# MonoWAD: Weather-Adaptive Diffusion Model for Robust Monocular 3D Object Detection – *Supplementary Material* –

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

Additional results and discussions of our supplementary are as follows:

- Section A: Additional details about Foggy KITTI dataset and MonoWAD.
- Section B: Additional results on KITTI 3D dataset (*i.e.,* weather-robustness, BEV results, fog density).
- Section C: Additional results on Virtual KITTI dataset.
- Section D: Additional results on Real-World dataset.
- Section E: Qualitative comparison with dehazing method.
- Section F: Additional visualization results.
- Section G: Video demo.


**Fig. 1:** Examples with fog densities $\delta = \{0.05, 0.1, 0.15, 0.3\}$.

## A. Additional Details about Foggy KITTI Dataset and MonoWAD

**Foggy KITTI Dataset.** Given image $I$, we adopt pre-trained DORN [2] to obtain depth map $I_D$ and calculate transmittance $T(I_D, \delta)$ using $I_D$ and fog density $\delta$. After estimating atmospheric light $I_A$ from $I$, foggy KITTI is obtained via Eq. 1. Following [1, 11], we can generate various foggy images via $\delta$={0.05,0.1,0.15, 0.3} (Fig. 1). In all experiments of our main paper, we set a fog density $\delta$=0.1.

$$I_F = (I * T(I_D, \delta) + I_A * (1 - T(I_D, \delta)). \tag{1}$$

In addition, unlike the Multifog KITTI dataset [8], our foggy KITTI utilizes depth information inferred from monocular images to generate photo-realistic fog data for monocular 3D object detection. Moreover, the various densities are provided separately, rather than integrated.

† Corresponding author

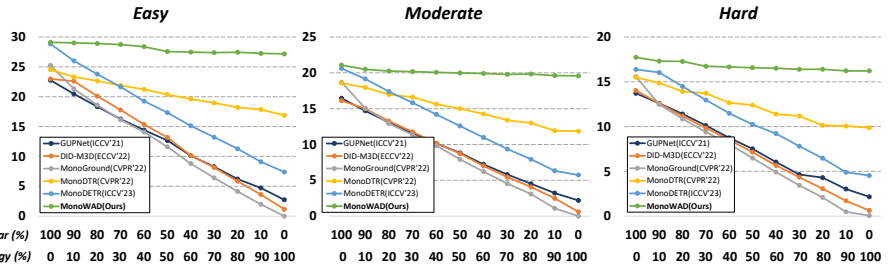


**Fig. 2:** Performance variations of car category on KITTI validation set under various weather conditions, including foggy weather (foggy) and clear weather (clear) based on its percentage. 'Clear (n%) and Foggy (m%)' indicates that n% images of the validation set correspond to clear weather, and m% images correspond to foggy weather.

**MonoWAD in Clear Weather.** In the training process, our weather codebook (WC) and weather-adaptive diffusion model (WAD) learn clear features via clear knowledge recalling (CKR) loss $\mathcal{L}_{ckr}$ and weather-adaptive enhancement loss $\mathcal{L}_{wae}$ to enhance the weather and emphasize feature by cross attention, and detection loss $\mathcal{L}_{OD}$ to enhance the features of the backbone for detection. This is different from performing detection by dehazing fog, as it serves to remove fog while emphasizing features. It also dynamically enhances the feature representation of input images (clear or foggy), allowing it to perform robustly in both clear and foggy weather conditions.

**Details of Weather Codebook.** We employ a single weather codebook in our MonoWAD. The weather codebook has 1.05M parameters, which is 1.9% of the total 54.25M parameters in the baseline model. With a single codebook, ours can learn to memorize the knowledge of clear weather using the clear knowledge recalling (CKR) loss $\mathcal{L}_{ckr}$ and generate reference features for other weather conditions (*e.g.,* foggy, rainy, snowy) (Eq.3 of our main paper).

