MonoWAD: Weather-Adaptive Diffusion Model for Robust Monocular 3D Object Detection

Youngmin Oh[1], Hyung-Il Kim[2], Seong Tae Kim[1†], and Jung Uk Kim[1†]

[1] Kyung Hee University, Yong-in, South Korea
{oym9104, st.kim, ju.kim}@khu.ac.kr
[2] ETRI, Daejeon, South Korea
hikim@etri.re.kr

Abstract. Monocular 3D object detection is an important challenging task in autonomous driving. Existing methods mainly focus on performing 3D detection in ideal weather conditions, characterized by scenarios with clear and optimal visibility. However, the challenge of autonomous driving requires the ability to handle changes in weather conditions, such as foggy weather, not just clear weather. We introduce MonoWAD, a novel weather-robust monocular 3D object detector with a weather-adaptive diffusion model. It contains two components: (1) the weather codebook to memorize the knowledge of the clear weather and generate a weather-reference feature for any input, and (2) the weather-adaptive diffusion model to enhance the feature representation of the input feature by incorporating a weather-reference feature. This serves an attention role in indicating how much improvement is needed for the input feature according to the weather conditions. To achieve this goal, we introduce a weather-adaptive enhancement loss to enhance the feature representation under both clear and foggy weather conditions. Extensive experiments under various weather conditions demonstrate that MonoWAD achieves weather-robust monocular 3D object detection. The code and dataset are released at https://github.com/VisualAIKHU/MonoWAD.

Keywords: Monocular 3D Object Detection · Weather-Adaptive Diffusion · Weather Codebook

1 Introduction

Monocular 3D object detection aims to detect 3D objects using only a single camera [16, 18, 28, 46, 55, 57]. In contrast to LiDAR-based methods that rely on expensive LiDAR sensors for depth estimation [24, 44, 45, 56], and stereo-based methods that require synchronized stereo cameras, monocular 3D object detection only requires monocular images, offering the advantage of computational cost-effectiveness and requiring fewer resources. Due to this characteristic, the monocular 3D object detection is applied to a wide range of real-world applications, such as autonomous driving [12, 19, 53] and robotics [50].

† Corresponding author

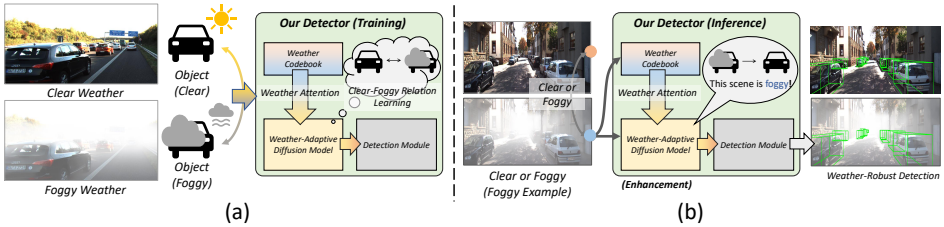

Fig. 1: Conceptual diagram of the proposed method (foggy example). (a) In the training phase, weather codebook learns the clear knowledge to transfer it to weather-adaptive diffusion model to enhance content related to the weather conditions. (b) By doing so, even with input images under various weather conditions (*e.g.,* foggy images), monocular 3D object detection becomes adaptable to various weather scenarios.

However, existing monocular 3D object detectors mainly focus on ideal autonomous driving environments (*i.e.,* clear weather). There are challenges in applying them to real-world scenarios with adverse weather conditions, such as fog and rain. Among these, fog poses the most significant challenge compared to other weather [14, 17]. This is due to the dense and diffuse nature of fog, which strongly scatters and absorbs light, leading to difficulties in object detection [2]. Since monocular 3D object detection relies solely on visual information from a single image, unlike LiDAR, it is crucial to design detectors to achieve enhanced performance in challenging visibility scenarios.

In this paper, we propose MonoWAD, a novel weather-robust monocular 3D object detector to address the aforementioned issues. As mentioned earlier, due to the inherent challenges posed by foggy weather among various adverse weather conditions, we focus primarily on clear and foggy weather (results for other weather conditions such as rainy and sunset are presented in Section 4). For weather-robust object detection, clear weather requires relatively modest improvements to enhance the visual representation of features. In contrast, significant enhancements in this feature representation are required for foggy weather. To address this, we consider two key aspects: (1) how to quantify the degree of improvement needed for the input image, and (2) how to guide the representation of the input image.

To address the two key aspects, as shown in Fig. 1, our MonoWAD consists of a weather codebook and weather-adaptive diffusion model. First, we introduce the weather codebook to generate a weather-reference feature that contains knowledge about reference weather in a given scene. The reference weather acts as a guide, indicating the degree of weather improvement required. Since the clear weather contains a richer visual representation of objects, we adopt it as a reference weather. At this time, we devise a clear knowledge recalling (CKR) loss to guide the weather codebook to memorize information about clear weather and generate a weather-reference feature for any input (clear or foggy). As a result, our detector can understand where improvements are needed in the input features based on the weather-reference feature.

Second, we propose a weather-adaptive diffusion model to effectively enhance feature representations in accordance with weather conditions. Given input fea-

ture (clear or foggy), the weather-adaptive diffusion model dynamically enhances the representation of the input feature based on the weather-reference feature. The weather-reference feature plays a role of attention, determining the extent to which the input features need improvement. At this time, we define the difference between clear and foggy weather (*i.e.,* weather changes) as the fog distribution to adopt it as the noise for our diffusion model. With fog distribution, our weather-adaptive diffusion model can enhance the feature representation according to the weather conditions through multiple steps of reverse processes. To achieve this goal, we introduce a weather-adaptive enhancement (WAE) loss. As a result, our MonoWAD performs weather-robust detection by adaptively improving feature representation according to the weather conditions.

To adaptively enhance the feature representation through the difference between weather conditions, we generate a new foggy KITTI dataset based on the KITTI dataset [12]. Comprehensive experimental results on several datasets [11, 12] show that our MonoWAD outperforms the existing state-of-the-art monocular 3D object detectors [16, 28, 37, 40, 55] under foggy weather, which is the most challenging weather condition. While our method primarily focuses on foggy, experiments conducted under various weather conditions (*e.g.,* foggy, rainy, and sunset) have demonstrated its applicability to other weather scenarios.

