

# RobuSTereo: Robust Zero-shot Stereo Matching under Adverse Weather

## Supplementary Material

To make our proposed method, **RobuSTereo**, self-contained, we provide more details in this document including: 1) dataset details, 2) model performance under normal weather, 2) more qualitative comparison results and visualization, 3) ablations on consistency module and data sources, 4) applications on visual SLAM, and 5) discussion on existing stereo datasets under adverse weather. We also provide videos and pointcloud visualization of autonomous driving scene, which shows that our method has more stable and accurate prediction results.

### A. Dataset Details

**SceneFlow** [7] is a widely used synthetic dataset for stereo matching. It contains approximately 39,000 stereo image pairs at a resolution of  $960 \times 540$  pixels, rendered from diverse synthetic indoor and outdoor scenes. In our experiments, all methods are initially trained on SceneFlow dataset [7] and subsequently fine-tuned on other datasets.

**KITTI 2012** [4] and **KITTI 2015** [8] are widely used stereo matching benchmarks under normal scenarios, each with 200 labeled training pairs and 200 testing pairs. The ground truth is derived from sparse LiDAR point clouds. In this work, we use the KITTI datasets to generate stereo images under adverse weather using our method, and they also serve as one of our comparison benchmarks.

**Virtual KITTI V2** [2], also known as vKITTI, consists of five sequence clones from the KITTI benchmark. The dataset includes various simulated weather conditions (e.g., fog, rain) and provides multiple data types for each sequence. vKITTI [2] represents a recent advancement in stereo matching simulation datasets for autonomous driving under adverse weathers. In this work, we also use vKITTI [2] as a training dataset for comparison.

**DrivingStereo** [10], introduced in 2019, contains 174,437 training pairs and 7,751 testing pairs, with an average image resolution of  $881 \times 400$ . Additionally, 2,000 frames with four different weather conditions (sunny, cloudy, foggy, rainy) are selected. Given its extensive data across various weather scenarios, we use DrivingStereo [10] as a test set and evaluate the performance of all models across different weather conditions for a detailed comparison.

**SeeingThroughFog** [1] provides a comprehensive benchmark for stereo matching under challenging adverse weather conditions. It includes data collected over 10,000 km and 12,000 samples of driving in northern Europe, covering diverse weather conditions such as fog, snow, and rain, with accurate Lidar point clouds. Given its extensive coverage of adverse weather scenes, we utilize this dataset

for both quantitative evaluations and visualization.

### B. Model Performance under Normal Weather.

Models trained on RobuSTereo demonstrate remarkable adaptability and do not exhibit significant accuracy degradation in normal scenarios. This robustness is attributed to the incorporation of prompts for normal scenarios, such as "cloudy," during the dataset generation process. These prompts ensure that the dataset encompasses a balanced variety of typical weather conditions, alongside adverse ones, enabling the model to maintain performance across diverse environments.

As shown in sTable 1, our method achieves comparable results to those on KITTI under "cloudy" and "sunny" weather conditions in the DrivingStereo dataset, which are representative of normal scenarios. Specifically, the performance gap between RobuSTereo-trained models and models trained on standard datasets is minimal, indicating that the inclusion of adverse weather data does not compromise the model's ability to handle regular conditions effectively.

Furthermore, we conducted fine-tuning experiments on KITTI [8] benchmark using IGEV [9] training on RST-datsset to validate the generalization capability of our dataset. The results confirm that training on our dataset does not significantly degrade performance in normal weather conditions, such as those encountered in KITTI, which predominantly features clear or mild weather scenarios. Instead, the fine-tuned models retain high accuracy, showcasing the robustness of our dataset design and its ability to support stereo matching tasks across both adverse and normal conditions. This balance makes RobuSTereo highly practical for real-world applications where models are required to perform reliably under a wide range of environmental conditions.