**Details of Weather-Adaptive Diffusion Model.** We further provide a more detailed explanation of our weather-adaptive diffusion model, including the noise in the forward process and enhancement in the reverse process. In the forward process, fog distribution $\mathcal{F} = x^f - x^c$ is the difference between clear and fog features, used as our diffusion noise. Fog variant $\epsilon_n$ is applied based on a fixed Markov Chain of $T$ timesteps determined by variance schedule $\beta_t$. During inference, our diffusion model estimates the mean $\boldsymbol{\mu}_\theta$ and variance $\boldsymbol{\Sigma}_\theta$ at each timestep, and $\mathcal{F}$ is estimated by aggregating them across all timesteps. Following [4], we set variance $\boldsymbol{\Sigma}_\theta(x_t^c, t)$ to be $\sigma_t^2 \boldsymbol{I}$, where $\sigma_t^2 = \beta_t$. In the reverse process, the weather-adaptive diffusion model consists of an autoencoder (U-Net) that has an encoder/mid-block/decoder with no additional backbone. In this architecture, cross-attention between the mid-block feature and the weather-reference feature from weather codebook is conducted. It takes the previous step $x_T^c$ as input and predicts the next step $x_{T-1}^c$. As shown in Fig. 4 of our main paper, we operate the same autoencoder at different timesteps to gradually enhance

**Table 1:** Performance ($AP_{3D}$) variations of car category on KITTI validation set under weather conditions, including foggy weather (foggy) and clear weather (clear) based on its percentage. 'Clear(n%)+Foggy(m%)' indicates that n% images of the validation set correspond to clear weather, and m% images correspond to foggy weather.

| Method | Clear(70%)+Foggy(30%) | | | Clear(50%)+Foggy(50%) | | | Clear(30%)+Foggy(70%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Easy | Mod. | Hard | Easy | Mod. | Hard | Easy | Mod. | Hard |
| GUPNet [7] (ICCV'21) | 16.25 | 11.59 | 10.11 | 12.67 | 8.83 | 7.51 | 8.28 | 5.81 | 4.65 |
| DID-M3D [9] (ECCV'22) | 17.78 | 11.77 | 9.83 | 13.18 | 8.74 | 7.14 | 8.14 | 5.44 | 4.34 |
| MonoGround [10] (CVPR'22) | 16.13 | 11.35 | 9.40 | 11.61 | 7.93 | 6.48 | 6.41 | 4.53 | 3.44 |
| MonoDTR [5] (CVPR'22) | 21.87 | 16.61 | 13.71 | 20.37 | 15.00 | 12.40 | 18.96 | 13.39 | 11.18 |
| MonoDETR [14] (ICCV'23) | 21.65 | 15.83 | 12.97 | 17.33 | 12.59 | 10.23 | 13.22 | 9.37 | 7.81 |
| **MonoWAD (Ours)** | **28.73** | **20.17** | **16.73** | **27.55** | **19.98** | **16.57** | **27.38** | **19.79** | **16.39** |

**Table 2:** Detection results ($AP_{BEV}$) of car category on KITTI validation set under foggy weather and clear weather conditions. **Bold**/underlined fonts indicate the best/second-best results.

| Method | Foggy ($AP_{BEV}$) | | | Clear ($AP_{BEV}$) | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|
| | Easy | Mod. | Hard | Easy | Mod. | Hard | Easy | Mod. | Hard |
| GUPNet [7] (ICCV'21) | 5.13 | 4.37 | 2.93 | 31.07 | 22.94 | 19.75 | 18.10 | 13.66 | 11.34 |
| DID-M3D [9] (ECCV'22) | 2.40 | 1.78 | 0.86 | 31.10 | 22.76 | 19.50 | 16.75 | 12.27 | 10.18 |
| MonoGround [10] (CVPR'22) | 0.00 | 0.00 | 0.07 | 32.68 | 24.79 | 20.56 | 16.34 | 6.20 | 10.32 |
| MonoDTR [5] (CVPR'22) | 22.01 | 14.84 | 12.74 | 33.33 | 25.35 | 21.68 | 27.67 | 20.10 | 17.21 |
| MonoDETR [14] (ICCV'23) | 11.03 | 7.26 | 5.69 | 37.86 | 26.95 | 22.80 | 18.12 | 17.11 | 14.25 |
| **MonoWAD (Ours)** | **35.70** | **25.31** | **21.43** | **38.07** | **26.97** | **23.04** | **36.89** | **26.14** | **22.24** |

the representation from $x_T^c$ to $x_0^c$. We trained the weather codebook, weather-adaptive diffusion model, and detection block as a single model in an end-to-end manner, without requiring any additional data.

