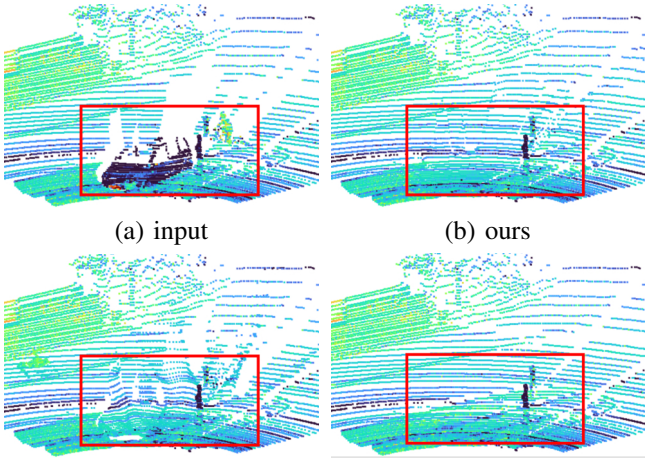
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

M²G-Net: Multimodal Mutual-Guidance Network for LiDAR Depth and Intensity Completion

Donghyun Choi, Sangmin Lee, and Jee-Hwan Ryu

***Abstract*— Autonomous driving has rapidly advanced with diverse sensors, especially Light Detection and Ranging (LiDAR), which provides precise geometry for tasks like simultaneous localization and mapping (SLAM). Recently, the performance of LiDAR-based SLAM has improved through studies leveraging intensity as a complementary cue to depth. However, in urban environments, dynamic objects occlude static scenes, degrading the stability and accuracy of LiDAR-based SLAM. While previous studies have focused mainly on completing occluded depth, they often disregard intensity, assuming it to be less critical or difficult to estimate due to inherent noise. This overlooks the strong complementary relationship between the two modalities, which can be exploited for effective multimodal completion. Furthermore, completing intensity alongside depth enables broader applicability to intensity-aware perception tasks. To address this issue, a Multimodal Mutual-Guidance (M^2G) module is proposed for the joint completion of occluded depth and intensity in LiDAR data. M^2G is integrated into a deep learning-based network that takes projected range and intensity images as input, enabling progressive cross-modal feature interaction. Leveraging the shared origin of LiDAR depth and intensity, M^2G balances noisy intensity and smooth depth via attention and structure-aware guidance. Experimental results demonstrate that the proposed method outperforms existing inpainting and depth completion approaches, validating its effectiveness for LiDAR completion.**



(a) input (b) ours (c) baseline (d) ground truth

Fig. 1: 3D point cloud completion from a range image. The proposed method more accurately completes regions occluded by dynamic objects (red boxes) compared to the baseline.

I. INTRODUCTION

Recent advancements in autonomous driving have progressed alongside the development of multiple sensing technologies, including cameras, radars, and Light Detection and Rangings (LiDARs). Among these, LiDAR provides robust, illumination-independent 3D geometry, which is crucial for reliable perception, high-definition mapping, and SLAM [1], [2]. Building on LiDAR's critical role in autonomous driving, recent studies have increasingly leveraged not only LiDAR depth but also intensity measurements, leading to enhanced performance in LiDAR-based SLAM, since intensity provides complementary cues by capturing surface reflectance and texture details that depth alone cannot convey [3]–[6].

However, LiDAR-based SLAM struggles in urban environments where dynamic objects like vehicles and pedestrians introduce noise and instability to pose estimation and mapping [7]–[9]. To mitigate these issues, recent studies have proposed methods for removing and filtering dynamic objects from LiDAR data [10]–[12]. Yet, removing dynamic points can reduce spatial cues for pose estimation, and filtering methods often fail to generalize in complex scenes [13]. Therefore, completing regions occluded by dynamic objects with static information is necessary to preserve spatial continuity and enhance LiDAR-based SLAM.

Various methods based on RGB and depth have been studied in computer vision to reconstruct occluded regions in visual scenes. RGB inpainting [14]–[16] focuses on reconstructing explicitly masked regions using surrounding visual context, whereas depth completion [17]–[19] aims to precisely estimate dense depth maps from sparse inputs, often using aligned RGB images as a guidance. More recently, RGB-D inpainting approaches that jointly reconstruct both RGB and depth have also been explored, aiming to exploit the complementary nature of the two modalities to generate structurally sound and visually plausible results [20]–[23]. These methods demonstrate that leveraging cross-modal relationships can significantly enhance reconstruction performance compared to unimodal approaches. However, these RGB and RGB-D inpainting methods typically rely on abundant RGB features, which limits their direct applicability to LiDAR data, since cameras often suffer from field of view mismatch with LiDARs and sensitivity to varying illumination conditions.