The main contributions of our paper can be summarized as follows:

- We introduce a new weather-robust monocular 3D object detector, called MonoWAD, that is robust to various weather conditions.
- We design a weather codebook with clear knowledge recalling loss for learning about clear weather, providing reference information for enhancement.
- We propose weather-adaptive diffusion model with weather-adaptive enhancement loss to dynamically enhance the feature representation of the input images according to the weather conditions.

2 Related Work

2.1 Monocular 3D Object Detection

Monocular 3D object detection task can be categorized into two directions according to the type of data used in the training phase: (1) using only a monocular image and (2) incorporating additional data, such as depth along with a monocular image. The first category relies on the geometric relationship between 2D and 3D [3, 7, 25, 26, 28, 31, 34, 40, 46, 55, 57]. For example, Deep3Dbox [34] utilizes the geometric information of 2D bounding boxes to predict 3D bounding boxes. In [3], M3D-RPN was proposed to understand the 3D scene from the depth-aware convolution. MonoRCNN [46] predicted 3D bounding boxes through geometry-based distance decomposition, and MonoCon [25] learns mono context for 3D object detection. MonoDETR [55] introduces a depth-guided transformer that utilizes geometric depth cues without requiring additional data.

Moreover, since monocular image contains limited information for estimating 3D object cues, monocular 3D object detectors have adopted the additional

data for more robust detection [4,8,16,27,32,37,41,48]. For example, the depth-conditioned dynamic message propagation (DDMP) [48] was proposed to integrate prior depth information with the image context. CaDDN [41] was introduced to utilize the depth distribution of predicted categories for each pixel to project the context information onto 3D space, deriving 3D bounding boxes. MonoDTR [16] employs the transformer architecture to integrate depth features and context, thus estimating more accurate depth information.

Despite recent progress, the existing monocular 3D object detectors mainly rely on the benchmark data collected under clear weather conditions. However, it is essential to account for challenging adverse weather conditions, such as fog, to more accurately reflect real-world scenarios. In this paper, we aim to introduce weather-robust 3D object detection by enhancing visual features with the proposed weather-adaptive diffusion model and weather codebook.

2.2 Computer Vision Tasks for Foggy Weather

There have been a lot of studies on improving performance in various weather conditions for real-world application of computer vision technology [2, 13, 20, 21, 33, 38, 43, 51, 54]. In particular, fog is considered one of the most critical issues due to its significant degradation of visual information compared to other weather conditions. To deal with the foggy weather conditions, the authors in [43] generated synthetic fog images (so-called Foggy Cityscapes) from clear weather images, and utilized them for training semantic segmentation and 2D object detection. Similar to Foggy Cityscapes, Martin *et al.* [13] naturally synthesizes fog into LIDAR to enhance the performance of LiDAR-based 3D object detectors in foggy weather. Bijelic *et al.* [2] use real dataset including foggy weather, for 2D detection with the multimodal fusion networks. Mai *et al.* [33] synthesize fog for LiDAR and stereo images and perform fusion-based 3D object detection using the SLS-Fusion network. Xin *et al.* [51] focused on 2D detection in foggy weather by applying domain adaptation. In this context, we focus on monocular 3D object detection in scenarios that rely solely on visual information in challenging foggy environments. To address this challenge, we propose a weather-robust diffusion model that dynamically improves features based on reference feature.

2.3 Diffusion Models

Recently, the diffusion model [15] has attracted considerable interest in computer vision due to its impressive progress in image generation [22, 29, 35, 42], as well as its potential application to other vision tasks such as segmentation [1] and image captioning [30]. Inspired by the remarkable generative ability, we design the diffusion model for robust monocular 3D object detection by considering a foggy effect (which is one of the challenging adverse weather conditions for monocular 3D object detection) as a form of noise in the model. That is, we propose a method in which visual features obscured by fog are progressively improved by training a diffusion model based on the forward/reverse diffusion process. In particular, we present an adaptive method that allows the diffusion model to control the degree of improvement by weather conditions.

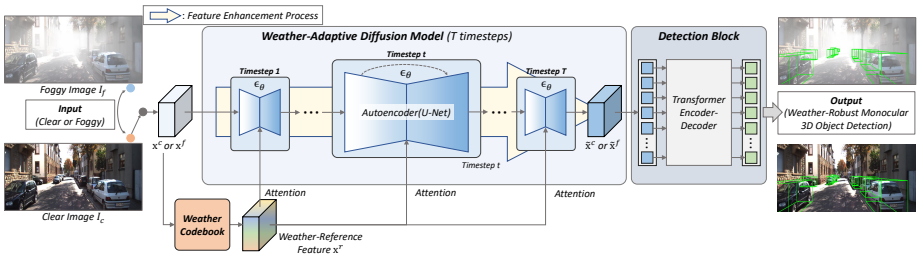

Fig. 2: Overview of our MonoWAD in the inference phase. It mainly contains three parts: weather codebook, weather-adaptive diffusion model, and detection block. Through the weather codebook and weather-adaptive diffusion model, our method can maintain robustness against various weather conditions (*i.e.,* clear or foggy).

3 Proposed Method

Fig. 2 shows the overall framework of the proposed MonoWAD in the inference phase. A backbone network receives an input image (clear image I_c or foggy image I_f) to encode the corresponding input feature (input clear feature x^c or input foggy feature x^f). It interacts with the weather codebook $\mathcal{Z}$ to generate weather-reference feature x^r, indicating the amount of enhancement for the given input feature. Subsequently, the weather-adaptive diffusion model attempts to enhance the input feature over T timesteps to obtain an enhanced feature $\tilde{x}^c$ or $\tilde{x}^f$. Finally, monocular 3D object detection is performed through the detection block. Note that, in the training phase, our MonoWAD use the clear feature x^c to train our diffusion as well as the detection block.

We address two key issues: (1) how to guide weather-reference feature x^r to serve as a reference feature, and (2) how to guide the weather-adaptive diffusion model to effectively enhance feature representations based on weather conditions. Details are in the following subsections.

3.1 Weather Codebook

In foggy weather conditions, the overall visual quality of the scene is generally poor, requiring a significant enhancement. Conversely, in clear weather, the amount of improvement is expected to be relatively minimal compared to foggy conditions. Therefore, inspired by [9,36], we devise a weather codebook $\mathcal{Z}$ to provide the reference knowledge about the weather for appropriate enhancement based on the weather conditions. At this time, as the clear weather contains abundant visual representations, we use it as a reference weather knowledge.