### C. More Comparison with Other Datasets

As illustrated in Figure 1, we present stereo prediction results under various adverse weather conditions, comparing our method (RobuSTereo) with StereoBase-SF [7], StereoBase-KITTI [8], and StereoAnything [6]. The visual results demonstrate that our method consistently outperforms the baselines across all tested scenarios, including rain, snow, fog, and low-light conditions. Existing methods, such as StereoBase-SF and StereoBase-KITTI, exhibit significant artifacts, missing disparities, and poor performance in occluded or low-visibility regions. StereoAnything performs better overall but struggles with fine-grained details and maintaining consistency under extreme weather condi-

Training Set	KITTI15			DS-sunny		DS-cloudy	
	D1-bg ↓	D1-fg ↓	D1-all ↓	EPE ↓	D1 ↓	EPE ↓	D1 ↓
KITTI	1.44	<b>2.31</b>	<b>1.59</b>	0.840	1.947	0.820	1.904
RST-Dataset	<b>1.43</b>	2.67	1.64	<b>0.834</b>	<b>1.834</b>	<b>0.797</b>	<b>1.764</b>

**Table 1.** Evaluation results on normal weather scenarios. Our method is called RobuSTereo (IGEV) on the benchmark website.

tions. In contrast, RobuSTereo effectively handles adverse weather effects, preserving fine object boundaries, maintaining depth continuity, and reconstructing occluded areas with high accuracy. These results highlight the robustness and adaptability of our approach, making it better suited for real-world applications in challenging environments.

## D. Ablations

In this section, we present visualization results from the ablation experiments, which include the analysis of the consistency module and comparative experiments involving different data sources.

### D.1. Ablation on Different Data Source

The comparison between SceneFlow [7] and vKITTI [2] data source highlights significant differences in image details and real texture. KITTI images, captured from real-world scenes, exhibit a high level of complexity and natural imperfections. For instance, in the “Cloudy city with overcast sky” scenario, KITTI shows intricate building textures, uneven lighting transitions, and natural variations in tree leaves, which are missing in the smoother, more uniform vKITTI counterparts. Similarly, in the “Foggy day with low visibility” scenario, KITTI captures the dynamic scattering of fog with varying densities, creating a more realistic atmospheric effect, whereas vKITTI applies synthetic fog in a uniform manner that lacks depth and randomness. The “Rainy day, wet pavement” scene in KITTI further demonstrates superior realism, with reflections on the wet road exhibiting complex interactions with vehicles and surrounding objects, compared to the simpler and less detailed reflections in vKITTI. Even in the “Light snowfall, gentle flakes” scenario, KITTI captures the irregular distribution of snowflakes and their interaction with the environment, while vKITTI renders the scene with consistent and predictable patterns. Across all scenarios, KITTI objects like vehicles and road features are seamlessly integrated into the environment, with natural imperfections such as dirt or minor deformations, whereas vKITTI objects appear smooth, polished, and artificial. These differences indicate KITTI’s greater authenticity and suitability for tasks requiring real-world generalization, while the images generated using vKITTI have poor authenticity and detailed texture, which affects the subsequent training results.