This research was supported in part by the Robot Industry Core Technology Development Program (20023294) and in part by the Robot Industry Core Technology Development Program (00423853) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea), and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-02213804). (*Corresponding author: Jee-Hwan Ryu.*)

Donghyun Choi, Sangmin Lee, and Jee-Hwan Ryu are affiliated with the Department of Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Korea (e-mail: choi23@kaist.ac.kr; iismn@kaist.ac.kr; jhryu@kaist.ac.kr)

In contrast, LiDAR-based inpainting methods [8], [9] have been proposed to reconstruct regions occluded by dynamic objects. However, most LiDAR inpainting approaches focus solely on depth, possibly assuming that intensity is either less critical or difficult to reconstruct, given the limitations of intensity measurement due to inherent noise [5], [24]. Although LiDAR intensity has seen growing use in recent SLAM and perception studies, complementary information from intensity is often overlooked, limiting applicability in downstream LiDAR-based tasks. Completing both depth and intensity jointly can not only improve the completion performance through their complementary relationship but also extend applicability to intensity-aware tasks. Moreover, although prior works adopt the term inpainting, LiDAR tasks require precise estimation of physical measurements. Therefore, the task needs to be defined as a completion, and this term is used throughout this paper.

To this end, a Multimodal Mutual-Guidance (M^2G) network is proposed, which jointly completes LiDAR depth and intensity by enabling mutual guidance through the co-located nature of both modalities, without relying on external RGB images. The contributions of the network are as follows:

1) Mutual modality guidance: The proposed method enhances completion quality by enabling mutual guidance, leveraging each modality's unique characteristics and their structural alignment.
2) Superior performance on LiDAR completion: The proposed method outperforms existing baselines, as shown in Fig. 1. Details of this visual comparison are provided in Section IV-B.

II. Related Works

A. Image Inpainting

Image inpainting techniques in computer vision have evolved from traditional approaches to deep learning-based methods. Traditional methods, including both diffusion [25]–[28] and patch-based approaches [29]–[32], focused on propagating information from surrounding pixels or patches, but struggled with large or semantically complex occluded regions. Deep learning has led to significant advances in image inpainting. Early works like Context Encoders [33] introduced adversarial training [34], while later methods improved robustness to irregular masks using gated or partial convolutions [16], [35]. More recent approaches incorporate structural priors [36], perceptual loss [37], [38], or semantic guidance [39], focusing on visual realism in RGB images. However, directly adopting image inpainting for LiDAR completion is challenging, as depth and intensity exhibit fundamentally different characteristics from RGB images, provide limited semantic cues, and their complementary information is difficult to leverage with unimodal methods.

B. Depth Completion

Depth completion refers to the task of completing missing regions in depth images or the accurate estimation of dense depth from sparse measurements. Many studies in this area have proposed methods that leverage additional modalities, such as RGB images, to guide the depth completion process [17]–[19]. These auxiliary modalities provide strong cues, as RGB images contain abundant visual information including textures and edges, which help infer smooth surfaces and accurate depth boundaries. In contrast, unguided methods relying solely on depth data tend to suffer from blurring and distortions around object boundaries. Therefore, guided approaches that exploit complementary information from other modalities are generally more robust and reliable [40]. However, methods that leverage abundant RGB cues to guide sparse depth are less suitable for LiDAR completion, where depth and intensity provide the same limited information.

C. RGB-D Inpainting

In addition to studies that inpaint a single modality, some studies have explored simultaneous inpainting of multiple modalities [20]–[23]. These studies, primarily conducted on RGB-D camera platforms, show that RGB and depth can effectively complement and guide each other, assuming a strong correlation exists between the two modalities [20]. For example, DynaFill [21] and SynerFill [22] proposed convolutional networks for RGB-D inpainting in autonomous driving scenarios, while MCD-Net [23] proposed an attention-based method that jointly inpaints RGB and depth in a coarse-to-fine manner. While RGB-D inpainting models effectively leverage the abundant textures of 3-channel RGB images and dense depth maps, LiDAR intensity differs from RGB in being single-channel, noisy, and lacking semantic cues. To address these limitations, effective LiDAR completion requires additional strategies to compensate for the limited information in LiDAR intensity.

D. LiDAR Inpainting

While much of the existing research on inpainting has focused on RGB or RGB-D data, recent interest in LiDAR inpainting has grown due to the importance in addressing challenges caused by dynamic objects in LiDAR scans [7]–[9]. To address these issues, SAM-Net [9] proposed a framework that jointly segments dynamic objects and inpaints occluded regions in LiDAR range images. The method estimates both a binary mask of dynamic regions and the static scene, focusing on geometric data rather than RGB imagery. A Structural Attention Module (SAM) is introduced to capture structural information and to inpaint depth more effectively. However, SAM-Net focuses only on depth and does not leverage the complementary benefits of intensity for more effective completion, which also limits applicability to intensity-aware downstream tasks.

III. Proposed Method

A. Problem Formulation

Let $X^i \in \mathbb{R}^{5\times H\times W}$ and $X^d \in \mathbb{R}^{5\times H\times W}$ denote the input images corresponding to the intensity and depth modalities, respectively, where H and W are the image height and width. Each five-channel input is constructed by augmenting the raw LiDAR data with additional information, as detailed in Section III-C.

