Exploiting Cross-modal Cost Volume for Multi-sensor Depth Estimation

Janghyun Kim[1][0009−0001−3525−062X], Ukcheol Shin[2][0000−0001−8363−9886], Seokyong Heo[1][0009−0008−0236−1308], and Jinsun Park[3⋆][0000−0002−2296−819X]

[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

Abstract. Single-modal depth estimation has shown steady improvement over the years. However, relying solely on a single imaging sensor such as RGB and near-infrared (NIR) cameras can result in unreliable and erroneous depth estimation, particularly in challenging lighting conditions such as low-light or sudden lighting change scenarios. Thereby, several approaches have leveraged multiple sensors for robust depth estimation. However, the effective fusion method that maximally utilizes multi-modal sensor information still requires further investigation. With this in mind, we propose a multi-modal cost volume fusion strategy with cross-modal attention, incorporating information from both cross-spectral and single-modality pairs. Our method initially constructs low-level cost volumes that consist of modality-specific (*i.e.*, single modality) and modality-invariant (*i.e.*, cross-spectral) volumes from multi-modal sensors. These cost volumes are then gradually fused using bidirectional cross-modal fusion and unidirectional LiDAR fusion to generate a multi-sensory cost volume. Furthermore, we introduce a straightforward domain gap reduction approach to learn modality-invariant features and depth refinement techniques through cost volume-guided propagation. Experimental results demonstrate that our method achieves SOTA (State-of-the-Art) performance under diverse environmental changes.

Keywords: Depth Estimation · Sensor Fusion · Cross-modal Attention

1 Introduction

Depth estimation is an essential technique in various real-world applications, such as robotics, augmented reality, and autonomous driving. There are various approaches including monocular depth estimation [8, 9, 1, 17], stereo depth estimation [2, 38, 25, 35, 36, 34], and multi-modal depth estimation [22, 26] to name a few. Among them, stereo and multi-modal methods often demonstrate robust performance compared to the monocular method in various environments.

⋆ Corresponding author.

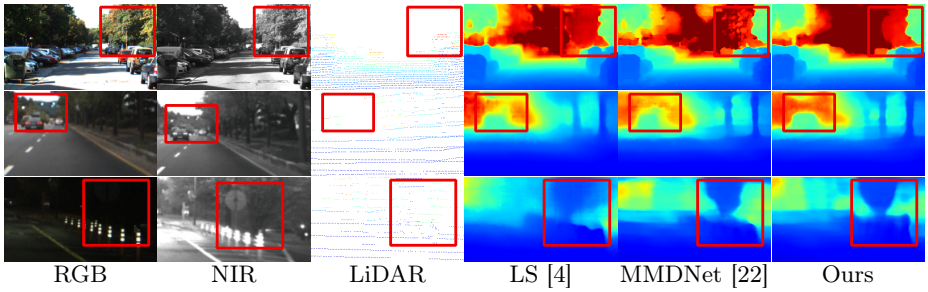


Fig. 1. Depth estimation results in changing environments. Top row: KITTI MMD, middle row: MMDCE day, and bottom row: MMDCE night. Note that the highlighted red box in the NIR image is shifted to match the reference view.

However, relying solely on a single imaging sensor such as RGB and near-infrared (NIR) cameras can result in unreliable and erroneous depth estimation, particularly in challenging lighting conditions like low-light or sudden lighting change scenarios (*e.g.*, driving at night or entering a tunnel). Therefore, previous studies have attempted to utilize diverse sensors for reliable and robust depth estimation against challenging lighting conditions, such as RGB-LiDAR fusion [5, 4] and RGB-NIR-LiDAR fusion [22]. These methods resolve sensor-fusion related issues, such as far-depth value [5], sensor misalignment [4], and redundancy of depth basis [22]. However, they have overlooked the potential benefits of exploiting multiple sensors in terms of reliability and precision. For example, learning modality-invariant features from heterogeneous sensors may allow a network to predict a reliable depth map less affected by lighting conditions. Additionally, diverse baselines between sensors enable the network to cover short-, middle-, and long-distance depth range searching.

In this paper, we propose a novel metric depth estimation framework that maximally exploits cross-modal cost volumes by learning modality-invariant features and considering diverse baselines. The proposed framework consists of cost volume generation module, multi-modal cost volume fusion module, and cost volume-guided propagation module. Given stereo RGB, stereo NIR, and LiDAR data, the framework constructs modality-specific cost volumes from each stereo pair and modality-invariant cost volumes from cross-spectral images. After that, a multi-sensory cost volume that includes modality-specific and modality-invariant properties is aggregated by the proposed cross-modal attention block. Lastly, the cost volume-guided propagation module predicts a reliable and precise depth map from the multi-sensory cost volume against challenging lighting conditions, as shown in Fig. 1. Our contributions are summarized as follows:

- We propose a novel framework that maximally exploits cross-modal cost volumes along with single-modal cost volumes by learning modality-invariant features and aggregating cost volumes constructed by diverse baselines.
- We design a cross-modal attention block that integrates modality-specific and modality-invariant properties of multiple sensors. (*e.g.*, RGB, NIR, and LiDAR).

– We introduce a simple yet effective modality-invariant feature learning method that is less affected by lighting conditions by utilizing structure consistency.
– The proposed network achieves state-of-the-art results on the KITTI MMD [29, 7] and MMDCE datasets [22] containing diverse environmental changes.

