Supplementary Material:
Deep Depth Estimation From Thermal Image

Ukcheol Shin
KAIST
shinwc159@gmail.com

Jinsun Park
Pusan National University
jspark@pusan.ac.kr

In So Kweon
KAIST
iskweon77@kaist.ac.kr

1. Multi-Spectral Stereo (MS^2) Dataset

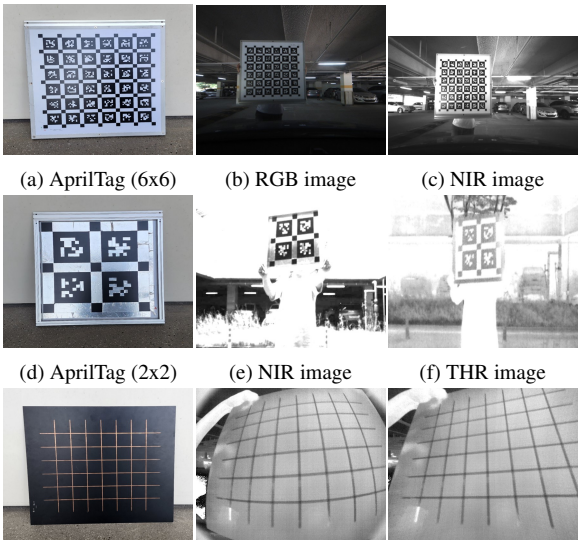
1.1. Sensor System Configuration

RGB stereo, NIR stereo, and GNSS/IMU sensor were installed inside the vehicle to ensure safe and reliable operation under adverse weather conditions such as rain, snow, fog, and haze, as shown in Fig. 2 of the main paper. A thermal camera cannot see through a glass, so we cannot install a thermal camera inside the vehicle. Therefore, LiDAR stereo and thermal stereo were built in outside of the vehicle. LiDARs are water-proof, and thermal cameras are covered with water-proof housing.

1.2. Calibration

We provide intrinsic and extrinsic parameters of all sensors built into our system to make our dataset applicable to various computer vision tasks. As shown in Fig. 1, we utilize 6x6 AprilTag [6] board for stereo RGB, stereo NIR, RGB-NIR, NIR-IMU, and NIR-LiDAR calibrations [4, 10]. Also, we utilize copper-coated line-board to estimate intrinsic matrices, radial distortion parameters, and extrinsic matrix of stereo thermal cameras. After that, we utilize a 2x2 AprilTag board with metallic tape attached to estimate an extrinsic matrix between NIR and thermal cameras. Before pattern board image acquisitions, both 2x2 board and line board were cooled down to obtain better thermal image contrast in the metallic and non-metallic regions.

The original RGB, NIR, and thermal image contain not necessary regions for various vision applications, such as car hood and sky area. Also, the projected LiDAR's depth points do not appear in the sky area and are invalid in the car hood. Therefore, after the calibration process, we rectified and cropped the original RGB, NIR, and thermal images to remain valid areas only, as shown in Fig. 2. After rectification and cropping, RGB, NIR, and thermal images provide 1224 x 384, 1280 x 352, and 640 x 256 spatial resolution, respectively.



(a) AprilTag (6x6) (b) RGB image (c) NIR image
(d) AprilTag (2x2) (e) NIR image (f) THR image
(g) Line-board (h) THR image (i) Rectified image

Figure 1. **Calibration pattern board for multi-sensor calibration**. We utilize a 6x6 AprilTag [6] board for stereo RGB, stereo NIR, RGB-NIR, NIR-Lidar, and NIR-IMU calibration. Also, a 2x2 AprialTage board is used to estimate the extrinsic matrix between NIR and thermal camera. We use a copper-coated line board for stereo thermal camera calibration.

(a) Original Images (RGB/NIR/THR)
(b) Rectified&Cropped Images (RGB/NIR/THR)

Figure 2. **Multi-spectrum images for MS^2 depth dataset**. We rectified and cropped the original RGB, NIR, and thermal images based on intrinsic and distortion parameters to make proper training data for MS^2 depth dataset. We cropped invalid areas such as the vehicle's hood and sky regions.