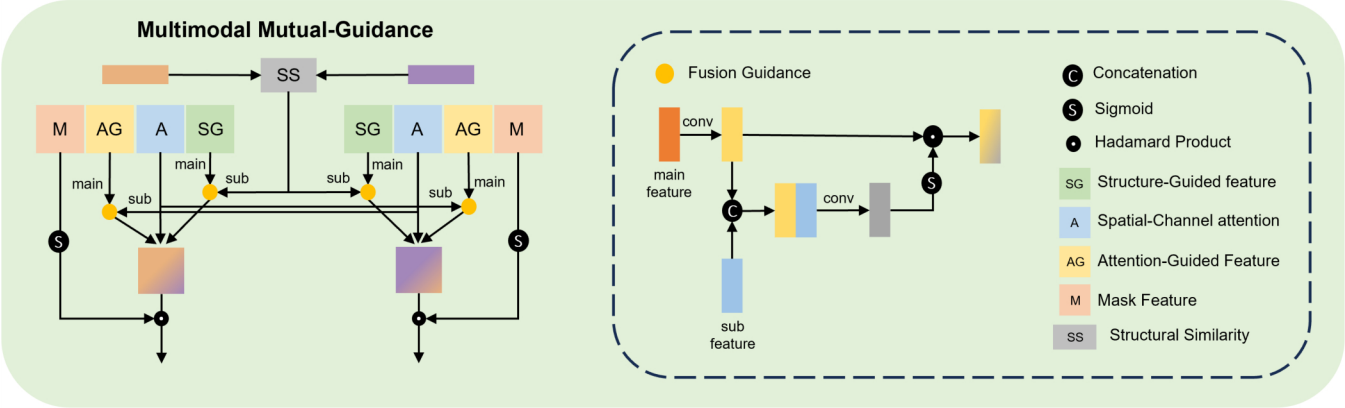


Fig. 2: Overall architecture of the M^{2}G. M^{2}G fuses depth and intensity features through structure-aware guidance, spatial-channel attention, and fusion cues at each encoding stage.

Given a binary mask $M \in \mathbb{R}^{1\times H\times W}$ indicating a dynamic object region (where 1 denotes masked pixels and 0 denotes background), the masked inputs are defined as:

$$X^i_m = X^i \odot (1 - M), \quad X^d_m = X^d \odot (1 - M), \tag{1}$$

where $\odot$ denotes Hadamard product and M is broadcasted across the channel dimension. The objective of the completion task is to estimate the masked regions in both modalities by learning the conditional distributions:

$$p(\hat{Y}^i \mid X^i_m), \quad p(\hat{Y}^d \mid X^d_m), \tag{2}$$

that approximate the true data distributions:

$$p(Y^i \mid X^i_m), \quad p(Y^d \mid X^d_m). \tag{3}$$

Note that although both inputs are 5-channel images, the desired outputs are modality-specific: $\hat{Y}^i \in \mathbb{R}^{1\times H\times W}$ for the completed intensity image and $\hat{Y}^d \in \mathbb{R}^{3\times H\times W}$ for the completed (x, y, z) coordinate images. Our network learns to complete static scenes by estimating the masked dynamic regions in a physically consistent and cross-modality-aware manner.

B. M²G: Multimodal Mutual-Guidance

Intensity and depth images, derived from the same LiDAR point cloud, provide different types of information: intensity as texture and depth as structure, while preserving structural consistency. Unlike prior rich-to-poor guidance methods, the Multimodal Mutual-Guidance (M^{2}G) module is proposed to enable symmetric, bidirectional feature enhancement between equally informative intensity and depth modalities. As shown in Fig. 2, M^{2}G operates in two steps: (1) extracting three feature types per modality and (2) fusing the extracted features via a fusion guidance module. Given intensity and depth features $F^i, F^d \in \mathbb{R}^{C\times H'\times W'}$, the following features are extracted for each modality. First, structure-guided features are convolutional features specifically extracted to facilitate later fusion guided by structure-aware information. Here, the structural similarity map S is computed by applying only the structural component of the Structural Similarity Index Measure (SSIM) [41] to F^i and F^d, leveraging the shared structural characteristics of both modalities. Second, spatial-channel attention features are applied to emphasize important spatial and channel-wise information within each modality. Finally, attention-guided features are convolutional features extracted to be guided later by attention maps from the other modality. Structure-guided and attention-guided features, which are initially convolutional features intended for subsequent guidance, are refined via a learnable fusion guidance function $FG(\cdot)$ that modulates each feature based on relevant cues. Specifically, the structure-guided features F^i_s and F^d_s are refined using the structural similarity map S:

$$F^i_{s,f} = FG(S, F^i_s), \quad F^d_{s,f} = FG(S, F^d_s), \tag{4}$$

where $F^i_{s,f}$ and $F^d_{s,f}$ denote the fused features. Similarly, the attention-guided features F^i_g and F^d_g are refined using spatial-channel attention features from the opposite modality, F^d_a and F^i_a, respectively:

$$F^i_{g,f} = FG(F^d_a, F^i_g), \quad F^d_{g,f} = FG(F^i_a, F^d_g), \tag{5}$$

where $F^i_{g,f}$ and $F^d_{g,f}$ are the fused features. The spatial-channel attention features F^i_a and F^d_a are computed using CBAM [42]. Finally, in each branch, the fused structure-guided features $F^i_{s,f}$ and $F^d_{s,f}$, fused attention-guided features $F^i_{g,f}$ and $F^d_{g,f}$, and the spatial-channel attention features F^i_a and F^d_a are concatenated and passed through gated convolutions to produce the next-level feature representations. This design combines both internal and complementary modalities, enhancing consistency and accuracy of completion.

