Supplementary Materials

Anonymous Authors

1 EVALUATION METRICS

Similar to [2], we employ the following evaluation metrics in our experiments,

- **AbsRel:** $\frac{1}{|\mathcal{M}|}\sum_{d\in\mathcal{M}}\left|d-d_{gt}\right|/d_{gt}$;
- **SqRel:** $\frac{1}{|\mathcal{M}|}\sum_{d\in\mathcal{M}}\left\|d-d_{gt}\right\|^2/d_{gt}$;
- **RMSE**: $\sqrt{\frac{1}{|\mathcal{M}|}\sum_{d\in\mathcal{M}}\left\|d-d_{gt}\right\|^2}$;
- **RMSElog**: $\sqrt{\frac{1}{|\mathcal{M}|}\sum_{d\in\mathcal{M}}\left\|\log(d)-\log(d_{gt})\right\|^2}$;
- $\mathbf{a}_t$: percentage of d such that $\max(\frac{d}{d_{gt}},\frac{d_{gt}}{d}) < 1.25^t$

where d_{gt} and d denote the GT and estimated pixel depth, $\mathcal{M}$ is the valid mask set to $1e-3 < d_{gt} < 80$. Note that we also use the median scaling technique introduced by [3] on d to recover the absolute depth scale.

2 FORMULA SYMBOL SUMMARY

In this paper, to explain our method clearly, we use numerous symbols in the formulas and the meaning of each symbol can be found at its first appearance. However, for the readers' convenience, we once again list the symbol meanings in Table. 1. Note that we only record meta symbols here and do not include compound symbols with superscripts and subscripts. The meaning of any symbol in this paper can be grasped through combinations. (e.g. $\mathbf{d}_{S,clr}$ denote the depth prediction of the clear images from the student model).

3 THE SNOWY DATASET

In WeatherDepth[1], Wang et al. use the CADC dataset to validate their model's performance in snowy scenes. However, as shown in Fig. 1(a), the LiDAR point clouds in this dataset have many errors on pedestrians and road signs. Furthermore, there is a vehicle bonnet at the bottom of each image, so Wang et al. have to crop them and the sky part, which greatly disrupts the depth clues brought by the vertical position. Hence, the experiments on this dataset can only give a rough result and we need a better depth-annotated snowy dataset.

In this work, we chose the Dense dataset. As depicted in Fig. 1(b), the LiDAR points in Dense are more accurate and only require cropping a small portion of the image. All of these factors make the comparison result reasonable.

4 ADDITIONAL QUALITATIVE RESULTS

To further vividly demonstrate the superiority of our method, as shown in Fig. 2, we present a visual comparison of each scene from both synthetic and real datasets. **(a)** is clear scene. The highlighted area in the dashed box exemplifies how pedestrians wearing white clothing are accurately distinguished from the background by D4RD. **(b)** is synthetic light rainy scene. Despite being affected by camera blur, D4RD still recognizes the correct shape of the streetlights. **(c)** is synthetic light snowy scene. D4RD stands out as the only RMDE model capable of not mistaking train windows for background depth. **(d)** is synthetic light foggy scene. D4RD correctly recognizes that the wall is a plane, avoiding concave depth errors. **(e)** is synthetic adverse rainy scene. The influence of railway tracks on depth should be negligible, but under the interference of raindrop particles, only D4RD exhibits correct depth structures. **(f)** is synthetic adverse snowy scene. D4RD not only identifies the white car far away but also recognizes the almost invisible road sign (on the left side of the image). **(g)** is synthetic adverse foggy scene. From the clear image, it can be seen that the object highlighted by the green box is a distant building roof, which should not be recorded in the depth map (given a maximum depth of 80m). Only D4RD can discern such details accurately. **(i)** is real rainy scene. Weatherdepth and Robust-Depth* identified lens blurring as infinity, while EC-Depth* did not recognize the sparse areas.

Table 1: The meanings of meta symbols. Because the same character can represent multiple meanings, please pay attention to font changes.

Symbol	Meaning
a	The auxiliary image from adjacent frame
α	A transformation of β
aug	The augmented image
β	Predefined variance noise schedule
c	The diffusion's condition
clr	The clear image
cst	Contrast (a learning method)
d	Pixel level depth
d	Map level depth prediction
D	Map level depth distribution (from GT)
dis	Distillation (a learning method)
e	The edges
ϵ	A noise that follows $\mathcal{N}(0, \mathrm{I})$ distribution
f, feat	The Depth feature
$\mathcal{F}$	A neural network, the concrete meaning is determined by its index
gt	The ground truth depth
I, img	The RGB images
L	The training loss
M	Adaptive distillation mask
$\mathcal{M}$	The valid mask of GT depth
$\mathcal{N}$	Gaussian sampling
nis	The noise
ph	Photometric consistency
S	Student model
T	Teacher model
$\mathcal{T}$	Relative camera poses transformation
$\mathbb{T}$	Overall diffusion training steps
t	The target image (The image to be estimated depth)
τ	The time step
$\mathcal{U}$	Uniformly sampling
λ	A constant, relative to adaptive mask
ω, η, ρ	Weights for various losses