As shown in Fig. 3, the reference knowledge embedding procedure involves receiving paired clear-foggy features during the training phase. The weather codebook $\mathcal{Z}$ consists of K learnable slots, denoted as $\mathcal{Z} = \{z_k\}_{k=1}^{K} (z_k \in \mathbb{R}^{1\times c})$, where c represents the dimensionality of each slot. The paired clear feature x^c and foggy feature x^f pass through a convolution layer to generate $\hat{x}^c \in \mathbb{R}^{h\times w\times c}$ and $\hat{x}^f \in \mathbb{R}^{h\times w\times c}$ (w denotes width and h indicates height). Each element of the

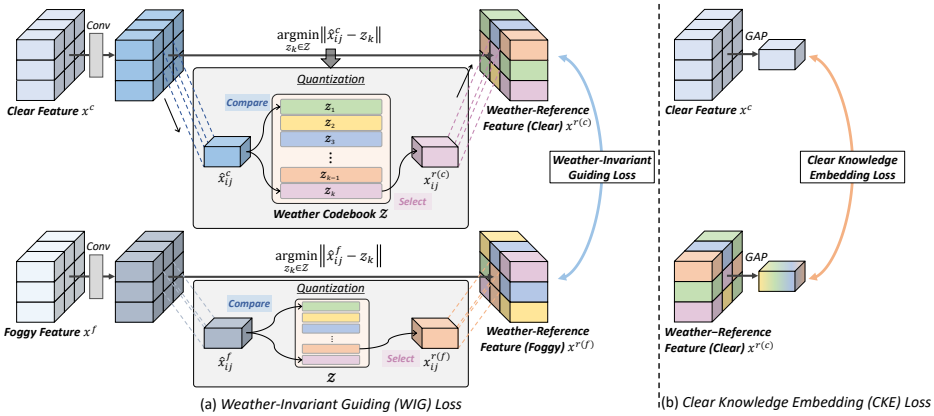

Fig. 3: Illustration of the proposed (a) weather-invariant guiding (WIG) loss and (b) clear knowledge embedding (CKE) loss. The clear knowledge recalling (CKR) loss, obtained from combining WIG and CKE, aims to memorize the knowledge of the clear weather and recall the same clear knowledge from the foggy weather.

feature denoted as $\hat{x}_{ij}^c \in \mathbb{R}^{1\times c}$ and $\hat{x}_{ij}^f \in \mathbb{R}^{1\times c}$. Subsequently, we obtain weather-reference feature for clear weather, $x^{r(c)} \in \mathbb{R}^{h\times w\times c}$, by conducting element-wise quantization process $\mathbf{q}(\cdot)$, calculated as:

$$x^{r(c)} = \mathbf{q}(\hat{x}^c) \coloneqq \left(\underset{z_k \in \mathcal{Z}}{\arg\min} \|\hat{x}_{ij}^c - z_k\| \right). \tag{1}$$

Utilizing x^c and $x^{r(c)}$, we introduce a clear knowledge embedding (CKE) loss $\mathcal{L}_{cke}$ to guide $x^{r(c)}$ to follow the representation of x^c. To this end, we perform global average pooling (GAP) for x^c and $x^{r(c)}$ with softmax, generating s^c and $s^{r(c)}$, respectively. Each element in the vector indicates the probability of the significance of each channel. With s^c and $s^{r(c)}$, we employ KL divergence $D_{KL}(\cdot)$ for $\mathcal{L}_{cke}$ to compare the probability distributions, formulated as:

$$\mathcal{L}_{cke} = D_{KL}(s^c || s^{r(c)}). \tag{2}$$

Through $\mathcal{L}_{cke}$, the weather codebook $\mathcal{Z}$ can memorize the knowledge of the clear weather, allowing it to effectively reconstruct the clear weather knowledge.

Additionally, as the paired clear-foggy images are identical except for weather conditions, the quantization process of the foggy feature x^f with the weather codebook should generate an equivalent weather-reference feature. For obtaining weather-reference feature for foggy $x^{r(f)}$, the element-wise quantization process is also conducted for $\hat{x}^f$ and $\mathcal{Z}$:

$$x^{r(f)} = \mathbf{q}(\hat{x}^f) \coloneqq \left(\underset{z_k \in \mathcal{Z}}{\arg\min} \|\hat{x}_{ij}^f - z_k\| \right). \tag{3}$$

Next, we introduce the weather-invariant guiding (WIG) loss $\mathcal{L}_{wig}$ to guide $\mathcal{Z}$ that the weather codebook recalls the same clear knowledge of clear feature for

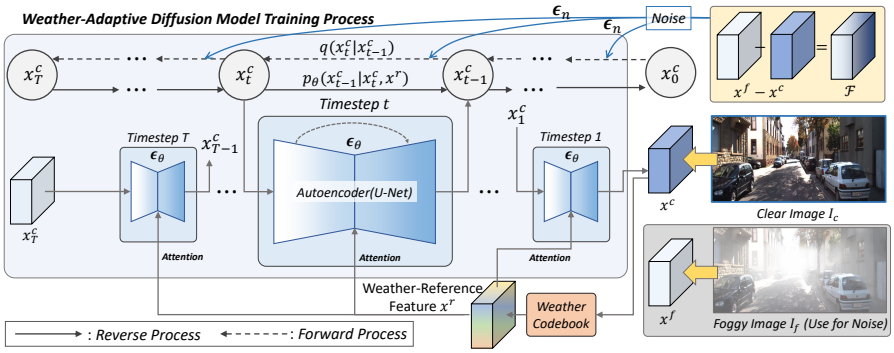

Fig. 4: Training process of the weather-adaptive diffusion model, which consists of two processes: (1) Adding fog variant $\boldsymbol{\epsilon}_n$ from input clear feature x^c (forward process) and (2) enhancing representation with weather-reference feature x^r (reverse process).

the foggy feature, which can be represented as follows:

$$\mathcal{L}_{wig} = \left\| x^{r(c)} - x^{r(f)} \right\|_2^2. \tag{4}$$

Finally, the clear knowledge recalling (CKR) loss $\mathcal{L}_{ckr}$ is obtained by adding $\mathcal{L}_{cke}$ and $\mathcal{L}_{wig}$, which is defined as:

$$\mathcal{L}_{ckr} = \mathcal{L}_{cke} + \mathcal{L}_{wig}. \tag{5}$$

In the training phase, the weight parameters of embedding K slots of weather codebook $\mathcal{Z}$ are initialized randomly and they are updated through Eq. (5). In the inference phase, all parameters are fixed to recall clear weather, generating weather-reference features for any weather conditions.