1.3. Sequence List

Multi-Spectral Stereo (MS^2) Dataset also provides 24 RAW ROSbag files that are taken under various locations (*e.g.*, city, campus, residential, road, and suburban), times (*e.g.*, morning, daytime, and nighttime), and weather conditions (*e.g.*, clear-sky, cloudy, and rainy), as listed in Tab. 1. Each ROSbag contains raw RGB stereo, NIR stereo, thermal stereo, LiDAR stereo, and GNSS/IMU data stream. Also, we can observe diverse characteristics of each sensor under each location, time, and weather condition.

Locations. The location "Campus" provides a moderate driving scenario with a moderate number of vehicles, bicycles, motorcycles, and riders. On the other hand, the location "City" provides a complex traffic situation with lots of dynamic objects (*e.g.*, numerous vehicles and pedestrians). Also, pedestrians and vehicles suddenly appear in the "Residential" scenario because of the sidewalk and road combined conditions and numerous closely located buildings and objects. The location "Suburban" usually contains only a few vehicles and provides a clean driving scenario. In the "Road" location, lots of fast-moving vehicles appear with a few buildings.

Lighting Condition. Also, as shown in Fig. 3, each sensor shows different aspects depending on lighting and weather conditions. RGB images can provide detailed object textures, color, and sharp structure information. However, as the lighting condition gets worse, this information is limited, saturated, and blurred. On the other hand, NIR images provide relatively better image quality in low-light conditions (*e.g.*, Night condition of "Road1" and "Road2"). But, the NIR spectrum is more sensitive to the light source, and NIR images are easily saturated in car head-light and street lamp regions. Also, if there are not enough external lighting sources, the quality also decreases (*e.g.*, Night condition of "Suburban" scenario). Thermal image shows lighting condition agnostic property since the principle of thermal imagining is not relevant to the lighting source.

Weather Condition. Under the daytime condition with clear-sky, RGB and NIR images show clear image quality with high contrast and details. Also, the thermal image shows high contrast compared to nighttime and rainy conditions. Because the sun acts as an external heat source to provide high thermal radiation values for all objects in daytime conditions with clear-sky. Each object can have more thermal radiation energy depending on heat absorption and reflectance ratios. Therefore, a thermal camera can acquire high-contrast images based on distinguishable thermal radiation values of each object (*e.g.*, Day and Night image of "Road2"). However, if there is no strong heat source or lots of clouds (*i.e.*, Night and Cloudy), the image contrast decreases. In rainy conditions, RGB, NIR, and thermal images suffer from light and heat reflections caused by wet roads.

Table 1. **Sequence list of MS^2 dataset.** The MS^2 dataset provides 24 RAW ROSbag files [5] taken under various locations, times, and weather conditions. Each ROSbag contains raw RGB stereo, NIR stereo, thermal stereo, LiDAR stereo, and GNSS/IMU data stream. We can observe diverse sensor characteristics according to the combination of time, weather, and location.

Index	ROSbag Name	Time	Weather	Loc	Duration
1	2021-08-06-10-59-33	Morning	Clear-sky	Campus	1071.0 sec
2	2021-08-06-17-44-55	Daytime	Cloudy&Rainy	Campus	1100.9 sec
3	2021-08-13-17-06-04	Daytime	Clear-sky	Campus	984.0 sec
4	2021-08-13-21-18-04	Nightime	Clear-sky	Campus	1040.0 sec
5	2021-08-06-11-23-45	Morning	Clear-sky	City	1118.9 sec
6	2021-08-06-16-19-00	Daytime	Cloudy&Rainy	City	1218.8 sec
7	2021-08-13-15-46-56	Daytime	Clear-sky	City	1201.0 sec
8	2021-08-13-21-36-10	Nightime	Clear-sky	City	1212.3 sec
9	2021-08-06-11-37-46	Morning	Clear-sky	Residential	599.9 sec
10	2021-08-06-16-45-28	Daytime	Cloudy&Rainy	Residential	665.5 sec
11	2021-08-13-16-14-48	Daytime	Clear-sky	Residential	929.6 sec
12	2021-08-13-22-03-03	Nightime	Clear-sky	Residential	773.9 sec
13	2021-08-06-12-06-20	Morning	Clear-sky	Suburban	632.73 sec
14	2021-08-06-17-10-27	Daytime	Cloudy&Rainy	Suburban	684.7 sec
15	2021-08-13-16-41-00	Daytime	Clear-sky	Suburban	579.06 sec
16	2021-08-13-22-27-31	Nightime	Clear-sky	Suburban	544.9 sec
17	2021-08-06-16-59-13	Daytime	Cloudy&Rainy	Road1	1177.9 sec
18	2021-08-13-16-31-10	Daytime	Clear-sky	Road1	579.65 sec
19	2021-08-13-22-16-02	Nightime	Clear-sky	Road1	543.47 sec
20	2021-08-06-17-21-04	Daytime	Cloudy&Rainy	Road2	614.2 sec
21	2021-08-13-16-50-57	Daytime	Clear-sky	Road2	883.8 sec
22	2021-08-13-22-36-41	Nightime	Clear-sky	Road2	434.6 sec
23	2021-08-13-16-08-46	Daytime	Clear-sky	Road3	259.9 sec
24	2021-08-13-21-58-13	Nightime	Clear-sky	Road3	261.3 sec