2 Related Work

2.1 Depth Estimation

Stereo Depth Estimation. Stereo depth estimation [2, 38, 25, 35, 26, 36] is a fundamental task in computer vision, involving the calculation of disparity from two images captured from the different viewpoints. In recent years, this technique has been extensively researched due to its broad applications [31]. PSMNet [2] employs a stacked hourglass to iteratively refine the cost volume in an end-to-end manner. AANet [38] proposed an effective intra- and cross-scale cost aggregation algorithm, achieving fast inference times. PCW-Net [25] introduced a multi-scale cost fusion method to cover diverse receptive fields and efficient warping volume-based disparity refinement. These depth estimation methods depend on a single modality, typically relying on RGB cameras to generate depth maps. While these stereo depth estimation networks perform well in general situations, they struggle to handle challenging lighting conditions.
Stereo-LiDAR Fusion. To address this limitation, several methods incorporate multi-modal sensors for depth estimation. LS [4] proposed an unsupervised depth estimation network based on noise-aware LiDAR and stereo fusion. VPN [5] developed a volumetric propagation network that can perform long-range depth estimation based on a stereo-LiDAR fusion network. Moreover, MMDNet [22] incorporated multiple sensors (*i.e.*, RGB-NIR-LiDAR) with adaptive cost volume to minimize computational cost and depth basis redundancy for robust performance. While these methods have shown promising performance in multi-modal depth estimation, they often struggle to effectively identify and utilize the most crucial factors among the various sensors in different scenarios. To address these challenges, we propose a novel multi-sensor stereo matching network that integrates single- and cross-spectral modalities with a single-scan LiDAR sensor. Our framework estimates depth values across short to far distances by leveraging diverse sensor configurations.
Spatial Propagation Network. Spatial Propagation Network (SPN) [3, 23, 20, 42, 32, 40, 15] is a widely adopted method in depth completion to refine initial depth maps through iterative propagation procedures. SPNs utilize affinities between neighboring pixels, where these affinities are typically extracted from geometric cues in visual information from RGB images. Consequently, they are not directly applicable to stereo matching and show poor performance on a NIR-LiDAR sensor configuration under nighttime conditions [22]. To address these limitations, we propose a cost volume-guided propagation method that does not rely on specific sensor characteristics, making it more versatile for refining depth in diverse scenarios.

2.2 Cross-Attention Algorithm

Recently, cross-attention algorithms [14, 12, 24, 17, 16] have been extensively researched and achieved high performance in various applications. In stereo depth estimation, several stereo matching networks have attempted to employ cross-attention algorithms. STTR [18] combined the benefits of CNNs and Transformers, using cross-attention to compute correspondences between epipolar lines in image pairs. ChiTransformer [27] introduced a pattern retrieval mechanism, employing cross-attention between features from two different views instead of traditional stereo matching. Xu et al. [37] simply utilized cross-attention in the feature extraction stage. However, these methods did not focus on cost volume generation, which is a crucial factor in stereo matching. Since the cost volume representation is closely related to the depth regression coefficients [2], accurate cost volume construction is essential for high-quality depth estimation. Therefore, we propose a novel cross-modal attention block that integrates modality-specific and modality-invariant cost volumes, incorporating reliable and crucial cues from various sensor pairs. This enhancement enables our network to achieve significantly higher performance across real-world changing environments.

2.3 Domain Gap Reduction

Several methods [19, 11, 44] have been proposed to reduce domain gaps in cross-spectral stereo matching. Liang et al. [19] and Guo et al. [11] attempted to minimize the domain gap using GANs [10] in an unsupervised manner. Zhi et al. [44] introduced a material-aware loss function to translate RGB images to pseudo-NIR images. However, these approaches primarily focus on enhancing visual similarity, while structural similarity is more critical for depth estimation. Therefore, we opt for a simpler approach, maintaining consistency between different modalities using SSIM [33] loss to create cross-modal cost volumes. This approach ensures the generation of modality-invariant cost volumes from cross-spectral images.

3 Method

In this section, we first provide an overview of our network architecture and describe the proposed cross-modal attention block designed for both low-level and high-level cost volumes. Next, we introduce a straightforward approach to reduce the domain gap while learning modality-invariant features for cross-spectral matching. Finally, we present a cost volume-guided propagation method to refine the initial predicted depth map for improved versatility across diverse scenarios.

3.1 Overall Architecture

Figure 2 illustrates the architecture of our model for densely predicting depth maps in diverse environments. In this work, we perform stereo matching not only

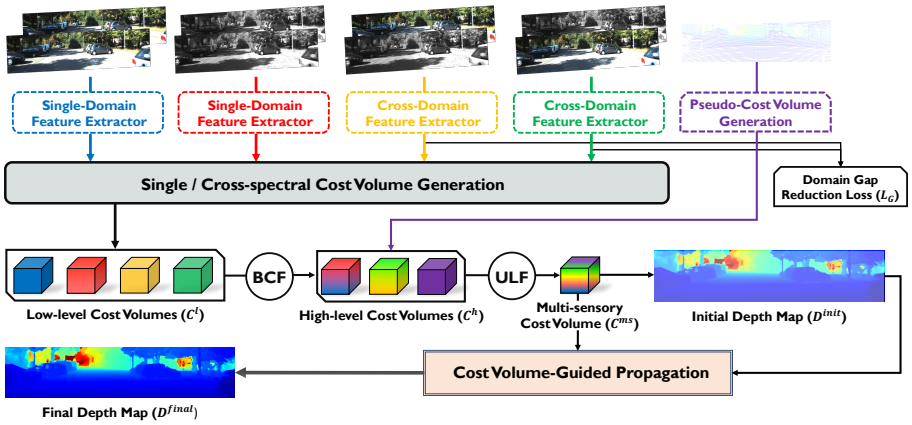


Fig. 2. Overview of the proposed architecture. Given stereo RGB, stereo NIR, and single-scan LiDAR, domain-specific encoders independently process each modality to capture the unique characteristics of those sensors. After that, we integrate all cost volumes through Bidirectional Cross-modal Fusion (BCF) and Unidirectional LiDAR Fusion (ULF) to extract valuable cues from each domain. Subsequently, the initial depth map D^{init} is obtained from the multi-sensory cost volume and refined through a cost volume-guided propagation module to predict the final depth map D^{final}.

within the same sensor modalities but also across different sensor modalities. We generate two low-level cost volumes from the single-stereo pair of RGB and NIR images, respectively. Additionally, we create two low-level cross-spectral cost volumes by pairing RGB and NIR images. Each pair of sensor configurations has a unique depth basis, which results in varying disparity ranges as shown in Fig. 3. Here, the depth basis $\mathcal{D} = \{d_m\}_{m=1}^{M}$ represents a set of depth candidate values with cardinality M, which is closely related to the coefficients for depth map D regression defined as follows:

$$D = \sum_{m=1}^{M} d_m \cdot \texttt{softmax}(C_m^a), \tag{1}$$

where C_m^a is the m-th cost slice of the aggregated cost volume C^a and $\texttt{softmax}(\cdot)$ is the softmax function. Note that disparity values to calculate each cost slice are determined corresponding to $\mathcal{D}$ [22]. Thus, utilizing various bases, including cross-spectral modalities, allows us to capture a broader spectrum of depth resolution, combining the strengths of each sensor pair to achieve more comprehensive and accurate depth estimation. This approach enhances the robustness of our cost volume generation strategy.