3.2 Weather-Adaptive Diffusion Model

Through Section 3.1, we now know that the weather codebook outputs the weather-reference feature $x^{r(c)}$ for the clear weather and $x^{r(f)}$ for the foggy weather through Eq. (5). From now on, since our method can receive any input images (clear or foggy), we denote the weather-reference feature as x^r.

Fig. 4 shows the training process of the proposed weather-adaptive diffusion model. The key idea of the diffusion model [15,42] is to gradually enhance x_T to x_0 with fixed Markov Chain of T timesteps. To this end, the forward and reverse processes are conducted in the training phase, and only the reverse process is used for the inference phase. Motivated by [15,42], we construct the weather-adaptive diffusion model to enhance representation related to the weather conditions. To this end, unlike traditional diffusion methods [15, 42] that adopt the Gaussian noise to the image or latent space, we adopt $\mathcal{F} = x^f - x^c$, called fog distribution, to guide our diffusion model to be aware of the foggy weather. Ideally, the information contained within $\mathcal{F}$ should include the information of fog, as it represents the difference between the foggy scene and the same scene with clear weather.

By doing so, our diffusion model learns the variation in weather by repeatedly adding and removing the fog. Note that, as we take clear feature x^c for x_0 to make a reference input for our diffusion model, we newly denote x_0 as x_0^c.

For the forward process at the t-th timestep, $q(x_t^c|x_{t-1}^c)$ takes the previous feature x_{t-1}^c and the noise related to the fog (*i.e.,* $\mathcal{F}$) as inputs to generate x_t^c. This procedure is repeated over T timesteps, which can be represented as:

$$q(x_t^c|x_{t-1}^c) = \mathcal{F}(x_t^c; \sqrt{1-\beta_t}x_{t-1}^c, \beta_t\mathbf{I}), \tag{6}$$

$$q(x_1^c, ..., x_T^c|x^c) = \prod_{t=1}^{T} q(x_t^c|x_{t-1}^c), \tag{7}$$

where β_t denotes the variance schedule.

Next, the reverse process at the t-th timestep aims to estimate fog variant $\boldsymbol{\epsilon}_n$ using x_t^c to enhance the feature representation of the foggy. To this end, we adopt conditional autoencoder $\boldsymbol{\epsilon}_\theta(x_t^c, t, x^r)$ that receives x_t^c and reference feature x^r from the weather codebook $\mathcal{Z}$. Specifically, $\boldsymbol{\epsilon}_\theta$ estimates mean $\boldsymbol{\mu}_\theta$ and variance $\boldsymbol{\Sigma}_\theta$ of the fog distribution at the t-th timestep, denoted as $\tilde{\mathcal{F}}(\cdot)$. The reverse process is also repeated over T timesteps, which can be represented as:

$$p_\theta(x_{t-1}^c|x_t^c, x^r) = \tilde{\mathcal{F}}(x_{t-1}^c; \boldsymbol{\mu}_\theta(x_t^c, t, x^r), \boldsymbol{\Sigma}_\theta(x_t^c, t)), \tag{8}$$

$$p_\theta(x^c, ..., x_T^c) = p(x_T^c) \prod_{t=1}^{T} p_\theta(x_{t-1}^c|x_t^c, x^r), \tag{9}$$

where $p_\theta(x_{t-1}^c|x_t^c, x^r)$ includes cross-attention layer in $\boldsymbol{\epsilon}_\theta$.

The cross-attention layer receives the flattened clear feature x_t^c and weather-reference feature x^r, *i.e.,* $\bar{x}_t^c$ and $\bar{x}^r$, respectively. The similarity between $\bar{x}^r$ and $\bar{x}_t^c$ is calculated in the cross-attention layer to transfer the enhancement to $\bar{x}_t^c$, which can be formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{10}$$

where $Q = W_i^q \cdot \bar{x}_t^c, K = W_i^k \cdot \bar{x}^r, V = W_i^v \cdot \bar{x}^r$ with learnable parameters W_i^q, W_i^k, W_i^v.

To ensure that the estimated fog variant $\boldsymbol{\epsilon}_\theta$ is similar to the fog variant $\boldsymbol{\epsilon}_n$ applied to make x_t^c from the input clear feature x^c, we propose a weather-adaptive enhancement loss $\mathcal{L}_{wae}$, which is formulated as:

$$\mathcal{L}_{wae} = \mathbb{E}_{x^c, \boldsymbol{\epsilon}_n \sim \mathcal{F}, t}\Big[\|\boldsymbol{\epsilon}_n - \boldsymbol{\epsilon}_\theta(x_t^c, t, x^r)\|_2^2\Big]. \tag{11}$$

With $\mathcal{L}_{wae}$, $\boldsymbol{\epsilon}_\theta$ can estimate the fog variant by leveraging the fog distribution as noise for our diffusion model through multiple forward/reverse processes. Additionally, the cross-attention layer within $\boldsymbol{\epsilon}_\theta$ dynamically enhances the feature representation based on combining knowledge of the input feature (foggy or clear) and the weather-reference feature. Since our diffusion model learns the

degree of improvement needed corresponding to weather conditions, it can improve the representation of any input, whether clear or foggy in inference phase. This leads our monocular 3D object detector can handle both clear and foggy input images, resulting in weather-robust detection.

3.3 Total Loss Function

The total loss function of our MonoWAD is represented as follows:

$$\mathcal{L}_{Total} = \mathcal{L}_{OD} + \lambda_1 \mathcal{L}_{ckr} + \lambda_2 \mathcal{L}_{wae}, \tag{12}$$

where λ_1 and λ_2 denote balancing hyper-parameters. In our experiments, we set $\lambda_1 = \lambda_2 = 1$. $\mathcal{L}_{OD}$ is the detection loss for 3D object detection [16,55]. It includes loss functions of classification, regression, and depth loss, that are similar to prior works [8, 16, 31]. The overall weight parameters are updated through $\mathcal{L}_{Total}$.