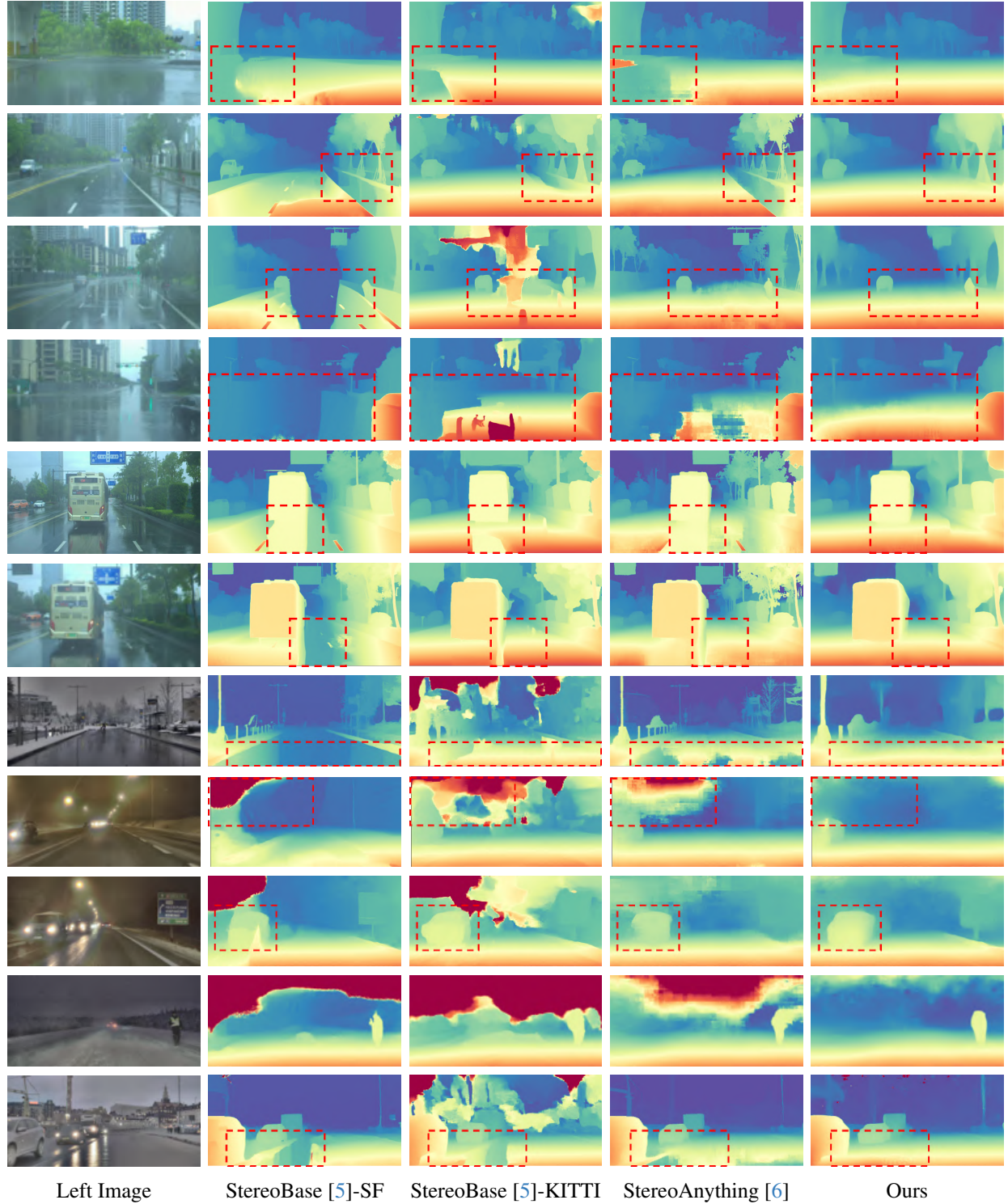
### D.2. Ablation on Consistency Module

The ablation study of the consistency module highlights its crucial role in generating coherent and realistic images. In the sunny road scene (first row), images without the consistency module show fragmented edges and inconsistent lighting, while our method delivers smooth textures and natural light transitions. In the wet urban scene (second row), reflections appear distorted without the module, whereas our method produces sharp, realistic reflections. In the foggy setting (third row), uneven fog intensity and abrupt transitions occur without the module, while our method achieves uniform fog integration. Finally, in the snowy environment (last row), snow coverage without the module is patchy, while our method ensures even distribution and smooth transitions. Overall, the consistency module enhances sharpness, stereo coherence, and environmental realism.

## E. Experimental Results on SLAM

SLAM (Simultaneous Localization and Mapping) is a critical downstream application of stereo depth estimation, where accuracy and robustness in depth predictions directly influence the quality of SLAM reconstruction. Comparative experiments across datasets highlight the limitations of SceneFlow [7] and KITTI [8] in adverse weather conditions, as shown in the magnified **brown** box in Figure 3. Both SceneFlow [7] and KITTI [8] introduce significant depth estimation errors, leading to incorrect SLAM reconstruction of intersections, particularly in challenging scenarios such as foggy or rainy weather. These errors compromise the reliability of SLAM-based systems.