The generated low-level cost volumes are then gradually combined using Bidirectional Cross-modal Fusion (BCF) to construct two high-level cost volumes. We also generate a LiDAR pseudo-cost volume to leverage the precise depth prior from point clouds acquired by the LiDAR. After that, we combine the LiDAR

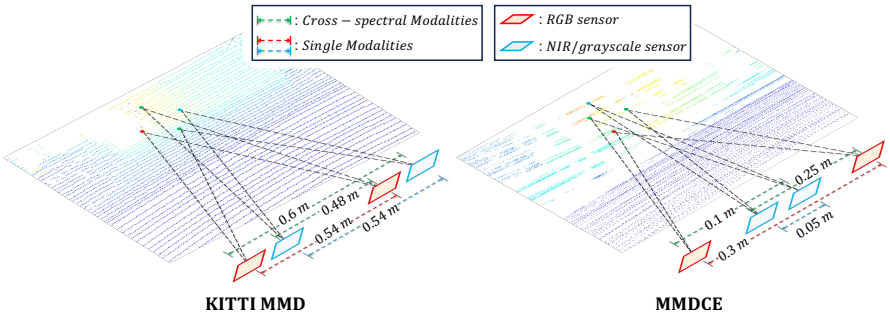


Fig. 3. Depth range variations. The depth range is determined by the configuration of stereo sensors, specifically basis value between them. Different sensor pairs have distinct depth ranges due to their unique sensor configurations. By leveraging these varying depth ranges, we can integrate the unique characteristics of all sensors without being limited by their individual mechanical structures.

pseudo-cost volume with the stereo cost volumes through our Unidirectional LiDAR Fusion (ULF) to obtain a final multi-sensory cost volume that contains the combined information from all sensor modalities. Subsequently, following the standard stereo estimation procedure [38], we go through the aggregation and estimation process to derive an initial depth map. We further refine the initial depth map through cost volume-guided propagation. This step involves non-local spatial propagation [23] with four iterations, leveraging both the initial depth map and the aggregated cost volume. With this novel architecture that maximally exploits cross-modal and single-modal cost volumes, our model can predict reliable and accurate depth maps even in challenging environments.

3.2 Multi-modal Cost Volume Fusion

We propose Bidirectional Cross-modal Fusion (BCF) and Unidirectional LiDAR Fusion (ULF) blocks to create a cost volume containing reliable information from various sensors. For these fusion approaches, we design the Cross-Modal Attention (CMA) block as shown in Fig. 4. Note that we employ geometry-aware warping [23] when fusing each cost volume for both BCF and ULF.

Cross-Modal Attention (CMA) Block. Our proposed CMA reinforces both the modality-specific and modality-invariant information of multiple sensors. The CMA is defined as follows:

$$Q_\psi = f_\psi^Q\left(C_\psi\right), \quad K_\psi = f_\psi^K\left(C_\psi\right), \quad V_\psi = f_\psi^V\left(C_\psi\right), \tag{2}$$

$$R_\psi = \texttt{softmax}\left(Q_\psi \otimes K_\psi^\top\right), \tag{3}$$

$$\hat{C}_\psi = \text{CMA}\left(C_\psi, C_{\psi'}\right) = R_{\psi'} \otimes V_\psi + f_\psi^{skip}\left(C_\psi\right), \tag{4}$$

where ψ denote *src* or *tgt* for source and target cost volumes, respectively, and ψ' denote the other domain relative to ψ. $f_\psi^Q(\cdot)$, $f_\psi^K(\cdot)$, $f_\psi^V(\cdot)$, $f_\psi^{skip}(\cdot)$, and $\otimes$ denote

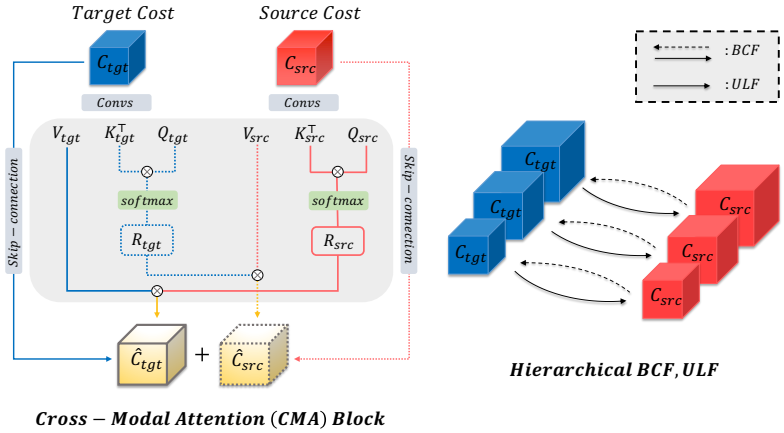


Fig. 4. Cross-Modal Attention (CMA) block. We extract the key, query, and value from each cost volume and calculate the correlation score. BCF is bidirectionally performed for single and cross-spectral modalities' cost volumes (denoted as solid and dotted rays). ULF is conducted similarly, using the LiDAR pseudo-cost volume as the target and the others as sources in a unidirectional manner (denoted as solid rays). We utilize CMA in a hierarchical manner across three different scale cost volumes.

convolution blocks for query Q, key K, value V, skip-connection, and element-wise multiplication, respectively. Note that C_{tgt} is the target cost volume from the reference view. For example, CMA generates $\hat{C}_{tgt}$ by first extracting Q_{src} and K_{src} from C_{src} and V_{tgt} from C_{tgt} using Eq. (2). Then, the relation score of the source feature R_{src} is calculated using Eq. (3). At the same time, the feature from skip-connection is extracted by $f_{tgt}^{skip}(C_{tgt})$. Afterward, the final cost volume $\hat{C}_{tgt}$ containing information from both the source and target cost volumes is obtained using Eq. (4). Since our stereo matching constructs cost volumes at three scales, we conduct hierarchical CMA for each scale.

