Empirical Study: Monocular Depth Estimation from RGB, NIR, Thermal Image in Adverse Weather Conditions

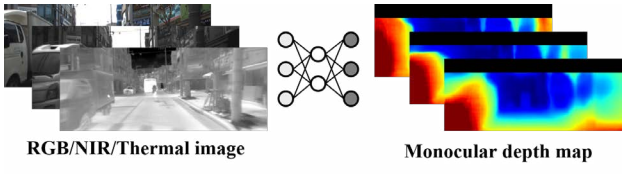
Ukcheol Shin
Carnegie Mellon University
shinwc159@gmail.com

Soonmin Hwang
Hanyang University
jjang9hsm@gmail.com

Jean Oh
Carnegie Mellon University
jeanoh@cmu.edu

Abstract

*Robust spatial understanding is one fundamental condition for safety-aware autonomous driving against adverse weather and lighting conditions, such as rain, fog, haze, snow, and low-light environments. Therefore, numerous autonomous vehicle platforms adopt various sensor modalities to ensure safety and reliability (*e.g.*, RGB camera, NIR camera, thermal camera, LiDAR, and RADAR). Among them, the RGB camera is a commonly adopted complementary sensor because it can provide dense spatial understanding ability compared to LiDAR and RADAR. However, RGB camera is known to be vulnerable to changes in lighting and weather conditions. In this paper, we empirically analyze the robustness of monocular depth estimation from RGB image in diverse seasonal, weather, and lighting conditions. Also, we investigate the robustness of depth estimation from NIR and thermal images in the same condition to find which sensor is robust to environmental changes and capable of dense spatial understanding even in extreme conditions. As a result, we found thermal cameras can provide reliable and robust dense spatial understanding against diverse seasonal, weather, and lighting condition changes.*



(a) Monocular depth estimation from RGB, NIR, and thermal image

(b) Challenging real-world environments in autonomous driving scenario

Figure 1. **Empirical study of monocular depth estimation from RGB, NIR, and thermal image in diverse real-world environments**. In this paper, we investigate the robustness and reliability of a modern deep neural network for monocular depth estimation from RGB, NIR, and thermal images (a) in diverse challenging real-world environments, such as rainy, snowy, and low-lighted conditions (b).

1. Introduction

Modern advancements in artificial intelligence have led to remarkable performance improvement in various computer vision tasks including semantic perception [1–4] and spatial perception [5–10]. Upon these perception abilities, an agent can perform high-level tasks such as path planning, obstacle avoidance, object grasping, navigation, and autonomous driving. However, when an agent encounters challenging real-world environments such as low-lighting, rainy, smoky, and foggy conditions, its semantic and spatial perception ability loses functionality.

Therefore, for the safety-aware autonomous driving system against adverse weather and lighting conditions, numerous autonomous vehicle platforms adopt various sensor modalities to ensure safety and reliability (*e.g.*, RGB camera, NIR camera, thermal camera, LiDAR, and RADAR). Among them, the RGB camera is a commonly adopted complementary sensor because it can provide dense spatial understanding ability compared to LiDAR and RADAR. However, RGB image is easily degenerated by changes in lighting and weather conditions. The degeneration also affects the reliability of a deep neural network's prediction result.

In this paper, we empirically analyze the robustness of monocular depth estimation from RGB image in di-

verse seasonal, weather, and lighting conditions, as shown in Fig. 1. Also, we investigate the robustness of depth estimation from NIR and thermal images in the same condition to find which sensor is robust to environmental changes and capable of dense spatial understanding even in extreme conditions. As a result, we found thermal cameras can provide reliable and robust dense spatial understanding against diverse seasonal, weather, and lighting condition changes.

2. Implementation Details

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

We utilize a Multi-Spectral Stereo (MS^2) dataset [11] to evaluate a network's robustness. The dataset provides about 184K multi-spectrum data pairs of RGB, NIR, thermal camera, LiDAR, and GPS/IMU taken under various locations, times, and weather conditions. For the monocular depth estimation, we utilize an MS^2-summer training split for training, MS^2-summer validation split for validation, and MS^2-summer evaluation splits for evaluation of daytime, nighttime, and rainy conditions. Also, we utilize the MS^2-spring, fall, and winter evaluation split for the zero-shot evaluation. Please refer to the paper [11] for additional details.