In contrast, models trained on our dataset demonstrate remarkable robustness, accurately reconstructing intersections even under adverse weather conditions. Additionally, as highlighted in the magnified **green** box, our method produces the densest and most consistent depth predictions for road surfaces, enabling precise and stable SLAM reconstruction. This superior performance stems from the ability of our dataset to handle diverse environmental complexities, such as occlusions, uneven lighting, and weather-induced visual artifacts. The enhanced robustness and precision of our method make it particularly valuable for ensuring the reliability of autonomous driving systems in real-world, weather-affected scenarios.

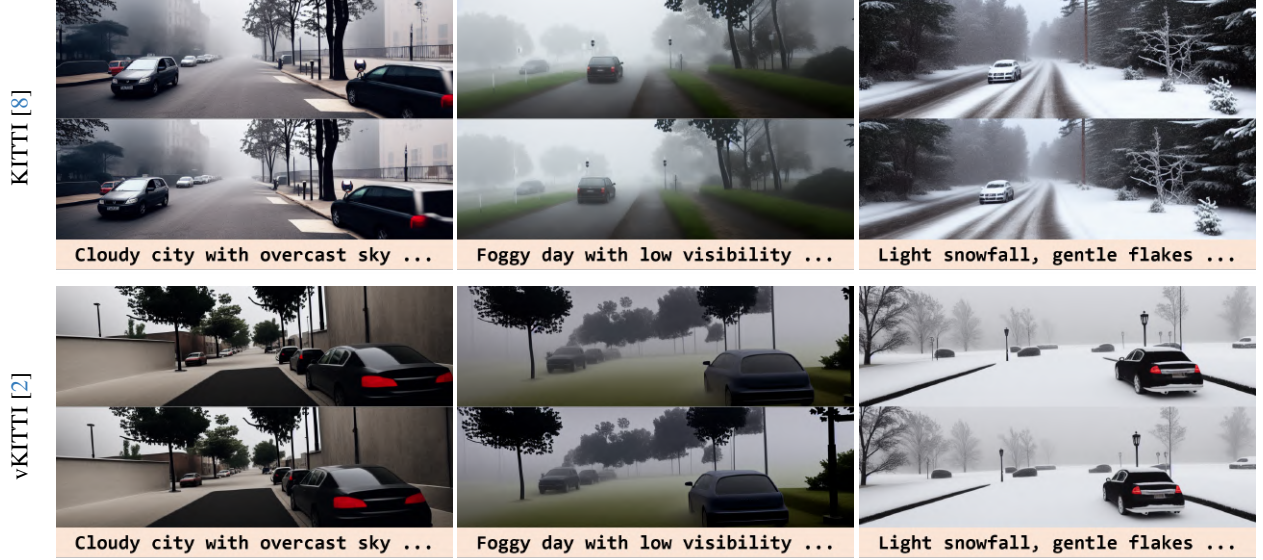


**Figure 1.** More comparison of the disparity predicted by different stereo matching models. StereoBase [5]-SF represents StereoBase model trains on SceneFlow [7], StereoBase [5]-KITTI represents StereoBase model trains on SceneFlow [8].

