# Supplementary Material: Exploiting Cross-modal Cost Volume for Multi-sensor Depth Estimation

Janghyun Kim[1], Ukcheol Shin[2], Seokyong Heo[1], and Jinsun Park[3]*

[1] Department of Information Convergence Engineering, Pusan National University, Republic of Korea
[2] Robotics Institute, Carnegie Mellon University, United States
[3] School of Computer Science and Engineering, Pusan National University, Republic of Korea
{jangjoa41,hsdr915,jspark}@pusan.ac.kr   ushin@andrew.cmu.edu

## 1 Overview

In this supplementary material, we provide further ablation study about the model components on the MMDCE day [2] dataset to prove the effectiveness of our proposed approaches. In addition, we visualize our depth estimation results on the KITTI MMD [3, 2] and MMDCE day-night [2] datasets.

## 2 Effectiveness of the Proposed Framework

**Table S1. Ablation study of model components on the MMDCE day dataset.**

| Modality | $\ell_1$ | $\ell_2$ | CMA | Cross-spectral | $L_G$ | SPN | RMSE (mm) | MAE (mm) |
|---|---|---|---|---|---|---|---|---|
| RGB-NIR-LIDAR | Base fusion ($\ell_1 + \ell_2$) | | | | | | 1230.5 | 589.9 |
| | ✓ | ✓ | ✓ | | | | 1142.3 | 557.4 |
| | ✓ | ✓ | | ✓ | | | 1301.6 | 616.1 |
| | ✓ | ✓ | ✓ | ✓ | | | 1131.7 | 534.2 |
| | ✓ | ✓ | ✓ | ✓ | ✓ | | 1124.1 | 525.9 |
| | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | **1092.3** | 507.4 |
| | ✓ | | ✓ | ✓ | ✓ | ✓ | 1175.1 | **482.1** |
| | | ✓ | ✓ | ✓ | ✓ | ✓ | 1138.5 | 562.1 |

**Bold**: The best, Underline: The second-best

To provide further insights into the proposed method, we conducted additional ablation studies on the MMDCE day [2] dataset as shown in Tab. S1. Our proposed approaches consistently demonstrate efficacy in multi-modal depth estimation, even in the MMDCE day dataset. Moreover, we conducted experiments

* Corresponding author.

on the exclusion of $\ell_1$ and $\ell_2$ losses from our final combination components. When using only one of these losses, we observed a performance degradation in terms of RMSE. While not using $\ell_2$ showed performance improvements from 507.4 to 482.1 in the MAE metric, this configuration led to significantly lower accuracy in terms of RMSE. Therefore, we chose to utilize both losses to balance performance across the two metrics. These results indicate that using both $\ell_1$ and $\ell_2$ losses is a confident choice for optimizing the model's performance.

## 3 Visualization of Depth Estimation Results

We provide more qualitative comparisons on the KITTI MMD and MMDCE day-night datasets [2]. As depicted in our main manuscript, our multi-sensory depth estimation network demonstrates superior performance compared to previous approaches [1, 4, 2].

**KITTI MMD Dataset.** Our network effectively captures background and foreground areas in the KITTI MMD dataset, as shown in scenes 1 and 2 of Fig. S1. These scenes demonstrate that our network accurately identifies objects even at far distances, unlike the other networks. This capability is attributed to our method's diverse depth range searching. Additionally, our network precisely distinguishes small objects (*e.g.*, human head), as illustrated in scene 3.

**MMDCE Day Dataset.** Our network is not interrupted by reflective areas (*e.g.*, car windows) due to the proposed cost volume-guided propagation, as shown in scene 1. Moreover, our network produces sharper predictions without blur effects, as shown in scenes 2 and 3 of Fig. S2.

**MMDCE Night Dataset.** As shown in scene 1 of Fig. S3, our network preserves the shape of cars by utilizing only reliable cues through the cross-modal attention block for the cost volumes. Furthermore, our network can capture the tiny objects, as illustrated in scenes 2 and 3 of Fig. S3.

These results highlight the effectiveness of our proposed network. The superior performance is achieved through our method's ability to utilize dynamic depth ranges from short to long distances and facilitate depth regression using only reliable cues while suppressing redundant and irrelevant information.

**Fig. S1. Depth map comparisons on the KITTI MMD [3, 2] dataset.** (a) RGB, (b) Grayscale, (c) LiDAR, (d) LS [1], (e) CompletionFormer [4], (f) MMDNet [2], (g) Ours, and (h) GT.