Bidirectional Cross-modal Fusion (BCF). We perform CMA bidirectionally for single and cross-spectral modalities' cost volumes. By adopting a bidirectional approach, we can incorporate the benefits of both cost volumes, allowing for a more comprehensive integration of reliable cues. Specifically, BCF can enhance the modality-specific and modality-invariant properties within the single and cross-spectral cost volumes, respectively. The proposed BCF can be formulated as follows:

$$C_{cs}^{h} = \mathrm{CMA}(C_{NIR-RGB}^{l}, C_{RGB-NIR}^{l}) + \mathrm{CMA}(C_{RGB-NIR}^{l}, C_{NIR-RGB}^{l}), \quad (5)$$

$$C_{s}^{h} = \mathrm{CMA}(C_{NIR}^{l}, C_{RGB}^{l}) + \mathrm{CMA}(C_{RGB}^{l}, C_{NIR}^{l}), \quad (6)$$

where C_{cs}^{h} and C_{s}^{h} denote fused high-level cost volumes generated from cross-spectral and single-domain volumes, respectively. C^{l} denotes low-level cost volumes generated by stereo matching in RGB and NIR images (*i.e*, C_{RGB}^{l}, C_{NIR}^{l}).

Unidirectional LiDAR Fusion (ULF). We propose the ULF strategy to fuse cost volumes generated from CMF with the LiDAR pseudo-cost volume C_{LiDAR}. Given the importance of the LiDAR sensor as a precise depth prior, ULF adopts a unidirectional fusion strategy. We regard the cost volumes obtained through BCF from all image sensors as the source cost volume, and C_{LiDAR} constructed from pseudo-cost volume generation [22] as the target cost volume. The proposed ULF is defined as follows:

$$\begin{aligned} C^{ms} &= \mathrm{ULF}(C_{LiDAR}, C_s^h, C_{cs}^h) \\ &= \mathrm{CMA}(C_{LiDAR}, C_s^h) + \mathrm{CMA}(C_{LiDAR}, C_{cs}^h), \end{aligned} \tag{7}$$

where C^{ms} denotes the multi-sensory cost volume that incorporates reliable information from all sensors. Note that in the previous fusion method [22], geometry-aware warping is solely applied to fuse multi-modal cost volumes using intrinsic and extrinsic parameters. However, naïve geometry-aware warping is prone to sensor calibration errors, causing misalignment during the fusion process. Therefore, we introduce our ULF block on top of geometry-aware warping to implicitly mitigate these errors through its attention-based suppression capability of irrelevant information.

3.3 Domain Gap Reduction

We opt for a straightforward approach to maintain modality-invariant features between different modalities (*i.e.*, cross-domain) by using the Structural Similarity Index (SSIM) [33] loss when creating cost volumes. It is effective in capturing changes in structural information, which is crucial for maintaining feature consistency across different modalities. Our domain gap reduction loss is defined as follows:

$$L_G = \frac{1}{N} \sum_{n=1}^{N} \left(1 - \mathrm{SSIM}\left(F_n^{RGB}, F_n^{NIR}\right)\right), \tag{8}$$

where n, N, F^{RGB}, and F^{NIR} denote the pixel index, the number of pixels, and features extracted from RGB and NIR images, respectively. Here, we employ SSIM loss with a 3×3 block filter inspired by Monodepth [8]. It evaluates the similarity between two images based on three components: luminance, contrast, and structure. By minimizing the domain gap reduction loss, we train the encoders to extract domain-invariant structural features that appear consistently across different domains (*e.g.*, strong edges, blobs, and corners). This consistency is crucial for accurate cross-spectral matching, as it aligns the feature representations across different sensor types. Using SSIM loss in our approach helps preserve the structural similarity between the RGB and NIR features, effectively reinforcing modality-invariant feature learning without requiring additional networks (*e.g.*, GANs). We argue that the introduction of L_G is computationally efficient as well as effective in maintaining the fidelity of feature representations across modalities.

3.4 Cost Volume-Guided Propagation

To further improve the depth estimation accuracy, we adopt an SPN-based depth refinement process after the initial depth regression from aggregated cost volume C^a (*cf.*, Eq. (1)). C^a is generated by the multi-scale cost volume aggregation [38] using C^{ms} at three different scales. Different from conventional SPNs generating affinities from visual cues [3, 23, 32], we extract them from the aggregated cost volume C^a and the initial depth map D^{init}. Since the cost volume contains probabilities for each disparity range (*cf.*, Eq. (1)), it can serve as a crucial cue for propagation. Our SPN process can be formulated as follows:

$$\mathcal{W}, \mathcal{N} = f^{SPN}\left(C^a, D^{init}\right), \quad \mathcal{W}, \mathcal{N} \in \mathbb{R}^{N \times V}, \tag{9}$$

where $f^{SPN}(\cdot)$ represents an encoder network, $\mathcal{W}$ and $\mathcal{N}$ denote the affinity and indices of non-local neighbors, respectively, and V is the number of neighbors for each pixel. The propagation process updates the depth prediction iteratively, which is defined at the time step t as follows:

$$D_n^{t+1} = w_n D_n^t + \sum_{v \in \mathcal{N}_n} \mathcal{W}_{n,v} D_v^t, \quad w_n = 1 - \sum_{v \in \mathcal{N}_n} \mathcal{W}_{n,v}, \tag{10}$$

where $D_n^t, w_n, \mathcal{N}_n, v, \mathcal{W}_{n,v}$ denote the depth value at n-th pixel at the time step t, the reference affinity, indices of neighbors of n-th pixel, neighbor index, and the v-th neighbor's affinity value of the n-th pixel, respectively. After this iterative refinement process, the final dense depth prediction D^{final} is obtained.