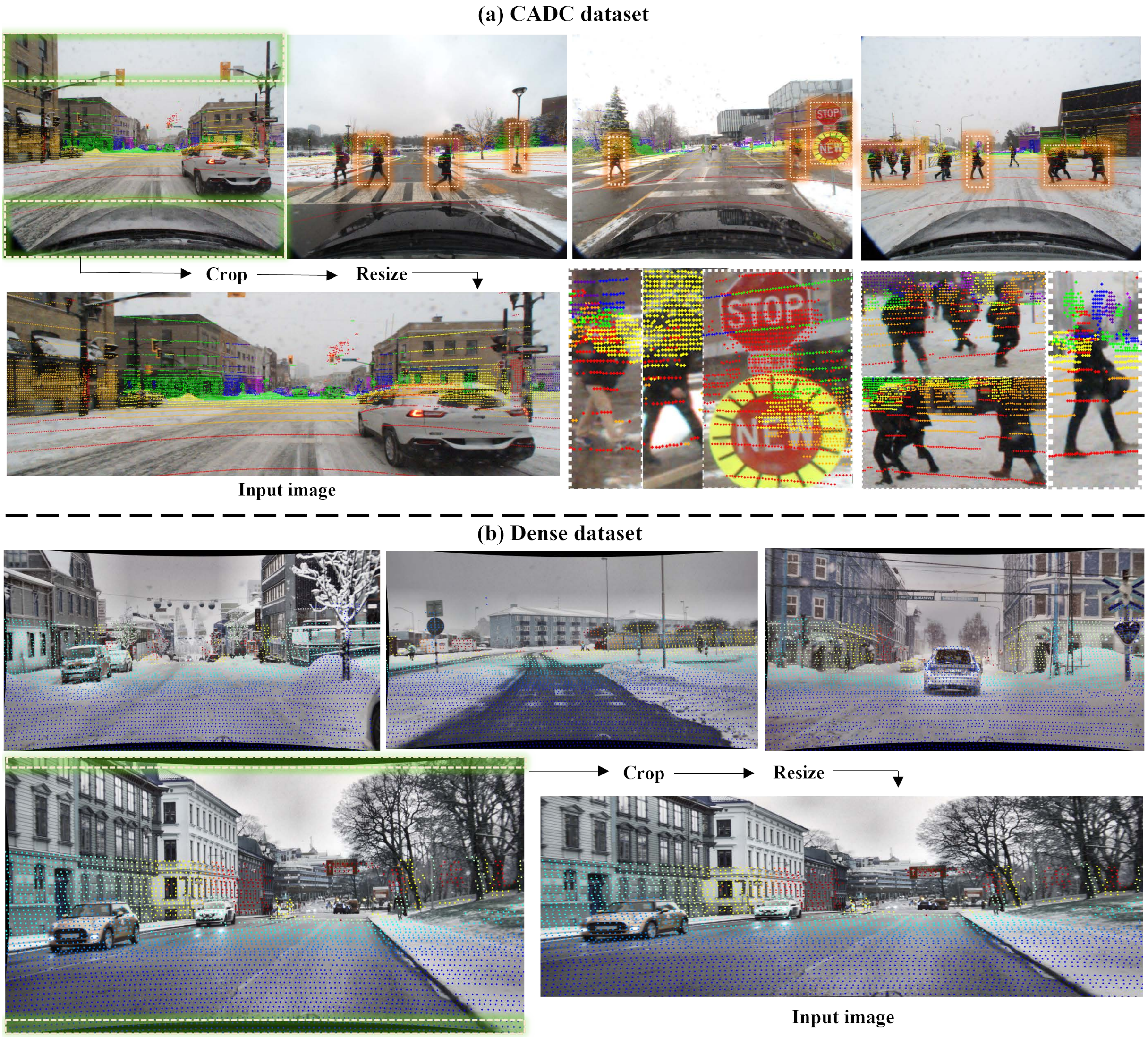

Figure 1: The LiDAR points projection on the CADC dataset and Dense dataset. The cropped areas and prominent errors are highlighted with green and orange boxes, respectively. For the CADC dataset, the colors from near to far are red, orange, yellow, green, blue, and purple in order. And for Dense, it is blue, light blue, yellow, orange, and red.

In contrast, D4RD successfully processed both challenges. **(j)** is real snowy scene. Once again, the distant cars and fences are accurately identified by D4RD. **(k)** is real cloudy scene. From the zoomed image, it can be seen that the building should be located behind a distant pole, which exceeds the range of 80m. However, D4RD can accurately identify this case. **(l)** is from a real sunny scene, specifically focusing on the unique condition of the bridge cave. Affected by this out-of-distribution environment, the depth estimation performance of other RMDE models was significantly impacted. Conversely, D4RD demonstrates robustness and maintains accurate depth estimation. **(m)** is real fog scene. From the zoomed image, it can be seen that the correct structural relationship of the highlighted area is 'a pole in front of the distant building'. Other models either estimate the depth of the building too closely or estimate the depth of the poles too far. Only D4RD is competent for this job.