4 Experiments

4.1 Dataset and Evaluation Metrics

Datasets. We utilize the KITTI 3D object detection dataset [12], the most widely adopted for 3D object detection. It contains 7,481 training images and 7,518 test images under clear weather conditions. Due to unavailable ground-truth annotations for test images and limited evaluation on the test server, we follow [6] by splitting 3,712 training images and 3,769 validation images. In addition, as our work requires paired images to learn about weather changes, we generate foggy images from all images in the KITTI dataset, called foggy KITTI, that emulate the foggy scene. Following the protocols of [2,43], we generate foggy images based on object distances using depth maps estimated by DORN [10]. Please refer to the supplementary materials for details.

In addition, as our work is focused on robust monocular 3D object detection in various weather conditions, we further adopt the Virtual KITTI dataset [5,11], which contains photo-realistic synthetic images under various weather conditions (*e.g.,* foggy, rainy, sunset). It is associated with the original real-world KITTI dataset, which has 3D annotations for each weather.

Evaluation Metrics. We adopt average precision in both 3D detection (AP_{3D}) and bird-eye view detection (AP_{BEV}) under three difficulty levels ('Easy', 'Moderate', 'Hard') according to size, occlusion, and truncation. Following [47], we use 40 recall position metric AP_{40} and report scores for the car category under the IoU threshold 0.7 for the KITTI dataset and 0.5 for the Virtual KITTI dataset.

4.2 Implementation Details

We use DLA-102 [52] as our backbone and adopt the transformer architecture of [16] for the detection block. We train MonoWAD on a single RTX 4090 GPU with a batch size of 4 over 120 epochs using Adam optimizer [23] (initial learning

Table 1: Detection results of car category on KITTI validation set under foggy weather and clear weather conditions. The results of the state-of-the-art methods under foggy weather are obtained through our reproduction with the official source code. **Bold**/underlined fonts indicate the best/second-best results.

Method	Foggy (AP_{3D})			Clear (AP_{3D})			Average		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
GUPNet [28] (ICCV'21)	2.74	2.19	2.16	22.76	16.46	13.72	12.75	9.33	7.94
DID-M3D [37] (ECCV'22)	1.15	0.61	0.64	22.98	16.12	14.03	12.07	8.37	7.34
MonoGround [40] (CVPR'22)	0.00	0.00	0.06	25.24	18.69	15.58	12.62	9.35	7.82
MonoDTR [16] (CVPR'22)	16.89	11.86	9.87	24.52	18.57	15.51	20.71	15.22	12.69
MonoDETR [55] (ICCV'23)	7.40	5.74	4.53	28.84	20.61	16.38	18.12	13.18	10.46
MonoWAD (Ours)	**27.17**	**19.57**	**16.21**	**29.10**	**21.08**	**17.73**	**28.14**	**20.33**	**16.97**

Table 2: Detection results of car category on KITTI test set under foggy weather. **Bold**/underlined fonts indicate the best/second-best results.

Method	AP_{3D}			AP_{BEV}		
	Easy	Mod.	Hard	Easy	Mod.	Hard
GUPNet [28] (ICCV'21)	3.01	2.42	1.13	4.90	3.02	2.91
DID-M3D [37] (ECCV'22)	3.10	2.39	2.19	5.34	3.01	2.91
MonoGround [40] (CVPR'22)	0.14	0.20	0.22	0.23	0.38	0.39
MonoDTR [16] (CVPR'22)	11.07	7.41	5.26	15.76	10.15	7.53
MonoDETR [55] (ICCV'23)	9.33	5.54	4.06	13.15	8.06	6.30
MonoWAD (Ours)	**19.75**	**13.32**	**11.04**	**27.95**	**19.06**	**15.61**

rate of 10^{-4}). For the weather codebook, we utilize embedding slot $K = 4096$ and set the dimension of each slot $D = 256$. We set 4 heads for cross-attention of the weather-adaptive diffusion model, and the timesteps for the forward and reverse process of diffusion default to 15. The number of channels for the diffusion output features is $C = 256$. As recent monocular 3D object detectors [16, 28, 37, 40, 55] have not been explored under adverse weather conditions, we implemented them with available official source code to faithfully reproduce them.

4.3 Comparison

Results on KITTI 3D Dataset. We compared MonoWAD with state-of-the-art monocular 3D object detectors [16, 28, 37, 40, 55] that do not use additional data (*e.g.,* depth maps or LiDAR) during inference on the KITTI and foggy KITTI validation sets. Table 1 shows the AP_{3D} results, and AP_{BEV} results are in the supplementary materials. While recent methods have shown improved performance in clear weather, their performance drops in foggy weather, limiting their applicability in real-world applications (*e.g.,* autonomous driving and robotics). In contrast, MonoWAD showed stable 3D detection performance under both foggy and clear weather. Since our weather codebook learns knowledge about clear weather and the weather-adaptive diffusion model can enhance the feature representation of the input images under both clear and foggy weather, MonoWAD shows a more weather-robust detection performance than that of the existing methods. Also, to explore the weather robustness of MonoWAD, we conducted experiments with mixed foggy and clear weather conditions at various ratios (please see the supplementary materials).

Table 3: Detection results of car category on Virtual KITTI under foggy, rainy, sunset conditions. **Bold**/underlined fonts indicate the best/second-best results.

Method	Foggy (AP_{3D})			Rainy (AP_{3D})			Sunset (AP_{3D})		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
GUPNet [28] (ICCV'21)	1.76	1.57	1.57	2.34	1.24	1.21	2.77	1.64	1.65
DID-M3D [37] (ECCV'22)	0.91	0.39	0.39	0.40	0.13	0.13	0.34	0.10	0.10
MonoGround [40] (CVPR'22)	0.29	0.30	0.25	5.49	2.82	2.77	7.68	4.24	4.20
MonoDTR [16] (CVPR'22)	8.79	5.75	5.72	11.73	6.25	6.74	9.86	5.42	5.42
MonoDETR [55] (ICCV'23)	4.50	2.99	2.96	6.61	3.46	3.42	7.08	4.17	4.16
MonoWAD (Ours)	**13.33**	**8.56**	**8.50**	**14.12**	**8.33**	**8.24**	**13.38**	**7.89**	**7.80**

Table 4: Effect of the proposed method on KITTI validation set for car category. **WC** denotes our weather codebook, **WAD** indicates our weather-adaptive diffusion model.