2.2. Network Architecture

Monocular Depth Estimation (MDE) network is adopted from NeWCRF [10] to evaluate the robustness and reliability of a modern deep neural network. We utilize off-the-shelf network architecture from the official source code and don't modify any architecture details. For the NIR and thermal images that have only a single-channel, we repeat them three times along the channel axis to be identical to the RGB image. All MDE networks are initialized with ImageNet pre-trained backbone model [12] by following the original implementations [10].

2.3. Training Details

We utilize the PyTorch library [13] to implement our proposed method. All models are trained for 30 epochs on a single RTX A6000 GPU with 48GB memory. We utilize AdamW optimizer [14] with an initial learning rate of $1e^{-4}$ for all model training. For the data augmentation, we apply random center crop-and-resize, brightness jitter, and contrast jitter for all modalities. Saturation and hue jitters [15] are additionally applied to the RGB modality. Also, the horizontal flip is applied to all modality training. We utilized the depth evaluation metrics (*i.e.*, RMSE, δ) commonly used to measure the accuracy and error of depth estimation results [16–18].

3. Experimental Results

3.1. Depth Estimation from Modality RGB, NIR, and Thermal image

Firstly, we evaluate the depth estimation performance from RGB, NIR, and thermal images in an in-distribution scenario. The Monocular Depth Estimation (MDE) network [10] was trained in MS^2-summer training dataset and evaluated in MS^2-summer evaluation splits (day, night, and rainy). The training and evaluation data have no overlap but the data distribution can be similar because of the same seasonal property.

The evaluation results are shown in Tab. 1 and Fig. 2. The first-row in Fig. 2 shows that monocular depth maps from a thermal image outperform in all day, night, and rainy conditions compared to depth results from RGB and NIR images. RGB and NIR images are suffered from unclear visibility, blur effect, and occlusion by rain or windshield wipers in rainy conditions. Therefore, the quantitative results are generally lower than the thermal image-based prediction result. In day and night conditions, thanks to the light sensitivity of the NIR spectrum, depth from a NIR image can achieve the second-best results.

3.2. Zero-shot Evaluation : Summer-to-Season X

Secondly, we conduct zero-shot evaluations of the MDE network [10] in out-of-distribution conditions. The MDE network was trained in MS^2-summer training dataset and evaluated in MS^2-spring (day,night,rainy), MS^2-fall (day,night,rainy), and MS^2-winter (day,night,snowy). The comprehensive comparison results are shown in Tab. 2 and Fig. 2. We measure the performance degradation for different conditions with RMSE difference between the Baseline (*i.e.*, MS^2-summer day evaluation split) and each test set.

In general, depth from thermal image shows robust performance against various domain shifts including seasonal change, temperature change, weather change, and lighting condition change. Depth from NIR images achieves the second best in most scenarios. We found the color space of RGB camera introduces further domain shift according to the changes in season, weather, and lighting conditions. The NIR camera has better light sensitivity and single-channel intensity information that are not affected by color filters. Therefore, depth from NIR images shows generally better robustness and performance than RGB images. But, still, NIR camera is affected by a light source and lighting effect. Therefore, the performance and reliability are easily degraded by insufficient lighting conditions, lighting effects (glare and blur) caused by water particles, and occlusion caused by windshield wipers and water particles. On the other hand, thermal image has less affected by the lighting effect, water particles, and the presence of lighting source.

Table 1. **Quantitative comparison of depth estimation results on the MS^2 dataset (Summer) [11]**. We train and evaluate a state-of-the-art monocular depth estimation network [10] on the MS^2 dataset. Monocular depth maps from thermal image outperform in all day, night, and rainy conditions compared to depth results from RGB and NIR images. Depth from NIR image shows runner-up results. In rainy conditions, RGB and NIR images are suffered from unclear visibility, blur effect, and occlusion by rain or windshield wipers. The best performance in each block is highlighted in **bold**.