## B. Additional Results on KITTI 3D Dataset

**Weather-Robustness Experiments.** As we have mentioned mixed foggy and clear weather conditions as an extension of the weather-robustness experiments in Section 4.3 (Results on KITTI 3D Dataset) of our main paper, we further compared the 3D detection performance under the clear and foggy validation set based on its percentage (Clear/Foggy: 100%/0% to 0%/100% balancing in 10% intervals). We conduct experiments by selecting random images from both foggy and clear weather according to predetermined seeds, ensuring that all models are tested under identical conditions. The results are shown in Fig. 2. As the ratio of the foggy increased, the performance of the existing methods gradually decreased. For example, when the clear/foggy ratio changed from 70%/30% (Table 1) to 30%/70% (Table 1), the performance of MonoDETR dropped significantly from (21.65, 15.83, and 12.97) to (13.22, 9.37, and 7.81) for ('Easy', 'Moderate', and 'Hard') settings, respectively. In contrast, the performance change of our method is marginal even when we vary the ratios of clear and foggy conditions. The experimental results demonstrate the weather-robustness property of our method.

**BEV Results on KITTI validation set.** We further compared the $AP_{BEV}$ on KITTI [3] and foggy KITTI validation set in Table 2. Similar to Table 1 of

**Table 3:** Detection results ($AP_{3D}$) of car category on foggy KITTI validation set under various foggy densities $\delta = \{0.05, 0.15, 0.3\}$ ($\delta = 0.1$ is in main paper). 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 | $\delta = 0.05$ | | | $\delta = 0.15$ | | | $\delta = 0.3$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Easy** | **Mod.** | **Hard** | **Easy** | **Mod.** | **Hard** | **Easy** | **Mod.** | **Hard** |
| GUPNet [7] (ICCV'21) | 7.29 | 5.16 | 4.16 | 0.64 | 0.93 | 0.88 | 0.00 | 0.00 | 0.00 |
| DID-M3D [9] (ECCV'22) | 9.66 | 6.90 | 5.46 | 0.50 | 0.74 | 0.77 | 0.00 | 0.00 | 0.00 |
| MonoGround [10] (CVPR'22) | 0.53 | 0.28 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| MonoDTR [5] (CVPR'22) | 22.42 | 16.24 | 13.09 | 11.38 | 7.27 | 5.74 | 2.24 | 1.89 | 1.85 |
| MonoDETR [14] (ICCV'23) | 15.06 | 10.70 | 8.89 | 3.61 | 2.92 | 2.02 | 0.36 | 0.36 | 0.36 |
| **MonoWAD (Ours)** | **26.99** | **19.19** | **15.88** | **15.48** | **10.71** | **8.60** | **9.66** | **6.90** | **5.46** |

**Table 4:** Detection results ($AP_{3D}$) of car category on Virtual KITTI under foggy, rainy, and sunset conditions, are based on an equal percentage mix of these weather conditions. **Bold**/underlined fonts indicate the best/second-best results.

| Method | **Foggy/Rainy/Sunset (33.3%)** | | | **Foggy/Rainy/Sunset (33.3%)** | | |
|---|---|---|---|---|---|---|
| | **Easy** | **Mod.** | **Hard** | **Easy** | **Mod.** | **Hard** |
| GUPNet [7] (ICCV'21) | 2.29 | 1.21 | 1.19 | 9.76 | 5.58 | 5.56 |
| DID-M3D [9] (ECCV'22) | 0.40 | 0.13 | 0.13 | 5.37 | 3.25 | 3.21 |
| MonoGround [10] (CVPR'22) | 4.39 | 2.50 | 2.43 | 17.27 | 11.29 | 11.21 |
| MonoDTR [5] (CVPR'22) | 10.27 | 5.88 | 5.84 | 22.09 | 14.24 | 14.21 |
| MonoDETR [14] (ICCV'23) | 6.17 | 3.31 | 3.28 | 15.84 | 9.77 | 9.79 |
| **MonoWAD (Ours)** | **13.69** | **8.22** | **8.14** | **29.46** | **18.81** | **18.76** |

our main paper, our MonoWAD outperforms the existing monocular 3D object detector under clear and foggy weather.

