatial alignment between the projected 2D bounding box $\tilde{B}$ derived from $\hat{B}_{3D}$ and the $\hat{B}_{2D}$ from the 2D detector. To obtain $\tilde{B}$, the 3D box $\hat{B}_{3D}$ is first converted into its eight 3D corner points. These corner points are projected onto the 2D image plane using Eqs. (10) and (11), yielding a

set of image coordinates $(u_k, v_k)_{k=1}^{8}$. The 2D bounding box enclosing the projected corners is then defined as follows:

$$\tilde{u} = \frac{\max(u_k)+\min(u_k)}{2}, \tilde{v} = \frac{\max(v_k)+\min(v_k)}{2}, \\ \tilde{w} = \max(u_k) - \min(u_k), \tilde{h} = \max(v_k) - \min(v_k). \quad (12)$$

where $\tilde{u}, \tilde{v}$ specify the box center and $\tilde{w}, \tilde{h}$ specify the width and height of the box. To ensure consistency and suppress outlier projections that extend beyond the expected 2D region, we define the alignment loss as:

$$\mathcal{L}_{align} = |\tilde{B} - \hat{B}_{2D}|_2. \quad (13)$$

In addition to minimizing the discrepancy between predicted and GT 3D bounding boxes via the regression loss $\mathcal{L}_{reg}$, the cross-entropy loss $\mathcal{L}_{cls}$ enhances object classification. The overall objective is defined as:

$$\mathcal{L}_{tot} = \mathcal{L}_{reg} + \mathcal{L}_{cls} + \lambda_{align}\mathcal{L}_{align}, \quad (14)$$

where λ_{align} controls the influence of $\mathcal{L}_{align}$. Using this combination, MonoEM improves performance while reducing ambiguity via alignment with 2D detection outputs.

IV. Experiments

A. Dataset

The KITTI dataset [41] provides 7,481 images, with 3,712 allocated for training and the remaining 3,769 used for validation purposes. Model effectiveness was evaluated using average precision metrics: AP_{BEV} for bird's-eye view detection and AP_{3D} for 3D detection. The evaluation followed over 40 recall thresholds, with performance reported for the car class using a threshold of 0.7 for intersection-over-union (IoU). In line with [43], we created synthetic weather-affected images, collectively termed S-KITTI.

The Dense dataset [44] targets 3D detection across diverse weather, such as clear, snowy, rainy, and foggy. A total of 808 clear-weather, 947 snowy, 57 rainy, and 388 foggy validation images were employed. These were reformatted to the KITTI style and evaluated at an IoU threshold of 0.5.

The CADC dataset [45], focused on snowy weather, includes 1,238 validation images, evaluated at an IoU of 0.5.

B. Implementation Details

In the range-aware dynamic denoising module, the τ was set to 4.0, while the total loss function incorporated a weighting factor λ_{align} of 1. MonoEM was trained for 250 epochs, configured with a learning rate initialized to 2e-4 and a batch size of 4 on an RTX-3080 GPU. The AdamW optimizer with weight decay was employed, and the learning rate was halved at epochs 85, 125, 165, and 225 to aid convergence.

For supervision during training, we apply synthetic weather degradations to KITTI images (S-KITTI) and use the corresponding original clear KITTI images as reconstruction targets. Ground-truth equirectangular maps are generated from KITTI LiDAR point clouds for supervision during training. The Equirectangular map translator network is trained to regress these range maps from RGB inputs. LiDAR is used only for supervision during training and is not required at inference.

TABLE I: Assessment of monocular 3D detectors on S-KITTI across weather types for car class at moderate level.

Metric	$AP_{BEV}(\%)$				$AP_{3D}(\%)$			
Weather	Clear	Rain	Snow	Fog	Clear	Rain	Snow	Fog
DDMP-3D	18.22	2.29	0.24	12.14	12.56	0.94	0.42	6.19
PL:F-Pointnet	21.56	4.95	1.72	13.32	14.92	2.18	0.76	8.46
MonoDTR	22.31	5.43	1.86	13.92	15.22	2.80	0.95	8.94
DEVIANT	20.54	2.89	0.34	13.00	14.77	1.02	0.08	7.63
MonoDETR	22.84	6.69	2.67	15.83	16.27	3.69	1.42	10.03
MonoCD	21.06	6.12	2.57	14.66	15.37	2.77	1.25	8.96
MonoWAD	24.48	11.93	5.83	17.20	17.96	7.45	3.55	13.57
Ours	**24.89**	**12.01**	**6.46**	**18.94**	**19.03**	**7.95**	**4.17**	**14.77**

TABLE II: Assessment of monocular 3D object detectors on real-weather datasets (Dense and CADC) across weather types for car class at moderate level.