Method	WC	WAD	Foggy (AP_{3D})			Clear (AP_{3D})		
			Easy	Mod.	Hard	Easy	Mod.	Hard
Baseline	-	-	13.75	9.61	8.10	22.63	17.16	14.28
MonoWAD (Ours)	-	✓	25.62	18.66	15.56	26.34	19.17	16.15
	✓	✓	**27.17**	**19.57**	**16.21**	**29.10**	**21.08**	**17.73**

Table 5: Detection results on KITTI validation set by changing diffusion timestep T.

Timestep T	Foggy (AP_{3D})			Clear (AP_{3D})		
	Easy	Mod.	Hard	Easy	Mod.	Hard
-	13.75	9.61	8.10	22.63	17.16	14.28
5	23.57	17.91	14.96	26.03	19.21	16.01
10	25.28	18.49	15.42	26.79	19.90	16.78
15	**27.17**	**19.57**	**16.21**	**29.10**	**21.08**	**17.73**
20	24.54	18.29	15.10	24.85	18.54	15.34

We further compared 3D detection performances on the foggy KITTI test set (Table 2). Similar to Table 1, existing monocular 3D object detectors show lower performance under foggy weather. Through the results, our method shows the weather-robust monocular 3D performance under foggy and clear weather.

Results on Virtual KITTI Dataset. We also conducted experiments on the Virtual KITTI dataset to validate the generalization ability of our method across various weather conditions (*i.e.,* foggy, rainy, sunset). The results are shown in Table 3. MonoWAD also shows the highest performance in foggy weather. Moreover, it outperforms the existing method in the rainy and sunset conditions. The results demonstrate that MonoWAD is robust and insensitive to various weather conditions that can be faced in real-world scenarios, not just clear weather.

4.4 Ablation Studies

We conducted ablation studies to examine (1) the effect of each proposed component (*i.e.,* weather codebook and weather-adaptive diffusion model) and (2) the effect of the weather-adaptive diffusion model by varying the timestep T. These experiments were performed on the KITTI and foggy KITTI 3D validation sets.

Effect of the Proposed Modules. The results regarding the effectiveness of the proposed weather codebook and weather-adaptive diffusion model are pre-

Table 6: Performance comparison on KITTI validation set under foggy and clear weather conditions. We compare our MonoWAD with MonoDTR [16], which demonstrates superior performance in foggy weather using state-of-the-art dehazing methods.

Method	Dehazing Method	Foggy (AP_{3D})			Clear (AP_{3D})			Average		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
MonoDTR($\mathcal{B}$) [16] (CVPR'22)	-	16.89	11.86	9.87	24.52	18.57	15.51	20.71	15.22	12.69
$\mathcal{B}$ + RIDCP [49] (CVPR'23)	Image-level	17.23	12.41	10.44	24.02	17.89	14.78	20.63	15.15	12.61
$\mathcal{B}$ + DENet [39] (ACCV'22)	Feature-level	22.35	17.44	14.47	7.10	5.70	4.53	14.73	11.57	9.50
$\mathcal{B}$ + Yang *et al.* [51] (ACCV'22)	Feature-level	22.87	15.21	12.17	17.96	13.10	10.64	20.42	14.16	11.41
MonoWAD (Ours)	-	**27.17**	**19.57**	**16.21**	**29.10**	**21.08**	**17.73**	**28.14**	**20.33**	**16.97**

Table 7: Comparison of diffusion models on KITTI validation set for car category: $\mathcal{B}$ is baseline detection block (transformer encoder-decoder), **CDM** (Conditional Diffusion Model [42]), **WC** (weather codebook), and **WAD** (weather adaptive diffusion model).

Method	Foggy (AP_{3D})			Clear (AP_{3D})		
	Easy	Mod.	Hard	Easy	Mod.	Hard
$\mathcal{B}$+DDPM [15]	5.32	3.84	2.77	18.31	12.71	10.20
$\mathcal{B}$+CDM [42]	2.74	2.10	2.00	20.50	14.51	11.63
$\mathcal{B}$+WC+CDM [42]	17.51	12.74	10.40	21.05	15.11	12.54
$\mathcal{B}$+WAD	25.62	18.66	15.56	26.34	19.17	16.15
MonoWAD ($\mathcal{B}$+WC+WAD)	**27.17**	**19.57**	**16.21**	**29.10**	**21.08**	**17.73**

sented in Table 4. Since our weather-adaptive diffusion model is designed to enhance feature representation, it alone has shown significant performance improvement. When the weather codebook is additionally considered, the weather-adaptive diffusion model can leverage the knowledge of the weather-reference feature for clear weather, leading to enhanced performance. It makes MonoWAD can be robust to various weather conditions for monocular 3D object detection.

Effect of Timestep *T*. We also conduct experiments by varying timestep T of the weather-adaptive diffusion model. Timestep is the number of steps for the forward and reverse process of the diffusion model. Table 5 indicates that the highest monocular 3D detection performance is achieved when $T = 15$ under both clear and foggy weather conditions. Our MonoWAD consistently outperforms the baseline and other existing methods across all timesteps.

4.5 Discussions

Comparison with Dehazing Methods. We investigate the weather-robustness of our method for monocular 3D object detection compared to the other monocular detectors using dehazing methods. To this end, we compared MonoWAD with MonoDTR [16], which shows the best performance in foggy weather through the application of state-of-the-art image-level and feature-level dehazing methods [39, 49, 51]. The results are shown in Table 6. Recent dehazing methods are primarily focused on specific weather conditions, such as foggy weather. While they have shown some improvement under foggy conditions, they exhibit reduced performances under clear weather conditions. In contrast, our MonoWAD shows robust performance across both foggy and clear weather conditions.