**Results under Different Fog Density.** We further compared the $AP_{3D}$ on foggy KITTI validation set under different fog density $\delta = \{0.05, 0.15, 0.3\}$. As shown in Table 3, even as the fog density $\delta$ increases, ours still outperforms the state-of-the-art methods. In Fig 1, we also visualize the 3D detection results on foggy KITTI images of various fog densities, demonstrating the robustness of our method under different visibility conditions.

## C. Additional Results on Virtual KITTI Dataset

**Weather-Robustness Experiments.** We also conducted a weather-robustness experiment under mixed foggy, rainy, and sunset weather conditions. Same as Table 1, we select random images, and we compared 3D detection performance under mixed weather conditions based on an equal percentage (percentage: 33.3%). As shown in Table 4, our MonoWAD outperforms the existing method in the coexisting of various weather conditions. These results demonstrate that our MonoWAD is still robust and insensitive to various weather conditions that can be faced in real-world autonomous driving.

## D. Additional Results on Real-World Dataset

We investigate the transferability to real-world conditions of our method compared to the application of other enhancement methods [12, 13] on two state-of-

**Table 5:** Detection results ($AP_{3D}$) of car category on Seeing Through Fog under various weather conditions (*e.g.,* clear, foggy, rainy, snowy). **Bold**/underlined fonts indicate the best/second-best results.

| Method | Clear ($AP_{3D}$) | | | Foggy ($AP_{3D}$) | | | Rainy ($AP_{3D}$) | | | Snowy ($AP_{3D}$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Easy | Mod. | Hard | Easy | Mod. | Hard | Easy | Mod. | Hard | Easy | Mod. | Hard |
| MonoDTR [5] | 10.08 | 8.71 | 6.98 | 19.26 | 16.66 | 15.37 | 5.30 | 4.99 | 3.53 | 9.05 | 7.24 | 6.35 |
| MonoDTR + RIDCP [13] | 9.44 | 8.57 | 6.95 | 17.22 | 14.48 | 13.48 | 3.85 | 4.32 | 3.67 | 8.12 | 6.66 | 5.28 |
| MonoDTR + ZeroScatter [12] | 7.54 | 7.08 | 5.68 | 13.30 | 11.99 | 10.89 | 3.27 | 3.47 | 2.77 | 6.15 | 5.25 | 4.62 |
| MonoDETR [14] | 17.09 | 12.26 | 9.49 | 26.78 | 18.44 | 16.41 | 11.12 | 7.09 | 5.39 | 15.94 | 10.20 | 8.66 |
| MonoDETR + RIDCP [13] | 16.66 | 11.07 | 9.19 | 25.05 | 17.52 | 15.67 | 9.83 | 6.24 | 4.96 | 14.92 | 9.69 | 8.18 |
| MonoDETR + ZeroScatter [12] | 14.05 | 10.22 | 7.61 | 19.47 | 13.61 | 12.07 | 6.39 | 4.16 | 3.14 | 11.70 | 7.87 | 6.54 |
| **MonoWAD (Ours)** | **20.44** | **14.24** | **10.95** | **30.31** | **20.51** | **18.68** | **15.10** | **9.15** | **6.86** | **19.04** | **12.03** | **10.19** |