Metric	$AP_{BEV}(\%)$					$AP_{3D}(\%)$				
Dataset	Dense				CADC	Dense				CADC
Weather	Clear	Rain	Snow	Fog	Snow	Clear	Rain	Snow	Fog	Snow
DDMP-3D	13.91	2.63	0.42	6.17	9.39	9.88	0.13	0.08	3.02	8.72
PL:F-Pointnet	15.23	3.86	0.91	8.11	13.13	12.81	1.22	0.11	6.05	9.21
MonoDTR	15.62	4.01	1.77	10.50	13.66	13.23	1.71	0.18	7.42	11.05
DEVIANT	14.83	1.63	0.27	7.95	13.68	11.56	0.83	0.07	5.22	9.81
MonoDETR	15.88	6.00	1.76	10.23	13.17	13.80	2.56	0.22	6.11	10.74
MonoCD	15.18	3.72	1.91	7.80	10.90	12.79	1.25	0.27	5.62	9.38
MonoWAD	17.76	5.71	3.36	11.03	14.24	14.96	5.97	1.22	8.60	11.73
Ours	**20.49**	**9.21**	**4.41**	**12.90**	**15.09**	**15.82**	**6.37**	**2.14**	**9.88**	**12.25**

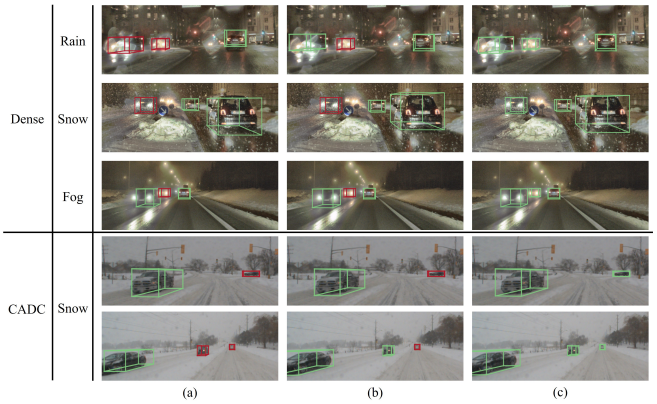

Fig. 7: Visual evaluation of the 3D object detectors, (a) MonoDETR, (b) MonoWAD, and (c) MonoEM, on Dense and CADC datasets. Correct predictions are highlighted in green, while missed ones are shown in red.

C. Main Results

We evaluate MonoEM against SOTA monocular 3D detection and restoration models trained on the S-KITTI dataset under all weather conditions. Bold text indicates the best performance in each table.

Table I presents a comparison between existing monocular 3D object detectors and MonoEM on the S-KITTI dataset for the car class at moderate level. Existing methods show notable performance drops under inclement weather, while our method consistently outperforms them in AP_{BEV} and AP_{3D}, including in clear weather scenarios. This robustness stems from the proposed range-aware dynamic denoising for harsh conditions, and improvements under clear weather are attributed to equirectangular map upsampling and 2D-3D box alignment. Additionally, our equirectangular map-based approach outperforms alternatives relying on pseudo-LiDAR (PL:F-PointNet) or depth maps (DDPM-3D).

Table II extends the comparison to the Dense and CADC

TABLE III: Assessment of 3D object detectors with restoration models on S-KITTI across weather types for car class at moderate level.

Metric		$AP_{3D}(\%)$			
Restoration	Detection	Clear	Rain	Snow	Fog
Restormer	MonoDETR	17.18	5.27	2.09	12.66
	MonoCD	16.89	4.60	1.82	12.25
FocalNet	MonoDETR	18.32	6.18	2.67	13.81
	MonoCD	17.61	5.86	2.30	13.38
Ours		**19.03**	**7.95**	**4.17**	**14.77**

TABLE IV: Assessment of denoising applied to equirectangular maps on the S-KITTI.