C. Overall Network

The proposed network adopts a dual-branch architecture based on gated convolution, where each branch independently processes either an intensity or a depth image to reflect distinct characteristics of each modality. Each branch follows a two-stage structure: a coarse completion network and a refine completion network, both implemented as encoder–decoder architectures. Both stages use the coarse inpainting network from [16] as the backbone. In the coarse

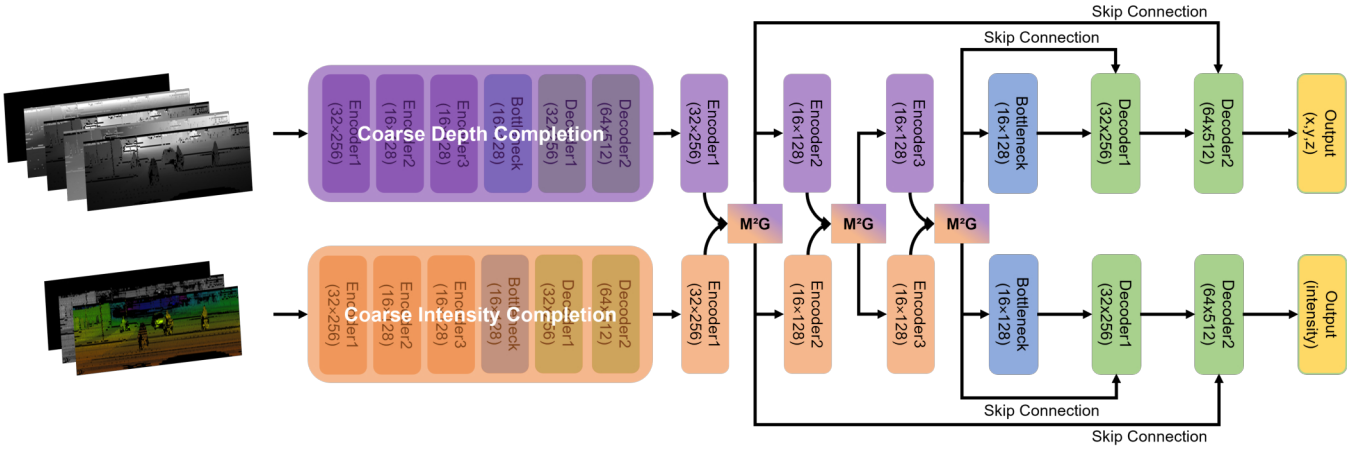
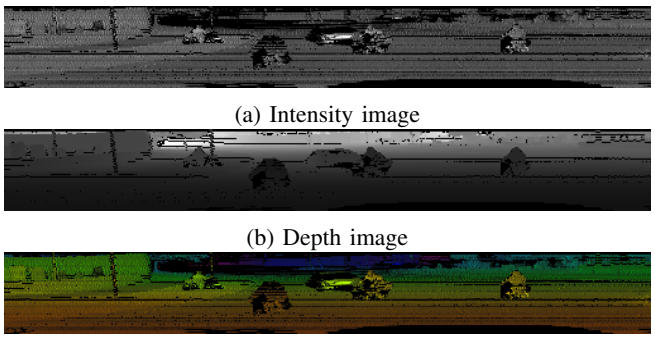


Fig. 3: Overview of the proposed architecture. The network takes two 5-channel inputs: (range, x, y, z, mask) for depth and (RGB, intensity, mask) for intensity, where RGB is derived from HSV. The M^{2}G module in the encoder fuses features and connects to the decoder via skip connections. The outputs are 3-channel point cloud coordinates and 1-channel intensity.



(a) Intensity image

(b) Depth image

(c) RGB input

Fig. 4: Example of RGB input generated from HSV representation of depth and intensity images. H: normalized depth, S: offset intensity (0.5 + intensity), V: intensity.

stage, the coarse inpainting network is applied as is to produce an initial estimation. In the refine stage, the backbone is restructured with three encoders, one bottleneck, and two decoders connected by skip connections (Fig. 3), enabling the M^{2}G module to facilitate cross-modal feature interaction.

In the depth branch, the input is a 5-channel image: one channel for the LiDAR range and three for the corresponding 3D coordinates (x, y, z). The coarse completion network produces an intermediate four-channel output without mask, which is refined by the refine network into a three-channel image containing only (x, y, z). This design aims to complete depth and estimate precise point-wise coordinates.

In the intensity branch, the input is a 5-channel image combining the original intensity image with a synthesized RGB image. Since intensity alone is noisy, an auxiliary RGB image is generated by transforming intensity and depth into an HSV representation. Fig. 4 shows how the enriched input preserves both depth structure and intensity variation. As in the depth branch, the coarse completion network produces an intermediate output, refined into the final 1-channel image.