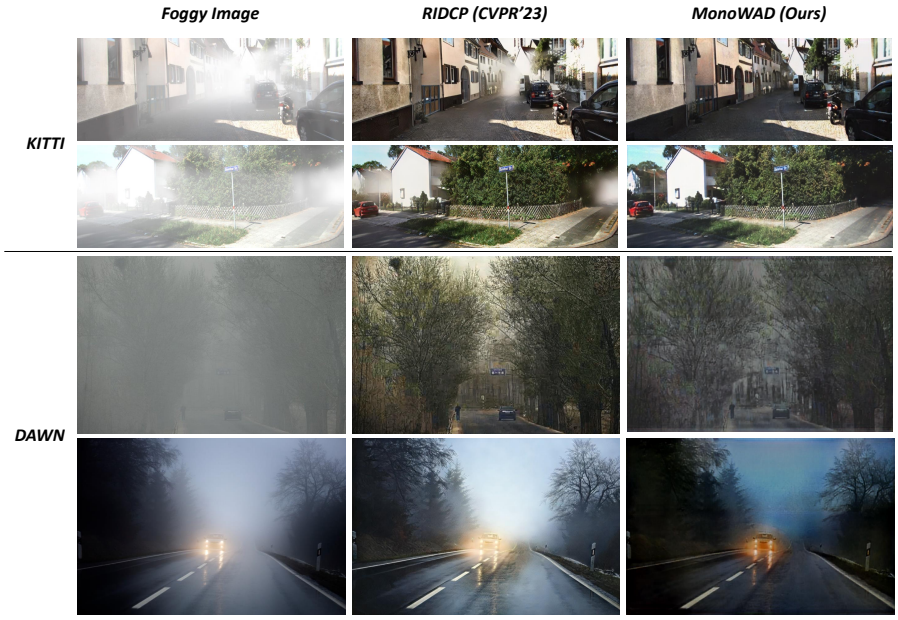


**Fig. 3:** Qualitative comparison on KITTI and DAWN dataset.

the-art detectors [5,14]. To this end, we compared the $AP_{3D}$ on real-world images from the Seeing Through Fog dataset [1]. As shown in Table 5, our MonoWAD consistently outperforms them in various weather, demonstrating its transferability to real-world conditions.

## E. Qualitative Comparison with Dehazing Method

In Section 4.5 (Comparison with Dehazing Methods) of our main paper, we applied dehazing method to an existing monocular 3D object detector [5]. Therefore, we show the results of the state-of-the-art image dehazing method, RIDCP [13], to the foggy KITTI validation set in Fig. 3. We further show the results of our MonoWAD in the dehazing application. Since our MonoWAD is designed for a weather-robust monocular 3D object detector, we performed dehaze by adding a simple decoder architecture to our weather-adaptive diffusion and weather

codebook. Fig. 3 further demonstrates that MonoWAD is effective not only on the foggy KITTI dataset but also on the DAWN dataset [6], which contains real foggy image from real-world scenarios. This shows that our proposed method for dynamically enhancing the feature representation of the input images according to the weather conditions works well and has the potential to be applied to other tasks beyond monocular 3D object detection.

### F. Additional Visualization Results

**Foggy Weather.** We further show the 3D detection results in foggy weather to compare our MonoWAD with MonoDTR [5] and MonoDETR [14], which exhibit the highest performance among existing methods [5, 7, 9, 10, 14] under various foggy scenarios. The results are shown in Fig. 4. The results demonstrate that the proposed MonoWAD effectively detects objects obscured by fog compared to existing methods.

**Clear Weather.** We also visualize the 3D detection results in clear weather to compare our MonoWAD (green) with ground-truth annotations (red). As shown in Fig. 5, the proposed MonoWAD effectively detects objects even in various scenes under clear weather conditions.

**Diverse Weathers on Real-World Images.** We also visualize the 3D detection results in diverse weather conditions (*i.e.,* foggy, rainy, snowy) using real-world images from Seeing Through Fog dataset [1]. As shown in Fig. 6, the proposed MonoWAD effectively detects objects even in various scenes under diverse weather conditions.

### G. Video Demo

We provide video materials to show the detection results of our method and existing methods under various weather conditions (clear and foggy). Please see the video in our official repository.

### References

1. 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)
2. 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)
3. 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

4. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Adv. Neural Inform. Process. Syst. vol. 33 (2020)
5. 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)
6. Kenk, M.A., Hassaballah, M.: Dawn: vehicle detection in adverse weather nature dataset. arXiv preprint arXiv:2008.05402 (2020)
7. 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)
8. 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)
9. 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)
10. Qin, Z., Li, X.: Monoground: Detecting monocular 3d objects from the ground. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
11. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. Int. J. Comput. Vis. **126** (2018)
12. Shi, Z., Tseng, E., Bijelic, M., Ritter, W., Heide, F.: Zeroscatter: Domain transfer for long distance imaging and vision through scattering media. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
13. 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)
14. 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)


**Fig. 4:** Comparison of 3D detection examples on foggy KITTI dataset (green: ground-truth, red: predicted 3D bounding-box) between our MonoWAD and two detectors, MonoDTR [5] and MonoDETR [14], that show the most improved performances among existing methods.