3.5 Loss Function

The total loss function L_{total} for our model is defined as follows:

$$L_{total} = \sum_{l \in \{1,2\}} \left(\frac{1}{N} \sum_{n=1}^{N} \left\| D_n^{GT} - D_n^{final} \right\|_l \right) + \alpha L_G, \tag{11}$$

where $D^{GT}, \|\cdot\|_l$, and α denote the ground truth (GT) depth, ℓ_l norm, and a balancing hyperparameter, respectively. Note that we utilize the combination of ℓ_1 and ℓ_2 norms to balance between average depth accuracy and sharp depth prediction boundaries [22].

4 Experiment

In this section, we first describe the implementation details and show the effectiveness of our method on the KITTI MMD and MMDCE datasets. We conduct experiments on these datasets containing environmental changes, following the same procedure as MMDNet [22]. Additionally, we provide extensive ablation studies on the proposed CMA, domain gap reduction, and cost volume-guided propagation approaches. We also analyze the performances of various sensor

Table 1. Quantitative performance comparison on the KITTI MMD dataset. (A): 90.0K, (B): 42.9K, and (C): 32.9K images are used for training, respectively.

Method			Input		RMSE	MAE	iRMSE	iMAE	Time
			RGB	Grayscale	(mm)	(mm)	(1/km)	(1/km)	(s)
Mono + LiDAR	(A)	DDP [41]	✓	-	1310.03	347.17	-	-	-
		NConv [6]			908.76	209.56	2.50	0.90	-
		S2D [21]			878.56	260.90	3.25	1.34	0.08
		DN [39]			811.07	236.67	2.45	1.11	-
		GuideNet [28]			777.78	221.59	2.39	1.00	0.14
		NLSPN [23]			771.80	197.30	2.00	0.80	0.22
		PENet [13]			757.20	209.00	2.22	0.92	0.03
		CompletionFormer [42]			741.44	194.99	2.01	0.84	0.12
		DySPN [20]			739.40	191.40	-	-	0.16
		LRRU [20]			723.40	<u>188.10</u>	1.90	0.80	0.13
		TPVD [40]			718.90	**187.15**	-	-	0.15
Stereo + LiDAR	(B)	SLFNet [43]	✓	-	641.10	197.00	1.77	0.87	0.16
		VPN [5]			<u>636.20</u>	205.10	1.87	0.99	1.41
	(C)	LS [4]	✓	-	832.16	283.91	2.19	1.10	0.34
		CCVN [30]			749.30	252.50	1.40	0.80	1.01
		MMDNet [22]		✓	673.34	202.56	1.69	0.80	0.12
		Ours			**622.14**	208.26	1.71	0.88	0.17

Bold: The best, <u>Underline</u>: The second-best

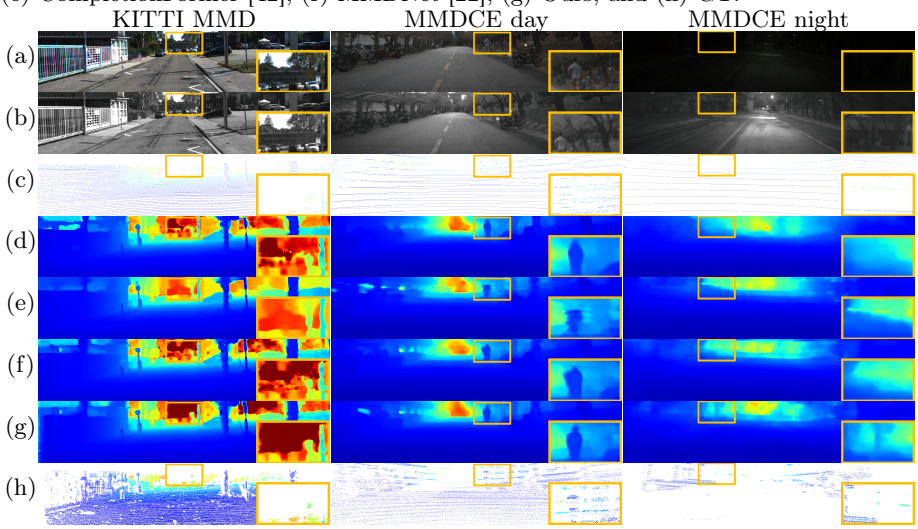
pairs to demonstrate the effectiveness of utilizing multi-sensory combinations with the proposed CMA. Our model is trained with a batch size of 12 for 25 epochs using PyTorch with 4 RTX A6000 GPUs and tested on an RTX 4090 GPU for all datasets. We set $\alpha = 0.5$ in Eq. (11) to balance the overall loss functions in all datasets. For the quantitative evaluation, we utilize metrics used in previous works [29]: RMSE, MAE, iRMSE, and iMAE.

4.1 KITTI MMD Dataset

The KITTI multi-modal depth (KITTI MMD) dataset [29, 7] contains over 37K pairs of RGB, grayscale, and LiDAR images. We utilize 32.9K images for training, 3.4K for validation, and 1K for testing, respectively. These samples are selected from traceable sequences in the KITTI raw dataset to provide the grayscale stereo pairs required by MMDNet. Note that the test dataset is the same as the original KITTI depth completion (KITTI DC) dataset [29]. Here, we utilize grayscale images in place of the previously defined NIR images.