Metric	PSNR (dB)	SSIM
w/o filtering	21.16	0.785
Median filtering (3×3)	22.85	0.779
Range-aware dynamic denoising	**24.22**	**0.803**

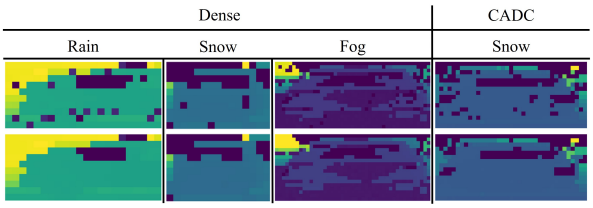

Fig. 8: Visual evaluation of the proposed range-aware dynamic denoising.

datasets, which reflect real-world weather scenarios. All models were trained on S-KITTI and tested on these datasets to assess generalization. MonoEM outperforms existing SOTA methods across all weather conditions, demonstrating generalizability. This result stems from explicitly modeling weather-induced noise patterns in the equirectangular map, which are consistently observed across diverse datasets.

Figure 7 shows visual comparisons between MonoEM and existing detectors on the Dense and CADC datasets. Existing models show limited performance in inclement weather. In contrast, MonoEM uses equirectangular map noise characteristics to guide object-level restoration, leading to more accurate detections. Although some failure cases remain for distant and heavily occluded objects, MonoEM demonstrates greater robustness overall.

Table III compares SOTA monocular 3D object detectors combined with image restoration approaches, including Restormer [37] and FocalNet [35], evaluated on the S-KITTI dataset in inclement weather. The detectors were trained using images enhanced by these restoration models. In contrast to the combined methods, MonoEM achieves the best detection performance across diverse conditions.

Table IV reports the denoising performance of the proposed range-aware dynamic denoising on S-KITTI, measured by PSNR and SSIM. Compared to 3×3 median filtering, our method yields higher scores, demonstrating superior effectiveness in mitigating noise in equirectangular maps.

Table V presents the PSNR and SSIM for the image-to-image translation performed on equirectangular maps on the S-KITTI dataset. Our method is compared against the

TABLE V: Assessment of the translator applied to equirectangular maps on the S-KITTI.

Metric	PSNR (dB)	SSIM
StegoGAN	28.49	0.798
Ours	**35.73**	**0.830**

TABLE VI: Complexity of monocular 3D object detectors on the S-KITTI.

Metric	Params (M)	Time (ms)	FLOPs (G)
DDMP-3D	285.50	182.27	158.03
PL:F-Pointnet	90.72	335.26	134.81
MonoDTR	54.27	62.81	117.96
DEVIANT	16.63	210.33	117.52
MonoDETR	35.93	38.87	119.44
MonoCD	41.96	42.36	171.20
MonoWAD	74.34	853.78	181.06
Ours	32.86	37.51	117.91

TABLE VII: Ablation study for the MonoEM on the S-KITTI at moderate level.

2D detector	Range-aware dynamic denoising	Equirectangular map upsampler	Image restoration	$AP_{3D}(\%)$
✓	✓	✓	✓	**11.48**
✓	✓	✓		10.34
✓	✓		✓	10.98
✓		✓	✓	9.69
	✓	✓	✓	10.74
				8.56

baseline model, StegoGAN [40]. StegoGAN, when trained on equirectangular maps, tends to suffer from structural distortion and introduces artifacts caused by abnormal pixels. The gradient loss term $\mathcal{L}_G$ we introduced enhances structural fidelity and suppresses artifacts. Consequently, our approach yields higher PSNR and SSIM values, indicating better handling of the inherent property of equirectangular maps.

Figure 8 presents the visual outcomes of applying range-aware dynamic denoising to equirectangular maps from real-world weather datasets (Dense and CADC). The results validate that the proposed method effectively suppresses weather-related noise while preserving the structural details, highlighting its robustness under various weather conditions.

Table VI presents a comparison of model complexity in terms of Params, inference time, and FLOPs on the S-KITTI dataset. Despite integrating restoration modules, the proposed method achieves the lowest inference time, underscoring its suitability for real-time applications. Additionally, its Params and FLOPs are on par with existing detection models. This efficiency is attributed to the object-level design of the reconstruction and detection components.

D. Ablation Study

Table VII exhibits the ablation study, analyzing the detection performance of the MonoEM across all weather conditions on the S-KITTI dataset. The symbol ✓ signifies that the module is active. The first row shows the AP_{3D} score of the complete model with all components enabled. The second to fourth rows present ablation results by individually removing the image restoration module, equirectangular map upsampler, and range-aware dynamic denoising, respectively. The fifth row assesses the impact of removing the object-

level pipeline, using the full image instead of 2D detector-guided regions. Finally, the last row reports the baseline performance of the backbone network. These results indicate that the proposed components contribute positively to the overall accuracy.