Based on the above discussion, we further prove the robustness and effectiveness of D4RD.

REFERENCES

[1] Jiyuan Wang, Chunyu Lin, Lang Nie, Shujun Huang, Yao Zhao, Xing Pan, and Rui Ai. 2023. WeatherDepth: Curriculum Contrastive Learning for Self-Supervised Depth Estimation under Adverse Weather Conditions. *ArXiv* abs/2310.05556 (2023). https://api.semanticscholar.org/CorpusID:263831385

[2] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. 2022. MonoViT: Self-Supervised Monocular Depth Estimation with a Vision Transformer. In *2022 International Conference on 3D Vision (3DV)*. IEEE. https://doi.org/10.1109/3dv57658.2022.00077

[3] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. 2017. Unsupervised Learning of Depth and Ego-Motion from Video. arXiv:1704.07813 [cs.CV]

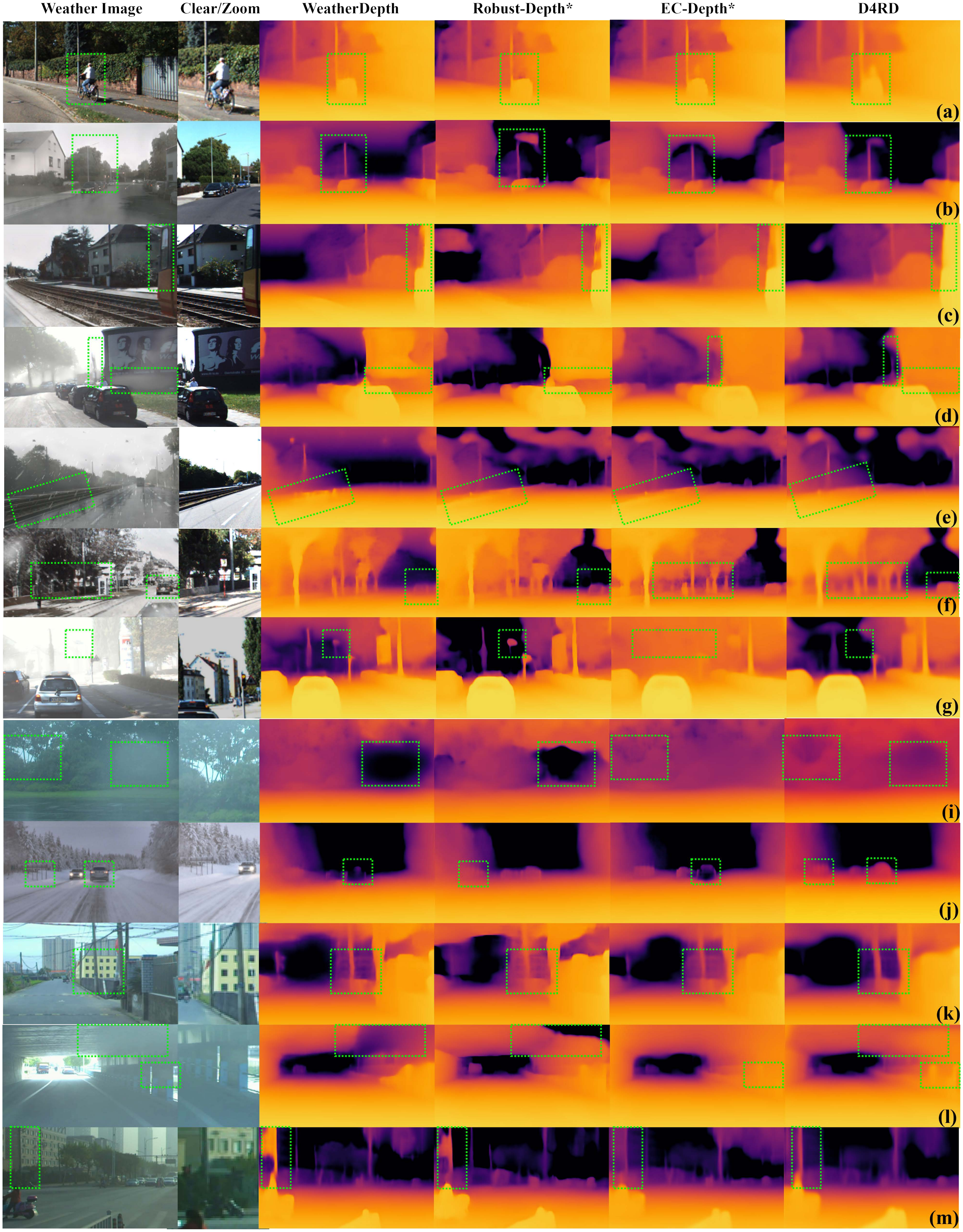

Figure 2: Extra visual comparisons on synthesis and real weather dataset. It is better viewed when zooming in. (a)-(g) are the WeatherKITTI dataset results and (i)-(m) are the real weathers.