Table 1 presents quantitative comparison results on the KITTI MMD test dataset. LS [4], CCVN [30], and MMDNet [22] were trained on a dataset with 32.9K samples, while other methods were trained with 42.9K and 90.0K samples. All datasets have identical validation and test splits because they follow the original configuration of the KITTI DC dataset. Compared to previous mono-LiDAR [41, 6, 39, 42, 40] and stereo-LiDAR depth estimation networks [43, 5, 4, 30, 22], our method brings substantial performance improvement in terms of RMSE. Our network shows significant improvement from 13.5% to 52.5% in the RMSE metric, compared to other state-of-the-art depth completion networks (*i.e*, mono-LiDAR depth estimation). Moreover, our method outperforms the

Fig. 5. Depth map comparisons on the KITTI MMD, MMDCE day, and MMDCE night datasets. (a) RGB, (b) NIR or Grayscale, (c) LiDAR, (d) LS [4], (e) CompletionFormer [42], (f) MMDNet [22], (g) Ours, and (h) GT.



baseline multi-sensor fusion model MMDNet [22] by approximately 7.6%. This is because our network effectively utilizes both single and cross-spectral modalities from multiple sensors. Our approach allows the network to capture a dynamic range of disparities corresponding to short to long distances for precise depth estimation. Furthermore, our method also exhibits superior speed compared to other stereo-LiDAR depth estimation networks except MMDNet and SLFNet.

Figure 5 visually illustrates the remarkable performance of the proposed network. Our network effectively distinguishes edge boundaries by leveraging CMA block and multiple modalities. In addition, by utilizing only reliable cues and performing depth refinement through cost volume-guided propagation, we can effectively separate foreground and background areas.

4.2 MMDCE Dataset

The Multi-Modal Depth in Changing Environments (MMDCE) dataset includes various day-night scenarios. We utilize 6,628 image pairs, comprising 5,876 for daytime (Train: 4,344, Validation: 656, Test: 876) and 752 for nighttime (Train: 601, Test: 151) data of RGB, NIR, and LiDAR sensors.

Table 2 provides quantitative comparison results on the MMDCE day-night dataset. Note that we obtained the model for the nighttime split by fine-tuning the model initially trained on the daytime images on the nighttime ones due to the lack of nighttime training data, following previous work [22]. To achieve optimal performance in both day and night scenes, it is crucial for the network to adaptively learn and utilize the most relevant information from different sensors.

Table 2. Performance comparison on the MMDCE day-night dataset (M: Mono, S: Stereo, L: LiDAR).

(a) Day

	Method	Input RGB	Input NIR	RMSE (mm)	MAE (mm)
M + L	NLSPN [23]	✓	-	1750.6	709.7
		-	✓	1791.4	831.8
	GuideNet [28]	✓	-	1486.7	697.0
		-	✓	1658.6	832.7
	CompletionFormer [42]	✓	-	1470.6	587.2
		-	✓	1583.0	734.8
S + L	LS [4]	✓	-	1759.6	939.8
		-	✓	13009.5	8353.5
	CCVN [30]	✓	-	2141.4	1046.2
		-	✓	5884.9	2379.0
	MMDNet [22]	✓	✓	<u>1226.2</u>	<u>610.4</u>
	Ours			**1092.3**	**507.4**

Bold: The best, <u>Underline</u>: The second-best

(b) Night

	Method	Input RGB	Input NIR	RMSE (mm)	MAE (mm)
M + L	NLSPN [23]	✓	-	1755.2	716.8
		-	✓	2126.5	1031.8
	GuideNet [28]	✓	-	1864.9	804.3
		-	✓	1892.9	963.8
	CompletionFormer [42]	✓	-	1386.0	682.4
		-	✓	2146.7	1058.8
S + L	LS [4]	✓	-	3589.8	1431.8
		-	✓	9289.3	6162.2
	CCVN [30]	✓	-	1722.4	727.0
		-	✓	3884.4	1569.2
	MMDNet [22]	✓	✓	<u>1371.3</u>	<u>663.6</u>
	Ours			**1295.3**	**627.9**

Bold: The best, <u>Underline</u>: The second-best

Each sensor has unique strengths and weaknesses for estimating depth under specific conditions. While previous depth completion networks and stereo matching networks do not consider which sensor data is crucial for different scenes, our network leverages the strengths of each sensor through the proposed hierarchical CMA strategy. This approach effectively selects and integrates the most valuable data from each sensor, enabling adaptation to a variety of lighting conditions and environmental challenges.

As a result, our network achieved high accuracy across all metrics compared to previous approaches in both scenarios. Specifically, our network achieves 10.9% and 16.9% higher performance than the previous SOTA approach [22] at daytime scenarios in terms of RMSE and MAE, respectively. Correspondingly, our network also exhibits improvements of 5.5% and 5.4% in RMSE and MAE for nighttime scenarios. Furthermore, SPN-based depth completion networks [23, 42] experience performance degradation in nighttime scenarios when using NIR-LIDAR sensors. This limitation primarily stems from these networks' inability to capture pertinent cues from blur artifacts caused by low-light conditions. However, our network achieves high performance even with NIR sensors by utilizing a comprehensive cost volume that integrates reliable cues for depth regression and facilitates additional propagation.

Figure 5 showcases qualitative comparisons on the MMDCE day-night datasets, highlighting the superior performance of our network compared to previous approaches. Especially in nighttime scenarios, our approach better distinguishes tiny objects and background areas compared to previous works. Furthermore, our network produces more sharper predictions without blur effects during daytime scenarios compared to SPN-based method [42], which suffers from lower sharpness due to the absence of LiDAR data. Our approach demonstrates robustness and adaptability across varying lighting conditions and scenarios.

Table 3. Ablation study results on the KITTI MMD dataset.

(a) Ablation on Model Components

Modality	CMA	Cross-spectral	L_G	SPN	RMSE (mm)	MAE (mm)
RGB-Gray-LIDAR	Base fusion				689.12	233.26
	✓				656.06	223.14
		✓			738.68	236.31
	✓	✓			650.20	220.92
	✓	✓	✓		645.14	220.03
	✓	✓	✓	✓	622.14	208.26

(b) Ablation on Modality Combinations

Modality	CMA BCF	CMA ULF	RMSE (mm)	MAE (mm)
RGB			885.71	335.14
Gray			922.84	336.75
LiDAR			817.25	269.47
RGB-Gray			888.32	327.25
	✓		821.79	306.47
RGB-LiDAR			721.41	247.81
		✓	662.49	225.87
Gray-LiDAR			755.05	255.44
		✓	687.88	225.72
RGB-Gray-LiDAR			689.12	233.26
	✓	✓	656.06	223.14

Table 4. Ablation study results on the MMDCE Day and Night datasets.

