2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

SLAM Based on Camera-2D LiDAR Fusion

Guoyu Lu

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10611546

Abstract— **The SLAM system plays a pivotal role in robotic mapping and localization, leveraging various sensor technologies to achieve precision. Traditional passive sensors, such as RGB cameras, offer high-resolution imagery at a lower cost for SLAM applications, yet they fall short in accurately estimating 3D positions and camera motions. On the other hand, LiDARs excel in generating accurate 3D maps but often come at a higher price and lower resolution. While active illumination sensors like LiDAR provide precise depth estimation, the prohibitive cost of high-resolution LiDAR systems restricts their widespread adoption across diverse applications. Although 2D single-beam LiDAR is more affordable, its limited depth sensing capability hampers comprehensive environmental perception. Addressing these limitations, this paper introduces a deep learning framework aimed at enhancing SLAM performance through the strategic fusion of camera and 2D LiDAR data. Our approach employs a novel self-supervised network alongside an economical single-beam LiDAR, striving to achieve or surpass the performance of more expensive LiDAR systems. The integration of single-beam LiDAR with our system allows for dynamic adjustment of scale uncertainty in depth maps generated by monocular camera systems within SLAM. Consequently, this fusion method enjoys the high-resolution and accuracy benefits of advanced LiDAR systems with the cost-effectiveness of 2D LiDAR sensors. Through this innovative combination, we demonstrate a SLAM system that not only maintains high fidelity in mapping and localization but also ensures affordability and broad applicability.**

I. Introduction

Simultaneous Localization and Mapping (SLAM) is a critical area of robotics, underpinning critical applications such as autonomous vehicles [1], [2], [3], interactive robotics [4], and 3D scene reconstruction [5], [6]. Its utility extends across diverse tasks including localization, mapping [7], [8], [9], and the facilitation of collaborative robotic interactions.

Traditionally, SLAM methodologies leverage both active sensors (e.g., RGB-D cameras and LiDAR) for direct depth estimation, and passive image matching techniques (e.g., stereo vision) to infer 3D spatial information. While RGB-D cameras offer utility in indoor settings within constrained ranges (e.g., 0.25-5.46m for Azure Kinect) due to IR signal sensitivity, their application is limited in outdoor environments. Conversely, LiDAR sensors, celebrated for their precision in depth measurement and extensive sensing range, have become staples in mobile robotics and autonomous driving. However, the adoption of high-resolution LiDAR systems is hindered by prohibitive costs (e.g., Ouster OS-2-64 LiDAR priced at approximately 24,000 USD) and the inherent sparsity of LiDAR-generated point clouds. The reliance on monocular cameras for passive depth estimation introduces scale ambiguity. The disparity in depth distribution and percentage errors for proximal scenes remains a challenge, underscoring the limitations of scale consistency in monocular visual SLAM. Though being able to estimate rough scale, a notable drawback in stereo vision systems is a fixed baseline for 3D reconstruction, limiting the valid range.

Current advances employing convolutional neural networks (CNNs) have not fully mitigated scale uncertainty in SLAM outputs relative to ground truth. While self-supervised learning schemes show promise in deducing relative depth maps and camera trajectories, achieving precise absolute depths and spatial positioning remains elusive. Moreover, supervised learning approaches necessitate densely labeled ground truth data, incurring large costs, and their accuracy is contingent upon the diversity of training scenarios. Such models often falter in untrained, novel environments, exacerbating inaccuracies in practical deployments.

In this paper, we introduce a cost-effective SLAM framework that synergistically combines a 2D single-beam LIDAR with a camera. Our approach is centered around a novel neural network architecture comprising a self-supervised CNN for single image depth estimation, and a multi-layer perceptron (MLP) network. The latter dynamically adjusts the depth map informed by 2D LiDAR inputs, enhancing the precision of the SLAM process. Specifically, we develop a deep learning model capable of extracting scene depth from consecutive images, facilitating the construction of 3D maps and accurate camera motion estimation. With sequential images, the camera trajectory is refined through a dedicated camera pose estimation network.