D. Loss Function

The overall training objective combines loss terms from the depth and intensity branches:

$$\mathcal{L} = \mathcal{L}^d + \mathcal{L}^i, \tag{6}$$

where each branch uses a set of losses tailored to the specific characteristics of the corresponding modality.

a) Depth Completion Loss: The depth branch completes the 3D scene by estimating real-world coordinates (x, y, z). To ensure both geometric accuracy and spatial coherence, the loss for this branch consists of four terms:

$$\mathcal{L}^d = \lambda_1 \mathcal{L}^d_r + \lambda_2 \mathcal{L}^d_c + \lambda_3 \mathcal{L}^d_s + \lambda_4 \mathcal{L}^d_g. \tag{7}$$

The first term, $\mathcal{L}^d_r$, is a pixel-wise L_1 reconstruction loss measuring the absolute error between the prediction $\hat{Y}^d$ and ground truth Y^d, normalized by a rectified mask R to exclude invalid pixels:

$$\mathcal{L}^d_r = \frac{\|(\hat{Y}^d - Y^d) \odot R\|_1}{\|R\|_1}, \tag{8}$$

where $R[i, j]$ is defined as:

$$R[i,j] = \begin{cases} 0, & M[i,j] > 0 \text{ and } Y[i,j] = 0 \\ M[i,j], & \text{otherwise.} \end{cases} \tag{9}$$

The second term, the depth loss $\mathcal{L}^d_c$, enforces 3D consistency by minimizing the depth error between predicted and ground truth 3D points:

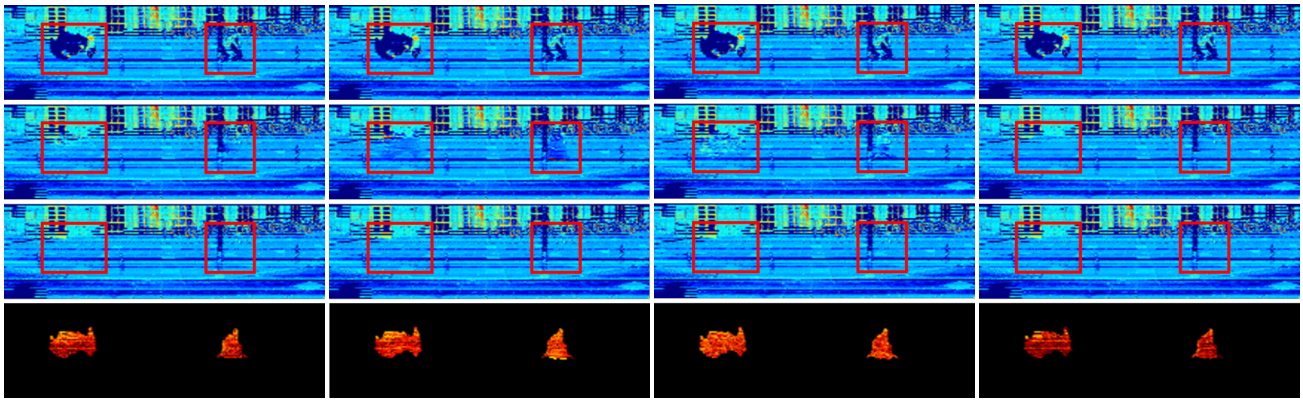
$$\mathcal{L}^d_c = \frac{\|(\hat{D} - D) \odot R\|_1}{\|R\|_1}, \quad D_{i,j} = \sqrt{x^2_{i,j} + y^2_{i,j} + z^2_{i,j}}. \tag{10}$$

To promote smoothness and suppress artifacts, a smoothness loss is adopted by comparing the spatial gradients of $\hat{D}$ and D over the rectified mask R:

$$\mathcal{L}^d_s = \frac{\|(\nabla\hat{D} - \nabla D) \odot R\|_1}{\|R\|_1}. \tag{11}$$

TABLE I: Quantitative Evaluation on LiDAR Inpainting Dataset.

Method	Depth				Intensity				Time[s]
	RMSE[m]	MAE[m]	iRMSE[1/m]	iMAE[1/m]	RMSE[m]	MAE[m]	PSNR	SSIM	
GatedConv [16]	2.8664	1.3537	0.0508	0.0109	0.1129	0.0807	19.1809	0.9649	0.0225
Latent-PIC [15]	4.2056	3.0516	0.0520	0.0307	0.1377	0.1050	17.4165	0.9528	0.1696
SPGAN [17]	3.5313	1.9008	0.1731	0.0273	–	–	–	–	0.0391
MCD-Net [23]	1.7599	1.0347	0.0074	0.0049	0.1182	0.0895	18.7189	0.9578	0.0411
MCD-Net* [23]	1.5792	0.9129	0.0068	0.0043	0.1255	0.0964	18.1868	0.9558	0.0340
Ours	**1.2103**	**0.5976**	**0.0052**	**0.0028**	**0.0886**	**0.0623**	**21.3276**	**0.9687**	0.0376



(a) GatedConv [16] (b) Latent-PIC [15] (c) MCD-Net* [23] (d) Ours

Fig. 5: Intensity completion comparison between the proposed network and other inpainting methods. The columns (a)–(d) represent the outputs from different models. From top to bottom, the rows show: input intensity, completed intensity, ground truth intensity, and the error map. In the error maps, brighter red indicates higher intensity error.