(a) Day

Modality	CMA BCF	CMA ULF	RMSE (mm)	MAE (mm)	iRMSE (1/km)	iMAE (1/km)
RGB			1813.7	916.5	11.1	6.5
NIR			1955.2	1064.4	15.1	9.2
LiDAR			1610.5	726.9	10.1	5.6
RGB-NIR			1766.2	912.4	11.5	6.7
	✓		1680.4	843.4	9.5	5.7
RGB-LiDAR			1293.6	605.5	8.3	4.6
		✓	1230.5	598.9	8.1	4.5
NIR-LiDAR			1378.5	655.1	8.6	4.9
		✓	1287.9	612.2	8.2	4.7
RGB-NIR-LiDAR			1230.5	598.9	8.5	4.9
	✓	✓	1142.3	557.4	8.0	4.6

(b) Night

Modality	CMA BCF	CMA ULF	RMSE (mm)	MAE (mm)	iRMSE (1/km)	iMAE (1/km)
RGB			2035.9	1119.3	18.1	11.2
NIR			2320.8	1459.5	24.9	16.0
LiDAR			1713.2	763.7	11.6	7.0
RGB-NIR			1986.6	1169.2	18.3	11.8
	✓		1954.3	1086.7	13.4	8.9
RGB-LiDAR			1378.6	695.3	10.7	6.8
		✓	1327.7	651.2	10.0	6.2
NIR-LiDAR			1392.8	740.3	11.2	7.0
		✓	1334.7	648.1	9.4	5.8
RGB-NIR-LiDAR			1364.7	687.7	10.5	6.4
	✓	✓	1322.3	637.9	9.1	5.6

4.3 Ablation Studies

Effectiveness of the Proposed Framework. We have performed ablation studies on the KITTI MMD test dataset to provide effectiveness of our proposed approaches as shown in Tab. 3. First, we evaluate the performance of the base fusion model [22] with ℓ_1 and ℓ_2 losses. When integrating cross-spectral modalities with single modalities without additional methods (*e.g.*, CMA and domain gap reduction), the model fails to replicate the performance of the base fusion model. This is because simply averaging all cost volumes [22] overlooks the varying significance of each cost volume. To address this issue, we employ CMA which can effectively adjust the importance values among the different cost volumes. As a result, our network further improves RMSE from 689.12 to 650.20 and MAE from 233.26 to 220.92 compared to the base fusion model. Additionally, we introduce domain gap reduction loss L_G to learn modality-invariant features between RGB and NIR features. This approach facilitated the minimization of feature discrepancies, leading to a further improvement in RMSE from 650.20 to 645.14. Lastly, our cost volume-guided propagation method brings 3.7% and 5.3% performance improvements in terms of RMSE and MAE, respectively. These consistent performance improvements strongly support the assertion that all components of the proposed framework contribute significantly to robust depth estimation.

Modality Combination. We evaluate the impact of our fusion strategy for stereo matching paired with LiDAR sensors across diverse environments as shown in Tab. 3 (b) and Tab. 4. Throughout the all datasets and modality pairs, utilizing CMA consistently demonstrates significant effectiveness in whole metrics. Compared with the base fusion model [22] (*i.e.*, RGB-NIR/Gray-LiDAR), adopting our CMF and MLF exhibits a 3.1% to 7.1% improvement in RMSE and a 4.3% to 7.2% improvement in MAE across the datasets. These results underscore the efficacy of CMA in enhancing depth estimation accuracy by effectively integrating information from multiple sensor modalities. In the KITTI MMD and MMDCE day datasets, employing CMF and MLF yields similar performance improvements. However, relying solely on images shows minimal performance enhancement in the MMDCE night scenes. This highlights the varying importance of different sensor pairs across different scenarios. In essence, utilizing our proposed CMA with all sensor pairs demonstrates the capability to achieve optimal performance across diverse conditions.

5 Conclusion

In this paper, we have proposed a novel multi-modal depth estimation framework that effectively exploits both cross-modal cost volumes (*i.e*, modality-invariant) and single-modal cost volumes (*i.e*, modality-specific). We also have introduced a straightforward approach to maintain the structure similarity of features within cross-modal cost volumes to maximize their invariant characteristics. Furthermore, we have designed a cross-modal attention block that consistently integrates modality-specific and modality-invariant properties with a LiDAR sensor to construct multi-sensory cost volume. This approach makes our network capture dynamic depth ranges from short to long distances and facilitates depth regression using only reliable cues while suppressing redundant and irrelevant information. To further improve the depth estimation robustness, we have proposed cost volume-guided propagation. It is noteworthy that all the proposed methods and network architecture represent a pioneering effort in multi-modal stereo matching. As a result, our network achieves state-of-the-art performance in the KITTI MMD and MMDCE datasets providing various environmental changes. In future work, it would be valuable to explore alternative methods for generating cost volumes from LiDAR sensors instead of relying on constructing pseudo-cost volumes.

Acknowledgments. This work was supported in part by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2023-00260098, 50%) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00217689, 50%).