1.4. Multi-Spectral Stereo (MS^2) Depth Dataset

Training Set Split. From the MS^2 dataset, we periodically sampled the thermal images and filtered out the static vehicle movement to make training, validation, and evaluation splits for the learning of monocular and stereo depth networks. We utilize "Road2", "Suburban", some of "City" (*i.e.*, index 5,8), and "Campus" (*i.e.*, index 1,2,3) as a training set. All "Residential", "Road1", and "Road3" are used for validation and testing sets. Also, some of "City" (*i.e.*, index 6,7) and "Campus" (*i.e.*, index 4) that have different lighting and weather condition with training set sequences are used for validation and testing set. Depending on each lighting and weather conditions, we divide them into a daytime evaluation set (*i.e.*, index 9,7,18), nighttime evaluation set (*i.e.*, index 4,12,19), and rainy evaluation set (*i.e.*, index 6,10,17). The remaining sequences are used for the validation set. In total, we utilize 26K data pairs for training, 4K pairs for validation, and 5.8K, 6.8K, and 5.2K pairs for evaluation of daytime, nighttime, and rainy conditions.

(a) Driving Scenario - Campus (Day/Cloudy/Night)

(b) Driving Scenario - Residential (Day/Cloudy/Night)

(c) Driving Scenario - City (Day/Rain/Night)

(d) Driving Scenario - Suburban (Day/Rain/Night)

(c) Driving Scenario - Road1 (Day/Rain/Night)

(d) Driving Scenario - Road2 (Day/Rain/Night)

Figure 3. **Data examples of Multi-Spectral Stereo (MS^2) outdoor driving dataset.** The collected dataset provides about 195K synchronized data taken under locations of campus, city, residential area, road, and suburban with various time slots (morning, day, and night) and weather conditions (clear-sky, cloudy, and rainy)). For each block, three row indicates RGB, NIR, and thermal images, respectively. According to the surrounding conditions, each spectrum sensor shows different aspects, advantages, and disadvantages induced by their sensor characteristics).

2. Qualitative Results

We provide a qualitative comparison results of various Monocular Depth Estimation (MDE) and Stereo Depth Estimation (SDE) networks, as shown in Fig. 4 and Fig. 5. RGB and NIR images are visualized as reference images to see lighting and weather condition. We visualize the inverse depth map of MDE networks and the disparity map of SDE networks. Generally, SDE networks (*e.g.*, GwcNet [3], AANet [8], ACVNet [7], Ours) show more accurate and clean disparity maps. Also, MDE networks (*e.g.*, AdaBins [1], NeWCRF [9]) are easily affected and overfitted by the GT disparity maps. However, MDE networks have the advantage to estimate depth maps of thin objects and par objects. Our proposed network leverages both advantages of MDE and SDE networks based on a monocular and stereo depth unification with conditional random field perspective.

3. Limitation&Future Plan

In this paper, we provide Multi-Spectral Stereo (MS^2) Dataset, including stereo RGB, stereo NIR, stereo thermal, stereo LiDAR data along with GNSS/IMU data. However, currently, we only provide Ground-Truth (GT) depth labels of monocular and stereo depth estimation tasks for thermal images. We plan to make GT depth labels for RGB and NIR images to investigate the possibility of depth estimation from multi-sensor under various conditions. Also, for the robust visual perception of a self-driving car, we plan to annotate object detection and segmentation labels. We keep collecting lots of driving scenarios to include more diverse locations and seasons in our dataset. We hope our dataset encourages active research of various computer vision algorithms from multi-spectral data to achieve high-level performance, reliability, and robustness against challenging environments.