Finally, to improve the perceptual realism of depth images, an adversarial loss inspired by SN-GANs [43] hinge loss is included:

$$\mathcal{L}_g^d = -\mathbb{E}_{\hat{Y}^d \sim p_G}[D^{sn}(\hat{Y}^d)], \tag{12}$$

where D^{sn} is a discriminator with spectral normalization.

b) Intensity Completion Loss: The intensity branch aims to reconstruct texture details with accurate values. The loss for the intensity branch integrates three objectives:

$$\mathcal{L}^i = \lambda_5 \mathcal{L}_r^i + \lambda_6 \mathcal{L}_s^i + \lambda_7 \mathcal{L}_g^i. \tag{13}$$

Similar to the depth branch, the reconstruction loss $\mathcal{L}_r^i$ is defined as:

$$\mathcal{L}_r^i = \frac{\|(\hat{Y}^i - Y^i) \odot R\|_1}{\|R\|_1}. \tag{14}$$

To enhance texture realism, the style loss $\mathcal{L}_s^i$ minimizes the L_1 norm between the Gram matrices derived from VGG-extracted features of both the prediction and ground truth:

$$\mathcal{L}_s^i = \sum_l \|G_l(\hat{Y}^i) - G_l(Y^i)\|_1. \tag{15}$$

Finally, the adversarial loss $\mathcal{L}_g^i$ follows the same formulation as in the depth branch:

$$\mathcal{L}_g^i = -\mathbb{E}_{\hat{Y}^i \sim p_G}[D^{sn}(\hat{Y}^i)]. \tag{16}$$

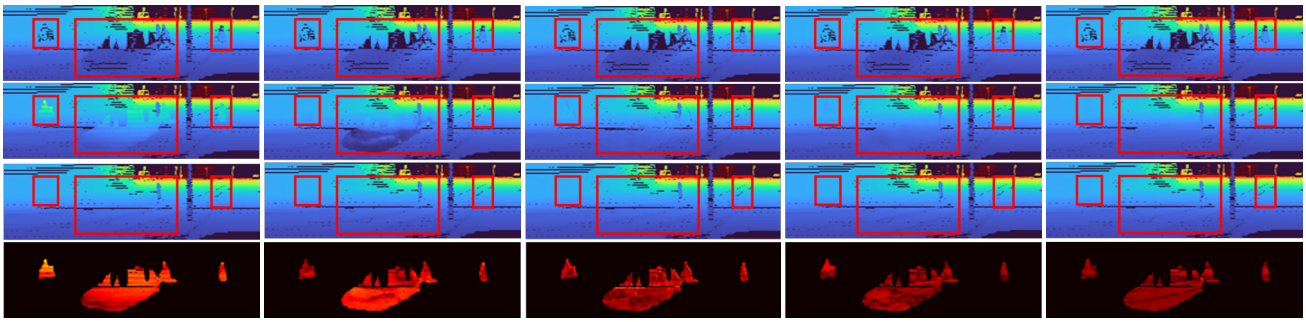
Together, these loss components encourage the network to generate geometrically accurate, visually coherent, and perceptually realistic completions for both depth and intensity modalities.

IV. EXPERIMENTS

A. Experimental Setup

The experiments in this study are divided into two main categories: (1) comparative evaluation of LiDAR completion performance against existing methods, and (2) ablation studies to verify the effectiveness of the proposed M^{2}G module. The LiDAR inpainting dataset introduced by SAM-Net [9] is used for training and testing, implemented in PyTorch [44]. All experiments are conducted on a workstation equipped with four NVIDIA TITAN V (12GB) GPUs and an Intel Xeon W-2133 CPU. Training proceeds for 100 epochs per model with batch size 8, applying the Adam optimizer at learning rate 1e$-$4. The loss weights for depth and intensity branches are empirically set to balance each objective: $\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 0.05$, $\lambda_4 = 0.1$ for depth, and $\lambda_5 = 1.0$, $\lambda_6 = 0.2$, $\lambda_7 = 0.1$ for intensity. To improve convergence and completion accuracy, depth values are preprocessed by filtering out extreme outliers based on a maximum range threshold. Intensity values are inherently represented in the [0, 1] range. Note that SAM-Net is excluded from direct comparison as its official implementation is not publicly available, and the preprocessing pipeline also differs.

For quantitative evaluation, standard metrics are adopted for each modality. For depth completion, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), inverse RMSE (iRMSE), and inverse MAE (iMAE) are measured. For intensity completion, RMSE, MAE, Peak Signal to Noise Ratio (PSNR), and SSIM are measured.



(a) GatedConv [16] (b) Latent-PIC [15] (c) SPGAN [17] (d) MCD-Net* [23] (e) Ours

Fig. 6: Depth completion comparison between the proposed network and other methods. The columns (a)–(d) represent the outputs from different models. From top to bottom, the rows show: input depth, completed depth, ground truth depth, and the error map. In the error maps, brighter red indicates higher depth error.