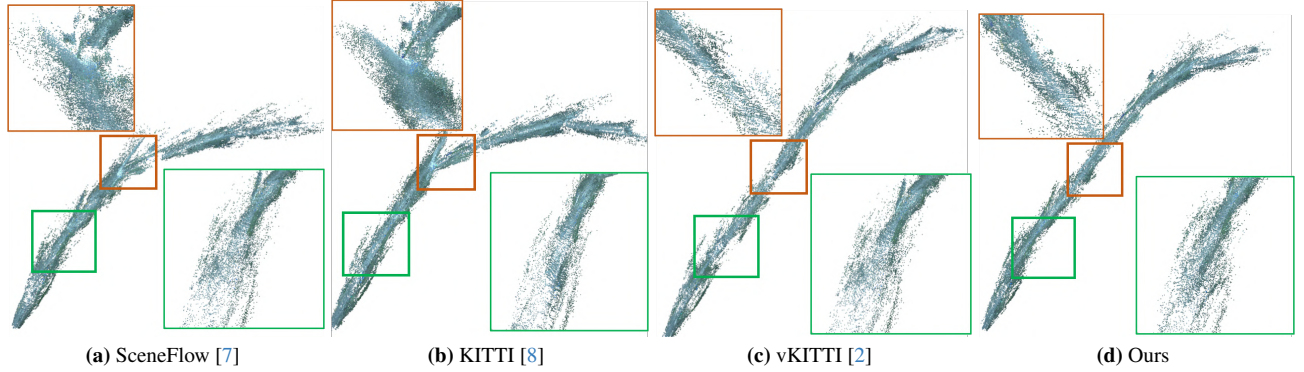
## F. Discussion on Existing Stereo Datasets under Adverse Weather

Several researchers [3, 10] have attempted to extend the scenarios covered by datasets by collecting real-world data un-





**Figure 2.** Comparison of the generated stereo images using KITTI [8] and vKITTI [2] as data source. Stereo images generated using vKITTI [2] often loses texture details in the images, compared to use KITTI [8] as datasource.



**Figure 3.** SLAM visualization results on DrivingStereo [10] dataset on rainy weather. We use different datasets to train StereoBase [5], and use the trained model to perform SLAM reconstruction on adverse weather scenes. The zoomed part shows that our method can achieve road reconstruction more stably, which is very important for autonomous driving.

der adverse weather conditions. However, as illustrated in Figure 5, limitations of LiDAR equipment result in datasets that either omit depth values in critical regions or produce excessively sparse data. These challenging regions are crucial for accurately modeling adverse weather scenarios, and the absence of such data significantly diminishes the practical utility of these datasets.

In contrast, our data generation method leverages depth data collected under normal conditions, ensuring denser and more accurate depth maps. Consequently, the data pairs generated by our method include dense ground truth (GT) data with minimal errors. Training stereo depth estimation models on our dataset enhances their robustness in adverse weather conditions. As shown in Figure 5, stereo models

trained on our dataset effectively address the limitations of LiDAR, particularly under challenging conditions.

## References

- [1] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 11682–11692, 2020. 1
- [2] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 1, 2, 4
- [3] Zaid A El-Shair, Abdalmalek Abu-raddaha, Aaron Cofield, Hisham Alawneh, Mohamed Aladem, Yazan Hamzeh, and Samir A Rawashdeh. Sid: Stereo image dataset for au-



**Figure 4.** Comparison of the generated stereo images with consistency module and without consistency module. Stereo images generation without consistency module often results in less consistent between left and right views.