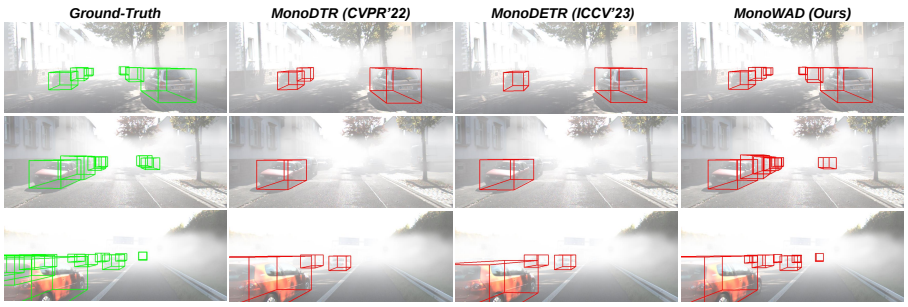

Fig. 5: Comparison of 3D detection examples (green: ground-truth, red: predicted 3D bounding-box) between our MonoWAD and two detectors, MonoDTR [16] and MonoDETR [55], that show the most improved performances among existing methods.

Fig. 6: 3D detection results on real-world images of various weather conditions.

Effect of Weather-Adaptive Diffusion Model. Table 7 shows the effectiveness of MonoWAD with existing diffusion models [15, 42] on the KITTI and foggy KITTI validation sets. Existing methods adopt Gaussian noise for forward and reverse processes, but they can not fully understand about weather. In contrast, our weather-adaptive diffusion understands weather variances, allowing our MonoWAD to surpass existing methods in clear and foggy weather.

4.6 Visualization Results

Results on KITTI Dataset. We visualize several 3D detection results on the KITTI 3D dataset, comparing MonoWAD with MonoDTR and MonoDETR, which exhibit the highest performances among existing methods under foggy condition (Fig. 5). Existing methods struggle to detect objects obscured by fog, indicating limitations in detecting only fully visible objects. In contrast, even

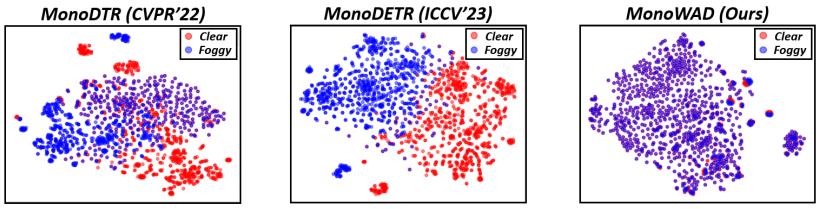

Fig. 7: t-SNE visualization results (Red: Clear, Blue: Foggy).

in dense fog, MonoWAD effectively detects both close and fog-obscured objects with the aid of the weather codebook and weather-adaptive diffusion model.

Results on Real-World Images. We further visualize 3D detection results on real-world images from the Seeing Through Fog dataset [2] under various weather conditions (*i.e.,* foggy, rainy, snowy). In Fig. 6, MonoWAD shows robust detection under diverse weather conditions. This demonstrates that our proposed method maintains weather-robustness even in real-world scenarios by dynamically enhancing the input scenarios.

t-SNE Visualization. We conducted t-SNE visualization to analyze feature representations of MonoDTR, MonoDETR, and our MonoWAD on the KITTI and foggy KITTI validation set. As depicted in Fig. 7, the existing methods exhibit distinct feature representations for foggy and clear weather conditions. In contrast, MonoWAD, leveraging weather-robust feature learning from the weather codebook and weather-adaptive diffusion model, demonstrates similar feature representations for both clear and foggy weather conditions.

4.7 Limitations

The experimental results show the weather-robustness of our method. However, due to the iterative nature of the diffusion model, our method shows 144ms/image at timestep $T = 15$ (110ms/image at $T = 5$), slower than the latest work, MonoDETR (38ms/image). Moreover, our method dynamically enhances the representation of the input feature based on the weather-reference feature and weather difference which needs paired images. Thus, exploring a method to achieve faster processing speeds while maintaining weather-robust performance without paired images could be an interesting direction for our future work.

5 Conclusion

We proposed MonoWAD, a novel weather-robust monocular 3D object detector to handle various weather conditions. Addressing challenges in applying existing monocular 3D object detectors to real-world scenarios with various weather, we design a weather codebook with clear knowledge recalling loss to memorize the knowledge of the clear weather and to generate a weather-reference feature from both clear and foggy features. Also, we design a weather-adaptive diffusion model with weather-adaptive enhancement loss to enhance feature representation according to the weather conditions. As a result, our MonoWAD can detect objects occluded by fog and perform well in clear weather.

Acknowledgements

This work was supported by the NRF grant funded by the Korea government (MSIT) (No. RS-2023-00252391), and by IITP grant funded by the Korea government (MSIT) (No. RS-2022-00155911: Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University), IITP-2023-RS-2023-00266615: Convergence Security Core Talent Training Business Support Program, No. 2022-0-00124: Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities).

References

1. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: Int. Conf. Learn. Represent. (2021)
2. Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., Heide, F.: Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
3. Brazil, G., Liu, X.: M3d-rpn: Monocular 3d region proposal network for object detection. In: Int. Conf. Comput. Vis. (2019)
4. Brazil, G., Pons-Moll, G., Liu, X., Schiele, B.: Kinematic 3d object detection in monocular video. In: Eur. Conf. Comput. Vis. Springer (2020)
5. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2 (2020)
6. Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., Urtasun, R.: Monocular 3d object detection for autonomous driving. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
7. Chen, Y., Tai, L., Sun, K., Li, M.: Monopair: Monocular 3d object detection using pairwise spatial relationships. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
8. Ding, M., Huo, Y., Yi, H., Wang, Z., Shi, J., Lu, Z., Luo, P.: Learning depth-guided convolutions for monocular 3d object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (2020)
9. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
10. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018)
11. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
12. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: IEEE Conf. Comput. Vis. Pattern Recog. (2012). https://doi.org/10.1109/CVPR.2012.6248074
13. Hahner, M., Sakaridis, C., Dai, D., Van Gool, L.: Fog simulation on real lidar point clouds for 3d object detection in adverse weather. In: Int. Conf. Comput. Vis. (2021)
14. Hamilton, B., Tefft, B., Arnold, L., Grabowski, J.: Hidden highways: Fog and traffic crashes on america's roads (november 2014). Montana **40** (2006)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Adv. Neural Inform. Process. Syst. vol. 33 (2020)