V. Conclusion

This work aims to mitigate the vulnerability of monocular 3D object detectors to a variety of weather conditions. MonoEM initiates the pipeline by employing a 2D object detector to generate region proposals, which simplifies subsequent 3D object detection and reduces both computational cost and ambiguity. To recover spatial geometry, the model leverages an estimated equirectangular map that encapsulates scene-level depth characteristics. Within the equirectangular map, MonoEM dynamically suppresses noise artifacts and adaptively enhances resolution in distant regions to retain structural fidelity. A visual-range fusion module further refines 3D box estimation by enforcing spatial coherence between RGB features and range cues, guided by a 2D-3D box alignment loss on the estimated bounding boxes. As a result, MonoEM achieves real-time speed while delivering superior detection accuracy across both synthetic and real-world benchmarks in adverse environmental settings.

References

[1] T. Dam, S. B. Dharavath, S. Alam, N. Lilith, S. Chakraborty, and M. Feroskhan, "Aydiv: Adaptable yielding 3d object detection via integrated contextual vision transformer," in *ICRA*. IEEE, 2024, pp. 10 657–10 664.

[2] J. Fu, C. Gao, Z. Wang, L. Yang, X. Wang, B. Mu, and S. Liu, "Eliminating cross-modal conflicts in bev space for lidar-camera 3d object detection," in *ICRA*. IEEE, 2024, pp. 16 381–16 387.

[3] A. C. Stutts, D. Erricolo, S. Ravi, T. Tulabandhula, and A. R. Trivedi, "Mutual information-calibrated conformal feature fusion for uncertainty-aware multimodal 3d object detection at the edge," in *ICRA*. IEEE, 2024, pp. 2029–2035.

[4] J. H. Yoon, J. W. Jung, E.-G. Lee, and S. B. Yoo, "Oprnet: Object-centric point reconstruction network for multimodal 3d object detection in adverse weathers," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 3954–3961.

[5] J. H. Yoon, J. W. Jung, and S. B. Yoo, "Equirectangular point reconstruction for domain adaptive multimodal 3d object detection in adverse weather conditions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, 2025, pp. 9553–9561.

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

[7] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[8] I. Lee, E. Lee, and S. B. Yoo, "Latent-ofer: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1536–1546.

[9] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *ICCV*, 2019, pp. 9287–9296.

[10] X. Liu, N. Xue, and T. Wu, "Learning auxiliary monocular contexts helps monocular 3d object detection," in *AAAI*, vol. 36, no. 2, 2022, pp. 1810–1818.

[11] Z. Liu, Z. Wu, and R. Tóth, "Smoke: Single-stage monocular 3d object detection via keypoint estimation," in *CVPRW*, 2020, pp. 996–997.

[12] S. Luo, H. Dai, L. Shao, and Y. Ding, "M3dssd: Monocular 3d single stage object detector," in *CVPR*, 2021, pp. 6145–6154.

[13] X. Shi, Q. Ye, X. Chen, C. Chen, Z. Chen, and T.-K. Kim, "Geometry-based distance decomposition for monocular 3d object detection," in *ICCV*, 2021, pp. 15 172–15 181.

[14] Y. Zhou, Y. He, H. Zhu, C. Wang, H. Li, and Q. Jiang, "Monocular 3d object detection: An extrinsic parameter free approach," in *CVPR*, 2021, pp. 7556–7566.

[15] X. Jiang, S. Jin, X. Zhang, L. Shao, and S. Lu, "Monomae: Enhancing monocular 3d detection through depth-aware masked autoencoders," in *NeurIPS*, vol. 37, 2024, pp. 11 392–11 411.

[16] Q. Yang, H. Chen, Z. Chen, and J. Su, "Uncertainty estimation for monocular 3d object detectors in autonomous driving," in *ICRAE*. IEEE, 2021, pp. 55–59.

[17] D. Park, J. Li, D. Chen, V. Guizilini, and A. Gaidon, "Depth is all you need for monocular 3d detection," in *ICRA*. IEEE, 2023, pp. 7024–7031.

[18] Y. Liu, Z. Xu, and M. Liu, "Star-convolution for image-based 3d object detection," in *ICRA*. IEEE, 2022, pp. 5018–5024.

[19] R. Zhang, H. Qiu, T. Wang, Z. Guo, Z. Cui, Y. Qiao, H. Li, and P. Gao, "Monodetr: Depth-guided transformer for monocular 3d object detection," in *ICCV*, 2023, pp. 9155–9166.