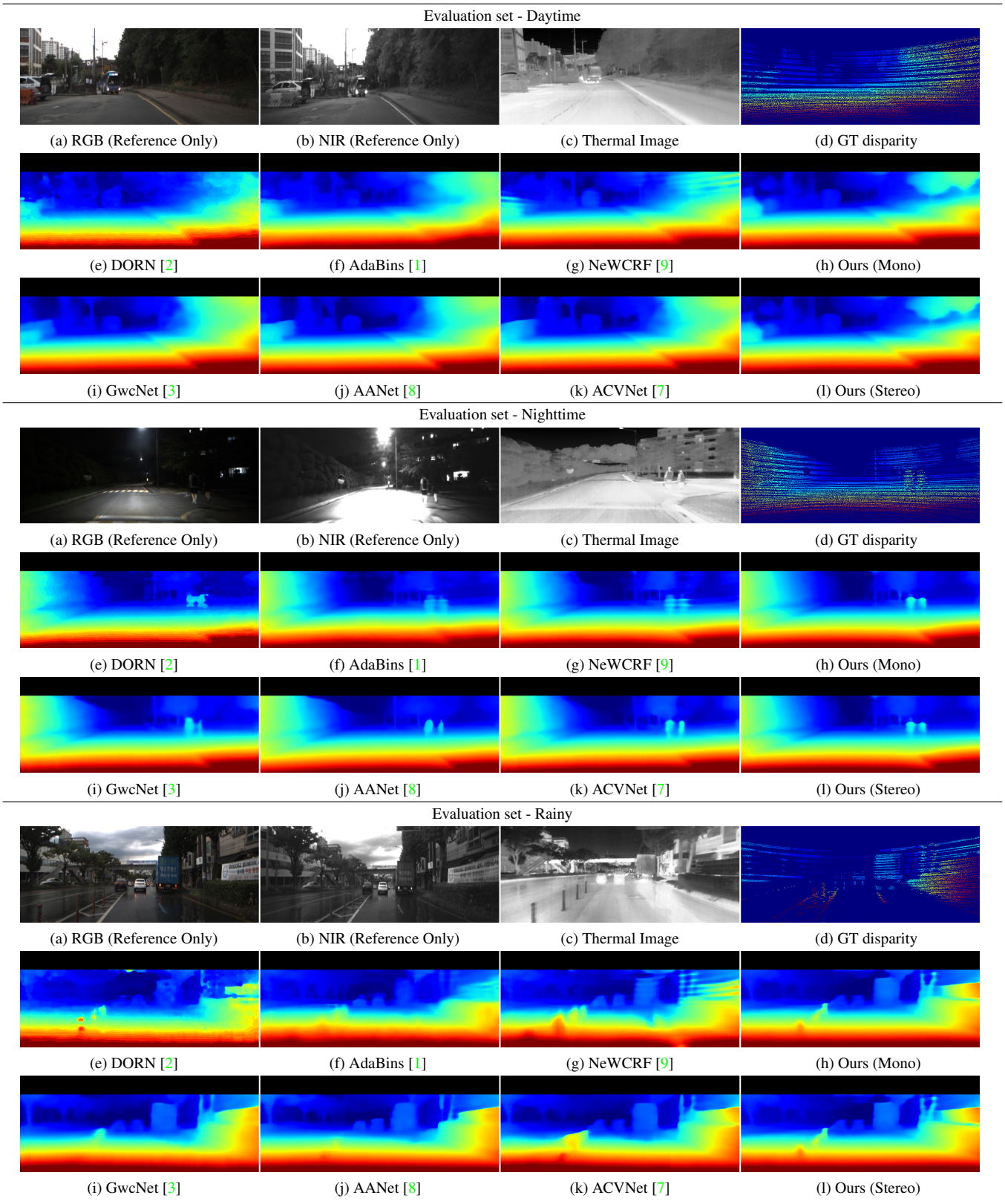


Figure 4. **Qualitative comparison of inverse depth and disparity maps on the MS^2 depth dataset**.