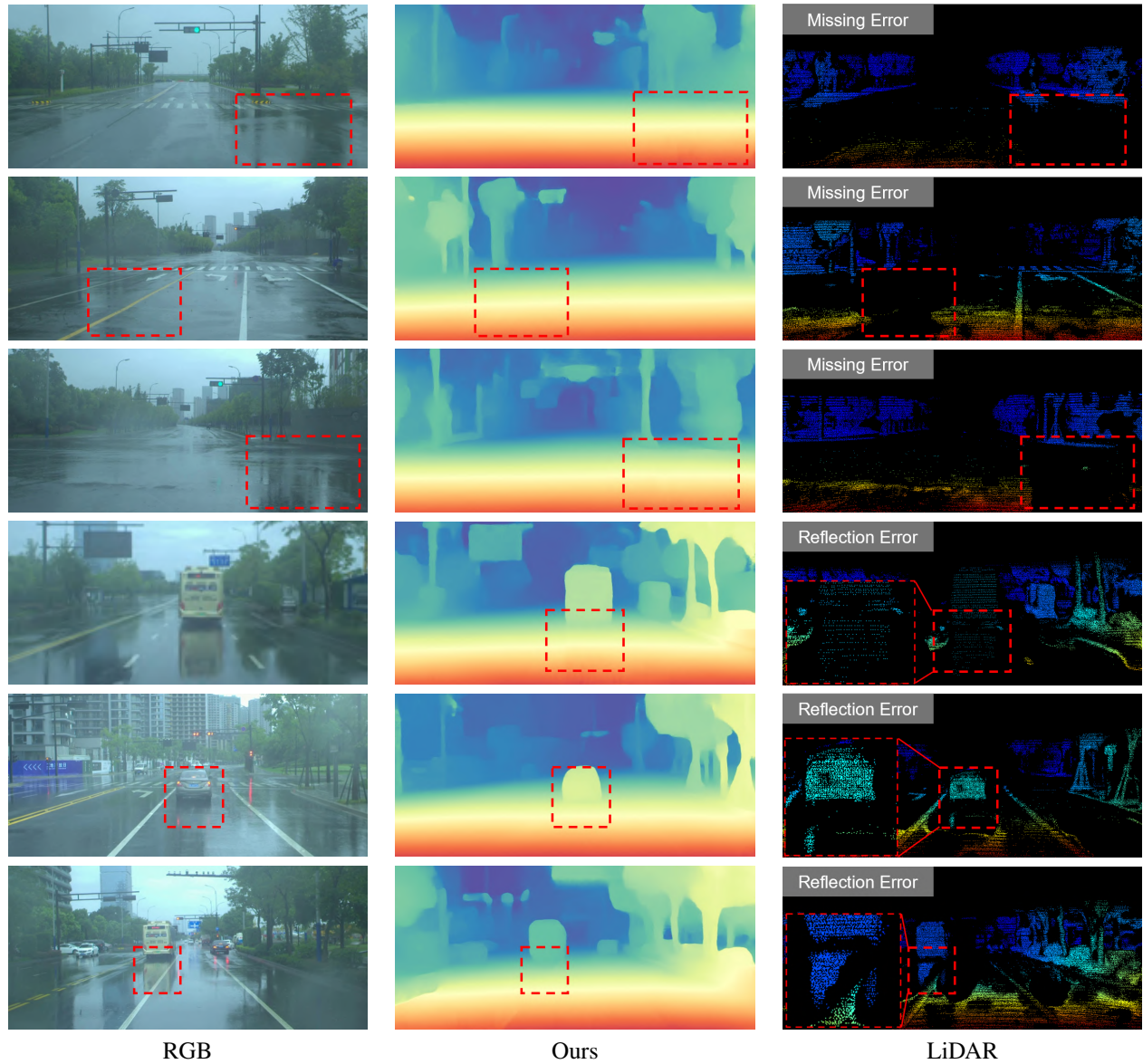
onomous driving in adverse conditions. In *NAECON 2024-IEEE National Aerospace and Electronics Conference*, pages 403–408. IEEE, 2024. 3

- [4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1
- [5] Xianda Guo, Juntao Lu, Chenming Zhang, Yiqi Wang, Yiqun Duan, Tian Yang, Zheng Zhu, and Long Chen. Openstereo: A comprehensive benchmark for stereo matching and strong baseline. *arXiv preprint*, 2023. 3, 4
- [6] Xianda Guo, Chenming Zhang, Youmin Zhang, Dujun Nie, Ruilin Wang, Wenzhao Zheng, Matteo Poggi, and Long Chen. Stereo anything: Unifying stereo matching with large-

scale mixed data. *arXiv preprint*, 2024. 1, 3

- [7] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016. 1, 2, 3, 4
- [8] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. 1, 2, 3, 4
- [9] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023.





**Figure 5.** Comparison of disparity maps predicted by the model and disparity maps collected by LiDAR in DrivingStereo [10]. **Black** pixel indicates invalid disparity pixels. Due to problems such as mirror reflection caused by adverse weather, LiDAR have obvious reflection errors. The method trained with our dataset can solve this problem.

1

- [10] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2019. 1, 3, 4, 6