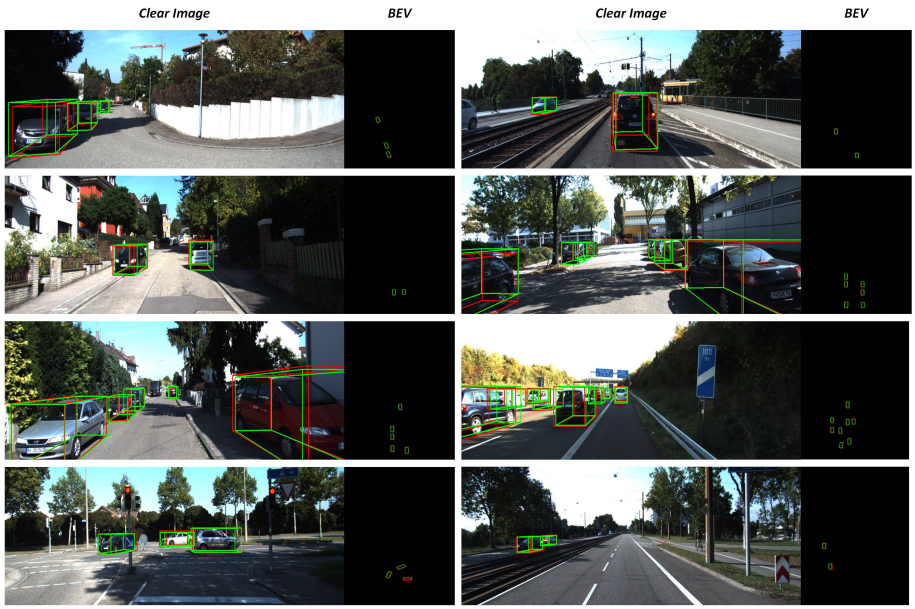


**Fig. 5:** Comparison of 3D detection examples in the image plane and BEV plane under clear weather KITTI dataset (red: ground-truth, green: predicted 3D bounding-box of our MonoWAD).

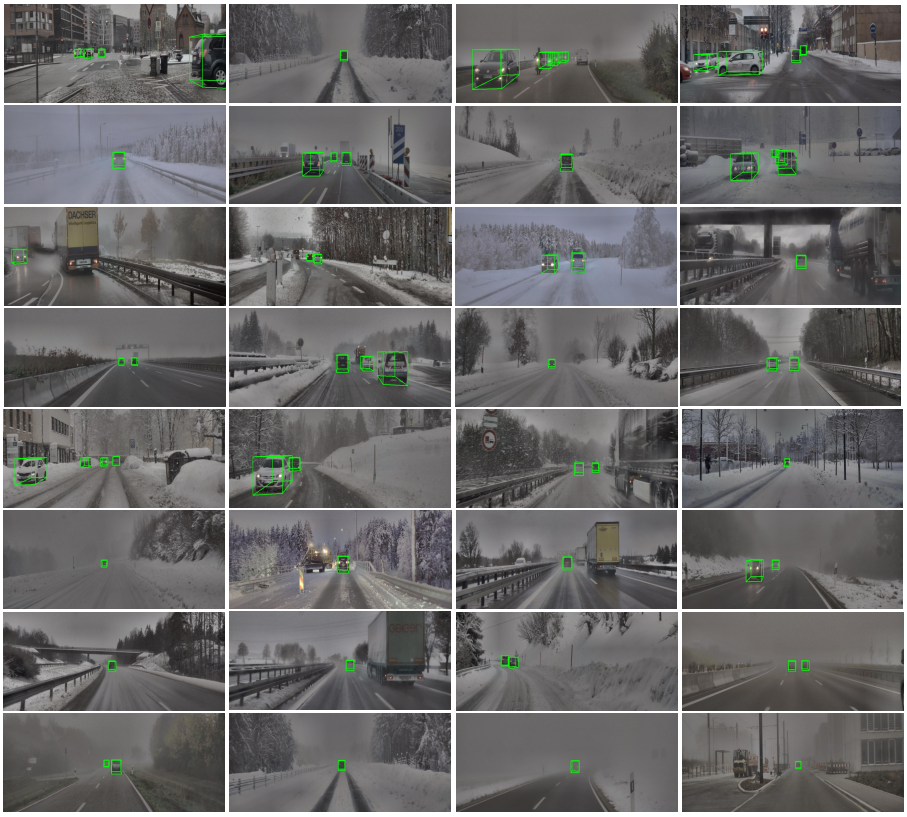

**Fig. 6:** 3D detection results on real-world images of various weather conditions (*e.g.*, foggy, rainy, snowy).