References

1. Bae, G., Budvytis, I., Cipolla, R.: Multi-view depth estimation by fusing single-view depth probability with multi-view geometry. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2842–2851 (2022)
2. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5410–5418 (2018)
3. Cheng, X., Wang, P., Guan, C., Yang, R.: Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In: AAAI. vol. 34, pp. 10615–10622 (2020)
4. 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)
5. Choe, J., Joo, K., Imtiaz, T., Kweon, I.S.: Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation. IEEE Robotics and Automation Letters **6**(3), 4672–4679 (2021)
6. Eldesokey, A., Felsberg, M., Khan, F.S.: Confidence propagation through cnns for guided sparse depth regression. IEEE Trans. Pattern Anal. Mach. Intell. **42**(10), 2423–2436 (2019)
7. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: IEEE Conf. Comput. Vis. Pattern Recog. (2012)
8. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 270–279 (2017)
9. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Int. Conf. Comput. Vis. pp. 3828–3838 (2019)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
11. Guo, Y., Jiang, H., Qi, X., Xie, J., Xu, C.Z., Kong, H.: Unsupervised visible-light images guided cross-spectrum depth estimation from dual-modality cameras. arXiv preprint arXiv:2205.00257 (2022)
12. Han, G., Ma, J., Huang, S., Chen, L., Chang, S.F.: Few-shot object detection with fully cross-transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5321–5330 (2022)
13. Hu, M., Wang, S., Li, B., Ning, S., Fan, L., Gong, X.: Penet: Towards precise and efficient image guided depth completion. In: IEEE Int. Conf. Robotics and Automation. pp. 13656–13662. IEEE (2021)
14. Ke, L., Li, X., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F.: Prototypical cross-attention networks for multiple object tracking and segmentation. Adv. Neural Inform. Process. Syst. **34**, 1192–1203 (2021)
15. Kim, J., Noh, J., Jeong, M., Lee, W., Park, Y., Park, J.: Adnet: Non-local affinity distillation network for lightweight depth completion with guidance from missing lidar points. IEEE Robotics and Automation Letters (2024)
16. Lee, S., Park, J., Park, J.: Crossformer: Cross-guided attention for multi-modal object detection. Pattern Recognition Letters (2024)
17. Li, R., Gong, D., Yin, W., Chen, H., Zhu, Y., Wang, K., Chen, X., Sun, J., Zhang, Y.: Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21539–21548 (2023)

18. Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F.X., Taylor, R.H., Unberath, M.: Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In: Int. Conf. Comput. Vis. pp. 6197–6206 (2021)
19. Liang, M., Guo, X., Li, H., Wang, X., Song, Y.: Unsupervised cross-spectral stereo matching by learning to synthesize. In: AAAI. vol. 33, pp. 8706–8713 (2019)
20. Lin, Y., Cheng, T., Zhong, Q., Zhou, W., Yang, H.: Dynamic spatial propagation network for depth completion. In: AAAI. vol. 36, pp. 1638–1646 (2022)
21. Ma, F., Karaman, S.: Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In: IEEE Int. Conf. Robotics and Automation. pp. 4796–4803. IEEE (2018)
22. 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)
23. Park, J., Joo, K., Hu, Z., Liu, C.K., So Kweon, I.: Non-local spatial propagation network for depth completion. In: Eur. Conf. Comput. Vis. pp. 120–136. Springer (2020)
24. Rho, K., Ha, J., Kim, Y.: Guideformer: Transformers for image guided depth completion. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6250–6259 (2022)
25. Shen, Z., Dai, Y., Song, X., Rao, Z., Zhou, D., Zhang, L.: Pcw-net: Pyramid combination and warping cost volume for stereo matching. In: Eur. Conf. Comput. Vis. pp. 280–297. Springer (2022)
26. Shin, U., Park, J., Kweon, I.S.: Deep depth estimation from thermal image. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1043–1053 (2023)
27. Su, Q., Ji, S.: Chitransformer: Towards reliable stereo from cues. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1939–1949 (2022)
28. Tang, J., Tian, F.P., Feng, W., Li, J., Tan, P.: Learning guided convolutional network for depth completion. IEEE Trans. Image Process. **30**, 1116–1129 (2020)
29. 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)
30. Wang, T.H., Hu, H.N., Lin, C.H., Tsai, Y.H., Chiu, W.C., Sun, M.: 3d lidar and stereo fusion using stereo matching network with conditional cost volume normalization. In: IEEE/RSJ Int. Conf. Intell. Robots and Systems. pp. 5895–5902. IEEE (2019)
31. Wang, Y., Lai, Z., Huang, G., Wang, B.H., Van Der Maaten, L., Campbell, M., Weinberger, K.Q.: Anytime stereo image depth estimation on mobile devices. In: IEEE Int. Conf. Robotics and Automation. pp. 5893–5900. IEEE (2019)
32. Wang, Y., Li, B., Zhang, G., Liu, Q., Gao, T., Dai, Y.: Lrru: Long-short range recurrent updating networks for depth completion. In: Int. Conf. Comput. Vis. pp. 9422–9432 (2023)
33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
34. Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B., Revaud, J.: Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In: Int. Conf. Comput. Vis. pp. 17969–17980 (2023)
35. Xu, G., Cheng, J., Guo, P., Yang, X.: Attention concatenation volume for accurate and efficient stereo matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12981–12990 (2022)

36. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21919–21928 (2023)
37. Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D., Geiger, A.: Unifying flow, stereo and depth estimation. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
38. Xu, H., Zhang, J.: Aanet: Adaptive aggregation network for efficient stereo matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1959–1968 (2020)
39. Xu, Y., Zhu, X., Shi, J., Zhang, G., Bao, H., Li, H.: Depth completion from sparse lidar data with depth-normal constraints. In: Int. Conf. Comput. Vis. pp. 2811–2820 (2019)
40. Yan, Z., Lin, Y., Wang, K., Zheng, Y., Wang, Y., Zhang, Z., Li, J., Yang, J.: Tri-perspective view decomposition for geometry-aware depth completion. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4874–4884 (2024)
41. Yang, Y., Wong, A., Soatto, S.: Dense depth posterior (ddp) from single image and sparse range. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3353–3362 (2019)
42. 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)
43. Zhang, Y., Wang, L., Li, K., Fu, Z., Guo, Y.: Slfnet: A stereo and lidar fusion network for depth completion. IEEE Robotics and Automation Letters **7**(4), 10605–10612 (2022)
44. Zhi, T., Pires, B.R., Hebert, M., Narasimhan, S.G.: Deep material-aware cross-spectral stereo matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1916–1925 (2018)