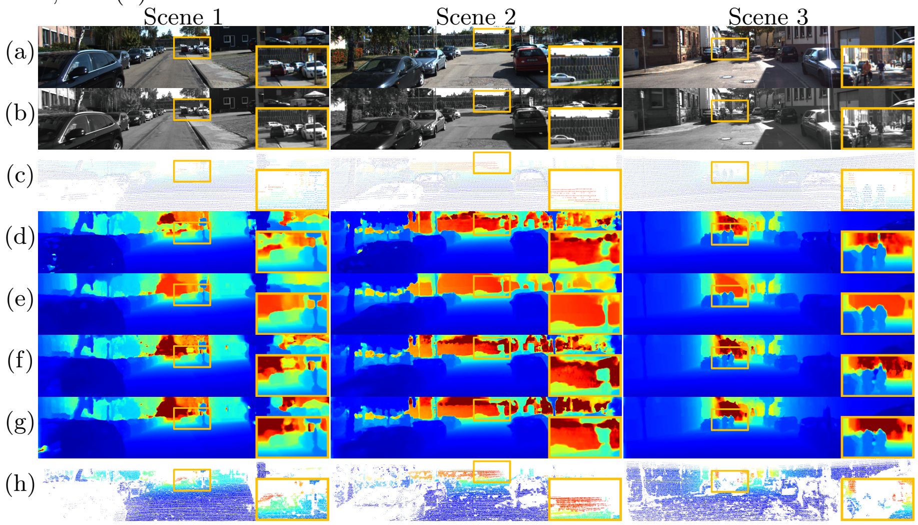


**Fig. S2. Depth map comparisons on the MMDCE day [2] dataset.** (a) RGB, (b) NIR, (c) LiDAR, (d) LS [1], (e) CompletionFormer [4], (f) MMDNet [2], (g) Ours, and (h) GT.

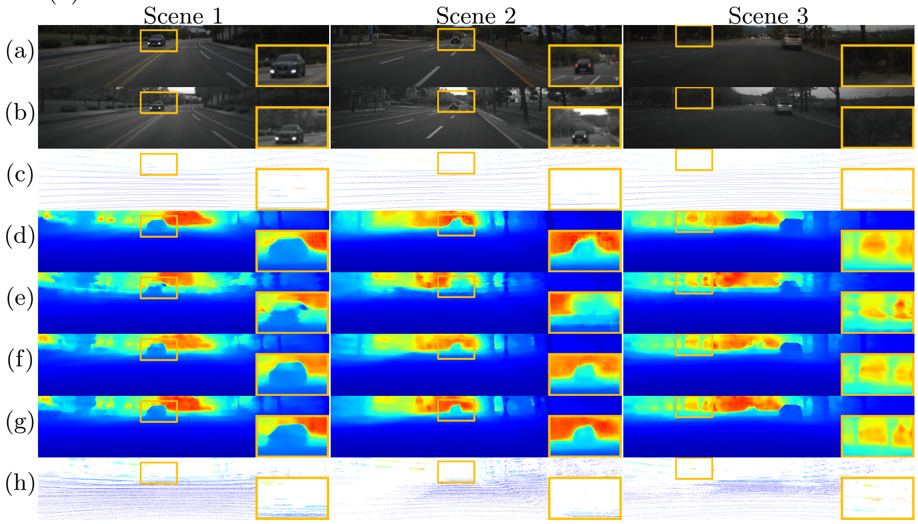
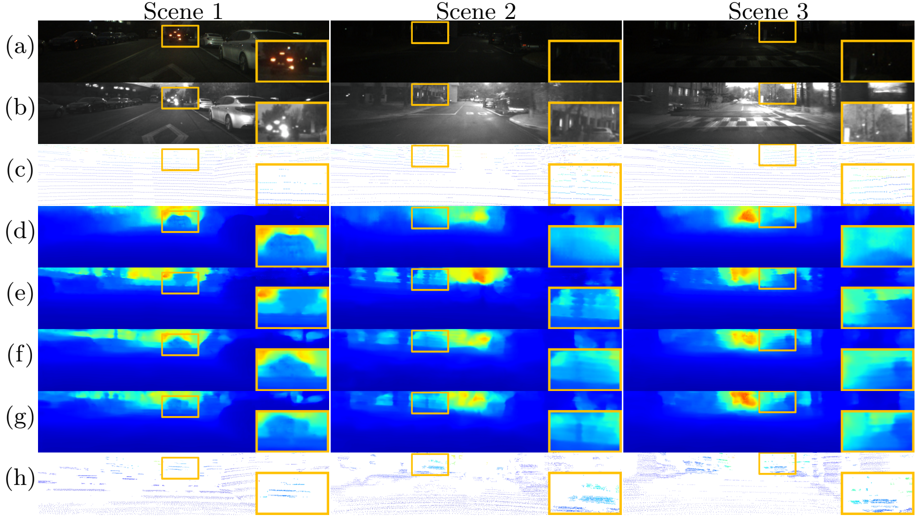


**Fig. S3. Depth map comparisons on the MMDCE night [2] dataset.** (a) RGB, (b) NIR, (c) LiDAR, (d) LS [1], (e) CompletionFormer [4], (f) MMDNet [2], (g) Ours, and (h) GT.



## References

1. Cheng, X., Zhong, Y., Dai, Y., Ji, P., Li, H.: Noise-aware unsupervised deep lidar-stereo fusion. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6339–6348 (2019)
2. Park, J., Jeong, Y., Joo, K., Cho, D., Kweon, I.S.: Adaptive cost volume fusion network for multi-modal depth estimation in changing environments. IEEE Robotics and Automation Letters **7**(2), 5095–5102 (2022)
3. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant cnns. In: Int. Conf. 3D Vis. pp. 11–20. IEEE (2017)
4. Youmin, Z., Xianda, G., Matteo, P., Zheng, Z., Guan, H., Stefano, M.: Completionformer: Depth completion with convolutions and vision transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18527–18536 (2023)