B. Comparison

The proposed method is compared with four baseline models: two RGB image inpainting methods, Gated Convolution [16] and Latent-PIC [15], one unguided depth completion method, SPGAN [17], and one RGB-D inpainting method, MCD-Net [23]. Since MCD-Net originally takes 3-channel RGB image and a 1-channel depth image as input, two variants are evaluated: one with intensity (replicated to 3 channels) and depth, and another modified version (denoted MCD-Net*) that matches our input structure. To ensure a fair comparison, GatedConv and Latent-PIC are trained separately on intensity and depth images, as these methods are designed for single-modality inpainting. SPGAN is trained only on the depth images, while MCD-Net and MCD-Net* are trained jointly on intensity and depth inputs.

As shown in Table I, Fig. 5, and Fig. 6, M^{2}G-Net outperforms all baselines on both depth and intensity modalities. Even compared with the strongest baseline, MCD-Net*, M^{2}G-Net achieves 35–40% lower MAE for depth and lower RMSE and MAE for intensity, while iMAE shows up to a 50% improvement, highlighting more accurate completion in short-range regions. To further assess 3D completion quality, range images are converted into point clouds, as shown in Fig. 1. Although RMSE gains are modest, this 3D evaluation demonstrates that MCD-Net* (corresponding to panel (c)) introduces spatial distortions, misaligning points in x, y, and z, indicating limited geometric consistency. In contrast, M^{2}G-Net preserves structural integrity across the scene, including occluded and distant areas, reconstructing all masked regions with higher fidelity. The model achieves approximately 26.6 Hz inference speed with 13.87M parameters (M²G modules: 1.69M, 12.2%), enabling real-time autonomous deployment.

C. Ablation Study

To evaluate the effectiveness of the M^{2}G module and additional RGB input (converted from HSV), an ablation study is conducted under four configurations by varying the use of M^{2}G and the RGB input. Specifically, we consider cases without either component, with only the RGB input, with only M^{2}G, and with both components combined. Each configuration is evaluated using RMSE and MAE for both the depth and intensity modalities. Table II summarizes the results. The proposed model (with M^{2}G and with RGB) achieves the best performance. Adding RGB input alone improves intensity but slightly degrades depth, suggesting an optimization imbalance. The M^{2}G module mitigates the imbalance through mutual-guidance, improving both modalities.

TABLE II: Ablation Study on the Proposed Network.

Configuration	Depth		Intensity	
	RMSE	MAE	RMSE	MAE
w/o M²G + w/o RGB	1.3287	0.6935	0.1221	0.0912
w/o M²G + w/ RGB	1.3774	0.7501	0.0908	0.0638
w/ M²G + w/o RGB	1.2723	0.6504	0.1081	0.0785
w/ M²G + w/ RGB	**1.2103**	**0.5976**	**0.0886**	**0.0623**

V. CONCLUSIONS

In this study, M^{2}G-Net is proposed as a network designed to enhance LiDAR-based SLAM in dynamic environments by completing the depth and intensity data of dynamic regions with static information. The network adopts a coarse-to-fine backbone with an M^{2}G module, which captures depth–intensity correlations during encoding and completes occluded regions through mutual-guidance. To enrich the intensity input, intensity and depth are transformed into an HSV representation to form RGB input, providing more informative cues. M^{2}G-Net outperforms existing methods by accurately completing (x, y, z) coordinates and intensities in dynamic regions and operates in real time, enabling practical deployment in SLAM systems. However, the proposed method assumes that dynamic object masks are given, which limits its applicability in fully autonomous settings. Developing a video-based framework that jointly estimates such masks while leveraging temporal information for improved completion performance remains as future work.

REFERENCES

[1] Y.-L. Zhao, Y.-T. Hong, and H.-P. Huang, "Comprehensive performance evaluation between visual slam and lidar slam for mobile robots: Theories and experiments," *Applied Sciences*, vol. 14, no. 9, p. 3945, 2024.

[2] P. Chen, X. Zhao, L. Zeng, L. Liu, S. Liu, L. Sun, Z. Li, H. Chen, G. Liu, Z. Qiao *et al.*, "A review of research on slam technology based on the fusion of lidar and vision," *Sensors*, vol. 25, no. 5, p. 1447, 2025.

[3] Y. S. Park, H. Jang, and A. Kim, "I-loam: Intensity enhanced lidar odometry and mapping," in *2020 17th international conference on ubiquitous robots (UR)*. IEEE, 2020, pp. 455–458.

[4] H. Wang, C. Wang, and L. Xie, "Intensity-slam: Intensity assisted localization and mapping for large scale environment," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1715–1721, 2021.

[5] W. Dai, S. Chen, Z. Huang, Y. Xu, and D. Kong, "Lidar intensity completion: Fully exploiting the message from lidar sensors," *Sensors*, vol. 22, no. 19, p. 7533, 2022.

[6] L. Lai, L. Li, H. Wang, J. Yuan, W. Fan, and D. Zhao, "Enhanced lidar-inertial slam with adaptive intensity feature extraction and fusion," *Measurement*, p. 117738, 2025.