TestSet	RGB image		NIR image		Thermal image	
	RMSE (↓)	$\delta < 1.25$ (↑)	RMSE (↓)	$\delta < 1.25$ (↑)	RMSE (↓)	$\delta < 1.25$ (↑)
Summer (Clear+Day)	3.111	94.8	3.071	93.3	**2.717**	**95.1**
Summer (Clear+Night)	3.573	89.9	3.157	91.2	**2.544**	**95.2**
Summer (Rainy+Day)	4.447	87.0	5.042	81.0	**3.503**	**90.9**

Table 2. **Zero-shot evaluation (Spring, Fall, Winter): Quantitative comparison of monocular depth estimation results from RGB, NIR, and thermal images**. We evaluate the monocular depth network (*i.e.*, NeWCRF [10]) trained with MS^2 dataset (summer) [11] on the out-of-distribution season, weather, and lighting condition. We measure the performance degradation for different conditions with RMSE difference between the Base and each test set. The estimated depth maps from thermal images show robust performance against various domain shifts, including seasonal changes, temperature changes, weather changes, and lighting condition changes. Depth from NIR images also shows better performance than RGB images. The best performance in each block is highlighted in **bold**.

TestSet	RGB image			NIR image			Thermal image		
	RMSE (↓)	$\delta < 1.25$ (↑)	Δ RMSE (↓)	RMSE (↓)	$\delta < 1.25$ (↑)	Δ RMSE (↓)	RMSE (↓)	$\delta < 1.25$ (↑)	Δ RMSE (↓)
Base: Summer (Clear+Day)	3.111	94.8	-	3.071	93.3	-	2.717	95.1	-
Spring (Clear+Day)	5.473	70.0	-2.362	4.157	77.4	-1.086	**3.810**	**84.9**	**-1.093**
Spring (Rainy+Day)	5.599	68.9	-2.488	5.470	65.8	-2.399	**3.207**	**85.5**	**-0.490**
Spring (Rainy+Night)	7.282	57.8	-4.171	7.207	52.2	-4.136	**3.848**	**81.6**	**-1.131**
Fall (Clear+Day)	5.260	80.3	-2.149	**3.814**	**89.6**	**-0.743**	4.290	88.1	-1.573
Fall (Rainy+Night)	5.017	75.4	-1.906	3.532	83.8	**-0.461**	**3.271**	**88.1**	-0.554
Winter (Snowy+Day)	5.092	72.9	-1.981	4.740	74.5	-1.669	**3.640**	**83.2**	**-0.923**
Winter (Snowy+Night)	6.154	73.0	-3.043	4.585	83.8	-1.514	**3.362**	**91.1**	**-0.645**
Average	5.555	72.5	-2.444	4.613	77.0	-1.542	**3.567**	**84.9**	**-0.850**

4. Conclusion

This paper empirically analyzes the robustness of monocular depth estimation from RGB, NIR, and thermal images in diverse seasonal, weather, and lighting conditions. As a result, we found the depth estimation from RGB images shows vulnerable properties in the changes of seasonal, weather, and lighting conditions. Depth estimation from NIR images generally shows better reliability than RGB image in most cases, thanks to its light sensitivity and single-channel intensity information that are not affected by color filters. Lastly, the thermal radiation spectrum is rarely affected by lighting source, lighting effect, and water particles. Therefore, depth from thermal images shows the high-level robustness and performance against changes in seasonal, weather, lighting conditions, and domain shift. We hope the empirical study is helpful to the reader who wants to develop their autonomous vehicle platforms.