[20] L. Peng, X. Wu, Z. Yang, H. Liu, and D. Cai, "Did-m3d: Decoupling instance depth for monocular 3d object detection," in *ECCV*. Springer, 2022, pp. 71–88.

[21] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *CVPR*, 2021, pp. 8555–8564.

[22] L. Wang, L. Du, X. Ye, Y. Fu, G. Guo, X. Xue, J. Feng, and L. Zhang, "Depth-conditioned dynamic message propagation for monocular 3d object detection," in *CVPR*, 2021, pp. 454–463.

[23] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, "Monodtr: Monocular 3d object detection with depth-aware transformer," in *CVPR*, 2022, pp. 4012–4021.

[24] Y. Lei, X. Li, Z. Jiang, X. Ju, and J. Liu, "Aeam3d: Adverse environment-adaptive monocular 3d object detection via feature extraction regularization," in *ICASSP*. IEEE, 2024, pp. 4135–4139.

[25] X. Li, J. Liu, Y. Lei, L. Ma, X. Fan, and R. Liu, "Monotdp: Twin depth perception for monocular 3d object detection in adverse scenes," *arXiv preprint arXiv:2305.10974*, 2023.

[26] Y. Oh, H.-I. Kim, S. T. Kim, and J. U. Kim, "Monowad: Weather-adaptive diffusion model for robust monocular 3d object detection," in *ECCV*. Springer, 2024, pp. 326–345.

[27] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *TPAMI*, vol. 45, no. 8, pp. 10 346–10 357, 2023.

[28] S. Kalwar, D. Patel, A. Aanegola, K. R. Konda, S. Garg, and K. M. Krishna, "Gdip: Gated differentiable image processing for object detection in adverse conditions," in *ICRA*. IEEE, 2023, pp. 7083–7089.

[29] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-in-one image restoration for unknown corruption," in *CVPR*, 2022, pp. 17 452–17 462.

[30] R. Li, R. T. Tan, and L.-F. Cheong, "All in one bad weather removal using architectural search," in *CVPR*, 2020, pp. 3175–3185.

[31] Y. Xia, J. Monica, W.-L. Chao, B. Hariharan, K. Q. Weinberger, and M. Campbell, "Image-to-image translation for autonomous driving from coarsely-aligned image pairs," in *ICRA*. IEEE, 2023, pp. 7756–7762.

[32] X. Guo, X. Fu, M. Zhou, Z. Huang, J. Peng, and Z.-J. Zha, "Exploring fourier prior for single image rain removal." in *IJCAI*, 2022, pp. 935–941.

[33] X. Wu, T. Huang, L. Deng, and T. Zhang, "A decoder-free transformer-like architecture for high-efficiency single image deraining," in *Proc. IJCAI*, 2022, p. 80.

[34] Y. Liang, B. Wang, W. Zuo, J. Liu, and W. Ren, "Self-supervised learning and adaptation for single image dehazing." in *IJCAI*, 2022, pp. 1137–1143.

[35] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Focal network for image restoration," in *ICCV*, 2023, pp. 13 001–13 011.

[36] H. Porav, T. Bruls, and P. Newman, "I can see clearly now: Image restoration via de-raining," in *ICRA*. IEEE, 2019, pp. 7087–7093.

[37] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *CVPR*, 2022, pp. 5728–5739.

[38] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel, "Transweather: Transformer-based restoration of images degraded by adverse weather conditions," in *CVPR*, 2022, pp. 2353–2363.

[39] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *CVPR*, 2018, pp. 2403–2412.

[40] S. Wu, Y. Chen, S. Mermet, L. Hurni, K. Schindler, N. Gonthier, and L. Landrieu, "Stegogan: Leveraging steganography for non-bijective image-to-image translation," in *CVPR*, 2024, pp. 7922–7931.
[41] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
[42] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*. Springer, 2016, pp. 694–711.
[43] Y. Dong, C. Kang, J. Zhang, Z. Zhu, Y. Wang, X. Yang, H. Su, X. Wei, and J. Zhu, "Benchmarking robustness of 3d object detection to common corruptions," in *CVPR*, 2023, pp. 1022–1032.
[44] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *CVPR*, 2020, pp. 11 682–11 692.
[45] M. Pitropov, D. E. Garcia, J. Rebello, M. Smart, C. Wang, K. Czarnecki, and S. Waslander, "Canadian adverse driving conditions dataset," *The International Journal of Robotics Research*, vol. 40, no. 4-5, pp. 681–690, 2021.