[7] H. Bavle, J. L. Sanchez-Lopez, C. Cimarelli, A. Tourani, and H. Voos, "From slam to situational awareness: Challenges and survey," *Sensors*, vol. 23, no. 10, p. 4849, 2023.

[8] C. Han, I. M. P. A. Winata, and J. Oh, "Lidar point inpainting model using smoothness loss for slam in dynamic environments," *International Journal of Advanced Robotic Systems*, vol. 21, no. 6, p. 17298806241297424, 2024.

[9] J. Lee, S. Hwang, W. J. Kim, and S. Lee, "Sam-net: Lidar depth inpainting for 3d static map generation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 12 213–12 228, 2021.

[10] Z. Wang, Z. Zhang, X. Kang, M. Wu, S. Chen, and Q. Li, "Dor-lins: Dynamic objects removal lidar-inertial slam based on ground pseudo occupancy," *IEEE Sensors Journal*, vol. 23, no. 20, pp. 24 907–24 915, 2023.

[11] Y. Jia, T. Wang, F. Cao, X. Chen, S. Shao, and L. Liu, "Trlo: An efficient lidar odometry with 3d dynamic object tracking and removal," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[12] H. Peng, Z. Zhao, and L. Wang, "A review of dynamic object filtering in slam based on 3d lidar," *Sensors*, vol. 24, no. 2, p. 645, 2024.

[13] K. Wang, J. Guo, K. Chen, and J. Lu, "An in-depth examination of slam methods: Challenges, advancements, and applications in complex scenes for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[14] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2149–2159.

[15] H. Chen and Y. Zhao, "Don't look into the dark: Latent codes for pluralistic image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7591–7600.

[16] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4471–4480.

[17] M. F. F. Khan, N. D. Troncoso Aldas, A. Kumar, S. Advani, and V. Narayanan, "Sparse to dense depth completion using a generative adversarial network with intelligent sampling strategies," in *Proceedings of the 29th acm international conference on multimedia*, 2021, pp. 5528–5536.

[18] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys, "Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3313–3322.

[19] J. Gu, Z. Xiang, Y. Ye, and L. Wang, "Denselidar: A real-time pseudo dense depth guided depth completion network," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1808–1815, 2021.

[20] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multi-modal multi-task masked autoencoders," in *European Conference on Computer Vision*. Springer, 2022, pp. 348–367.

[21] B. Bešić and A. Valada, "Dynamic object removal and spatio-temporal rgb-d inpainting via geometry-aware adversarial learning," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 2, pp. 170–185, 2022.

[22] K. Liu, Y. Zhang, Y. Xie, L. Li, Y. Wang, and L. Chen, "Synerfill: A synergistic rgb-d image inpainting network via fast fourier convolutions," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 69–78, 2023.

[23] J. Hou, Z. Ji, J. Yang, C. Wang, and F. Zheng, "Mcd-net: toward rgb-d video inpainting in real-world scenes," *IEEE Transactions on Image Processing*, vol. 33, pp. 1095–1108, 2024.

[24] W. Y. Yan and A. Shaker, "Airborne lidar intensity banding: Cause and solution," *ISPRS journal of photogrammetry and remote sensing*, vol. 142, pp. 301–310, 2018.

[25] Levin, Zomet, and Weiss, "Learning how to inpaint from global image statistics," in *Proceedings Ninth IEEE international conference on computer vision*. IEEE, 2003, pp. 305–312.

[26] J. Shen and T. F. Chan, "Mathematical models for local nontexture inpaintings," *SIAM Journal on Applied Mathematics*, vol. 62, no. 3, pp. 1019–1043, 2002.

[27] J. Sun, L. Yuan, J. Jia, and H.-Y. Shum, "Image completion with structure propagation," in *ACM Siggraph 2005 Papers*, 2005, pp. 861–868.

[28] M.-J. Fadili, J.-L. Starck, and F. Murtagh, "Inpainting and zooming using sparse representations," *The Computer Journal*, vol. 52, no. 1, pp. 64–79, 2009.

[29] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. IEEE, 1999, pp. 1033–1038.

[30] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *IEEE transactions on image processing*, vol. 10, no. 8, pp. 1200–1211, 2001.

[31] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.

[32] K. He and J. Sun, "Image completion approaches using the statistics of similar patches," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 12, pp. 2423–2435, 2014.

[33] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[34] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[35] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 85–100.

[36] C. Cao, Q. Dong, and Y. Fu, "Zits++: Image inpainting by improving the incremental transformer on structural priors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 10, pp. 12 667–12 684, 2023.

[37] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[38] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," *Advances in neural information processing systems*, vol. 30, 2017.

[39] C. Zhang, W. Yang, X. Li, and H. Han, "Mmginpainting: Multi-modality guided image inpainting based on diffusion models," *IEEE Transactions on Multimedia*, vol. 26, pp. 8811–8823, 2024.

[40] J. Hu, C. Bao, M. Ozay, C. Fan, Q. Gao, H. Liu, and T. L. Lam, "Deep depth completion from extremely sparse data: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8244–8264, 2022.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[43] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[44] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.