Acknowledgment

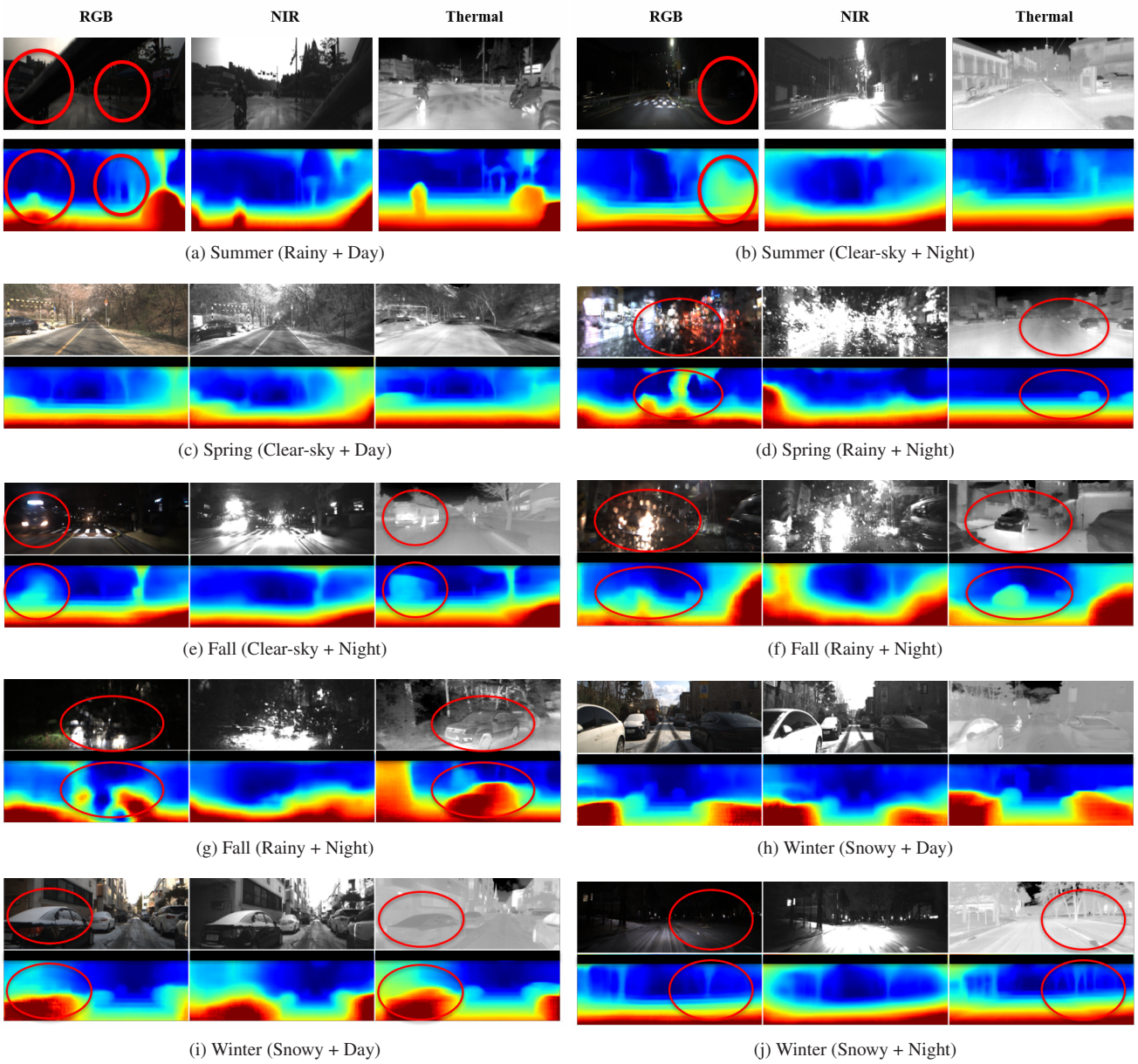
This work was supported by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute of Advancement of Technology (KIAT) through the International Cooperative R&D program: P0019782, Embedded AI Based fully autonomous driving software and Maas technology development.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.



(a) Summer (Rainy + Day)

(b) Summer (Clear-sky + Night)

(c) Spring (Clear-sky + Day)

(d) Spring (Rainy + Night)

(e) Fall (Clear-sky + Night)

(f) Fall (Rainy + Night)

(g) Fall (Rainy + Night)

(h) Winter (Snowy + Day)

(i) Winter (Snowy + Day)

(j) Winter (Snowy + Night)

Figure 2. **Quantitative comparison of monocular depth maps from RGB, NIR, and thermal image in diverse seasonal, weather, and lighting conditions.** In a clear-sky and well-lit condition, the estimated depth map shows similar prediction results regardless of input modality difference. However, for challenging weather and lighting conditions, a depth map estimated from thermal image generally provides high-level accuracy and robustness against various domain shifts, including seasonal, temperature, weather, and lighting condition changes, compared to RGB and NIR images. RGB and NIR images are suffered from unclear visibility, blur effect, and occlusion by rain or windshield wipers.

[4] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

[5] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022.

[6] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision*, 129(9):2548–2564, 2021.

[7] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7286–7291. IEEE, 2018.

[8] Fei Xue, Xin Wang, Shunkai Li, Qiuyuan Wang, Junqiu Wang, and Hongbin Zha. Beyond tracking: Selecting memory and refining poses for deep visual odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8575–8583, 2019.

[9] Nan Yang, Lukas von Stumberg, Rui Wang, and Daniel Cremers. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1281–1292, 2020.

[10] 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.

[11] Ukcheol Shin, Jinsun Park, and In So Kweon. Deep depth estimation from thermal image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1043–1053, 2023.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[16] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[17] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[18] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4796–4803. IEEE, 2018.