16. Huang, K.C., Wu, T.H., Su, H.T., Hsu, W.H.: Monodtr: Monocular 3d object detection with depth-aware transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
17. Juneja, A., Kumar, V., Singla, S.K.: A systematic review on foggy datasets: Applications and challenges. Archives of Computational Methods in Engineering **29**(3) (2022)
18. Kim, J.U., Kim, H.I., Ro, Y.M.: Stereoscopic vision recalling memory for monocular 3d object detection. IEEE Trans. Image Process. **32** (2023). https://doi.org/10.1109/TIP.2023.3274479
19. Kim, J.U., Park, S., Ro, Y.M.: Robust small-scale pedestrian detection with cued recall via memory learning. In: Int. Conf. Comput. Vis. (2021)
20. Kim, J.U., Park, S., Ro, Y.M.: Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection. IEEE Trans. Circuit Syst. Video Technol. **32**(3) (2021)
21. Kim, J.U., Park, S., Ro, Y.M.: Towards versatile pedestrian detector with multisensory-matching and multispectral recalling memory. In: AAAI. vol. 36 (2022)
22. Kim, S.W., Brown, B., Yin, K., Kreis, K., Schwarz, K., Li, D., Rombach, R., Torralba, A., Fidler, S.: Neuralfield-ldm: Scene generation with hierarchical latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019)
25. Liu, X., Xue, N., Wu, T.: Learning auxiliary monocular contexts helps monocular 3d object detection. In: AAAI. vol. 36 (2022)
26. Liu, Z., Wu, Z., Toth, R.: Smoke: Single-stage monocular 3d object detection via keypoint estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. (2020)
27. Liu, Z., Zhou, D., Lu, F., Fang, J., Zhang, L.: Autoshape: Real-time shape-aware monocular 3d object detection. In: Int. Conf. Comput. Vis. (2021)
28. Lu, Y., Ma, X., Yang, L., Zhang, T., Liu, Y., Chu, Q., Yan, J., Ouyang, W.: Geometry uncertainty projection network for monocular 3d object detection. In: Int. Conf. Comput. Vis. (2021)
29. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
30. Luo, J., Li, Y., Pan, Y., Yao, T., Feng, J., Chao, H., Mei, T.: Semantic-conditional diffusion networks for image captioning. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
31. Luo, S., Dai, H., Shao, L., Ding, Y.: M3dssd: Monocular 3d single stage object detector. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
32. Ma, X., Liu, S., Xia, Z., Zhang, H., Zeng, X., Ouyang, W.: Rethinking pseudo-lidar representation. In: Eur. Conf. Comput. Vis. Springer (2020)
33. Mai, N.A.M., Duthon, P., Khoudour, L., Crouzil, A., Velastin, S.A.: 3d object detection with sls-fusion network in foggy weather conditions. Sensors **21**(20) (2021)
34. Mousavian, A., Anguelov, D., Flynn, J., Kosecka, J.: 3d bounding box estimation using deep learning and geometry. In: IEEE Conf. Comput. Vis. Pattern Recog. (2017)

35. Ni, H., Shi, C., Li, K., Huang, S.X., Min, M.R.: Conditional image-to-video generation with latent flow diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
36. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning (2018)
37. Peng, L., Wu, X., Yang, Z., Liu, H., Cai, D.: Did-m3d: Decoupling instance depth for monocular 3d object detection. In: Eur. Conf. Comput. Vis. Springer (2022)
38. Pfeuffer, A., Dietmayer, K.: Robust semantic segmentation in adverse weather conditions by means of sensor data fusion. In: 2019 22th International Conference on Information Fusion (FUSION) (2019). https://doi.org/10.23919/FUSION43075.2019.9011192
39. Qin, Q., Chang, K., Huang, M., Li, G.: Denet: Detection-driven enhancement network for object detection under adverse weather conditions. In: Asian Conf. Comput. Vis. (2022)
40. Qin, Z., Li, X.: Monoground: Detecting monocular 3d objects from the ground. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
41. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
43. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. Int. J. Comput. Vis. **126** (2018)
44. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
45. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019)
46. Shi, X., Ye, Q., Chen, X., Chen, C., Chen, Z., Kim, T.K.: Geometry-based distance decomposition for monocular 3d object detection. In: Int. Conf. Comput. Vis. (2021)
47. Simonelli, A., Bulo, S.R., Porzi, L., Lopez-Antequera, M., Kontschieder, P.: Disentangling monocular 3d object detection. In: Int. Conf. Comput. Vis. (2019)
48. Wang, L., Du, L., Ye, X., Fu, Y., Guo, G., Xue, X., Feng, J., Zhang, L.: Depth-conditioned dynamic message propagation for monocular 3d object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
49. Wu, R.Q., Duan, Z.P., Guo, C.L., Chai, Z., Li, C.: Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
50. Yang, S., Scherer, S.: Cubeslam: Monocular 3-d object slam. IEEE Transactions on Robotics **35**(4) (2019). https://doi.org/10.1109/TRO.2019.2909168
51. Yang, X., Mi, M.B., Yuan, Y., Wang, X., Tan, R.T.: Object detection in foggy scenes by embedding depth and reconstruction into domain adaptation. In: Asian Conf. Comput. Vis. (2022)
52. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018)
53. Yurtsever, E., Lambert, J., Carballo, A., Takeda, K.: A survey of autonomous driving: Common practices and emerging technologies. IEEE Access **8** (2020). https://doi.org/10.1109/ACCESS.2020.2983149

54. Zhang, C., Wang, H., Cai, Y., Chen, L., Li, Y., Sotelo, M.A., Li, Z.: Robust-fusionnet: Deep multimodal sensor fusion for 3-d object detection under severe weather conditions. IEEE Transactions on Instrumentation and Measurement **71** (2022). https://doi.org/10.1109/TIM.2022.3191724
55. Zhang, R., Qiu, H., Wang, T., Guo, Z., Cui, Z., Qiao, Y., Li, H., Gao, P.: Monodetr: Depth-guided transformer for monocular 3d object detection. In: Int. Conf. Comput. Vis. (2023)
56. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. (2018)
57. Zhou, Y., He, Y., Zhu, H., Wang, C., Li, H., Jiang, Q.: Monocular 3d object detection: An extrinsic parameter free approach. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)