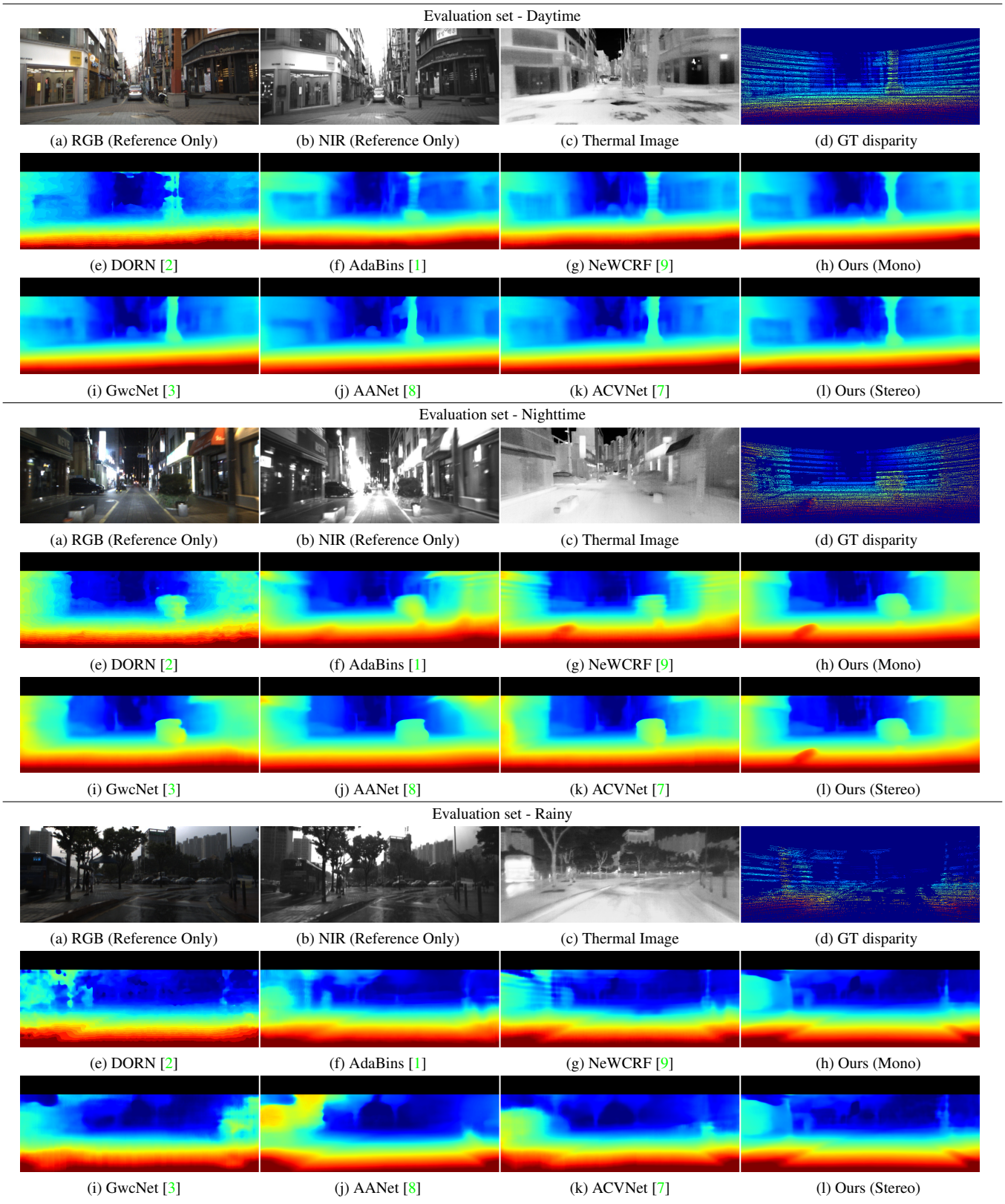


Figure 5. **Qualitative comparison of inverse depth and disparity maps on the MS^2 depth dataset**.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 3, 4, 5

[2] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 4, 5

[3] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. 3, 4, 5

[4] Joern Rehder, Janosch Nikolic, Thomas Schneider, Timo Hinzmann, and Roland Siegwart. Extending kalibr: Calibrating the extrinsics of multiple imus and of individual axes. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4304–4311. IEEE, 2016. 1

[5] Stanford Artificial Intelligence Laboratory et al. Robotic operating system. 2

[6] John Wang and Edwin Olson. Apriltag 2: Efficient and robust fiducial detection. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4193–4198. IEEE, 2016. 1

[7] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12981–12990, 2022. 3, 4, 5

[8] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 3, 4, 5

[9] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3916–3925, 2022. 3, 4, 5

[10] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000. 1