To overcome the limitations of conventional global scaling methods, we employ an MLP network trained to accurately predict the scale of depth maps derived from the CNN. This network leverages image pixels corresponding to single-beam LiDAR points for fine-tuning, subsequently applying these refined scales to enhance all pixel depth values across the image, thereby enabling the recovery of a comprehensive 3D map alongside the estimated camera trajectory. Our framework is rigorously evaluated on the public raw KITTI dataset, where we simulate single-beam LiDAR data from sparse LiDAR inputs to test our 3D maps and camera motion estimation methods.

The primary contributions of our work are fourfold: 1) We propose a feasible SLAM solution that achieves high-precision 3D mapping and camera motion estimation at low cost. 2) Our framework effectively extends the utility of 2D single-beam LiDAR to full image resolution, presenting a viable low-cost alternative to expensive high-beam LiDAR systems for accurate, real-world 3D mapping. 3) We introduce an innovative approach utilizing an MLP network to dynamically learn scaling factors for significantly improving the accuracy of 3D mapping in a variety of unseen

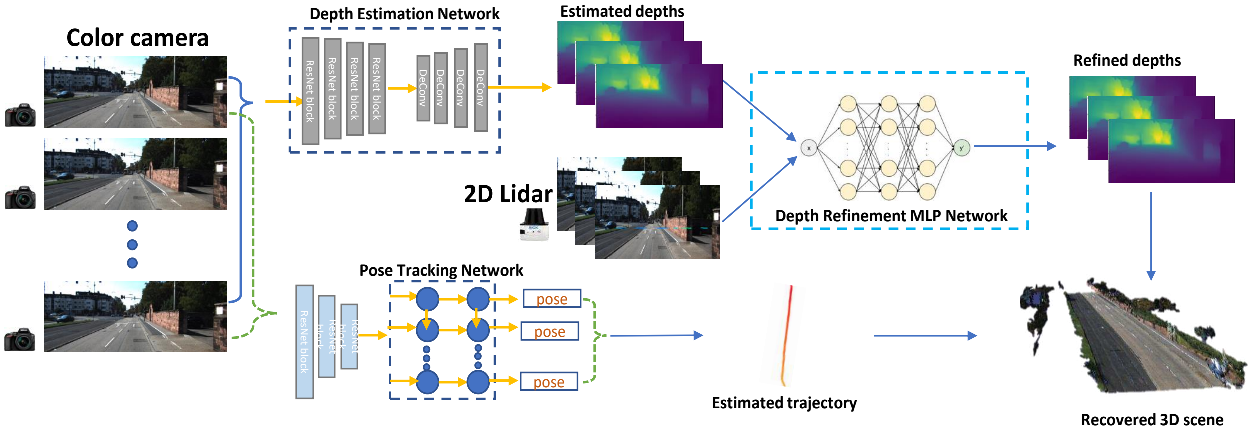

Fig. 1. An overview of our training framework designed for use with monocular sequential images. Our architecture incorporates a 2D LiDAR point array and sequential images as inputs. The system comprises a depth estimation network that generates initial depth maps. These initial maps, along with single-beam LiDAR points, are input into a refinement MLP network designed to dynamically adjust the depth scale within the image, resulting in refined depth maps. Additionally, our pose estimation component utilizes a CNN-RNN structure to estimate relative poses throughout the image sequence. The integration of estimated trajectory data with refined depths enables the reconstruction of a comprehensive and detailed 3D scene.

scenarios, which correspondingly corrects the scales of motion estimation and enhances its accuracy. 4) Our proposed model demonstrates robust performance on both monocular sequences and stereo sequences as input, showcasing its potential for broad application in self-supervised depth estimation strategies. The different training frameworks are illustrated in Figures 1 and 2.

II. Related Work

Self-supervised SLAM systems have made significant advances by learning depth estimation and camera egomotion from monocular video sequences through unsupervised methods [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. The incorporation of Generative Adversarial Networks (GANs) has further refined depth estimation accuracy [25], [26], [27]. Moreover, recent efforts have expanded to joint learning with ancillary tasks such as optical flow [15], keyframe detection [28], and modeling uncertainty [29], enhancing the robustness and utility of SLAM systems. A few pioneering studies have begun to address these challenges. Zhou et al. [18] introduced the use of optical flow as a supervisory signal specifically to mitigate the difficulties encountered in low-texture environments. Bian et al. [30] developed a weak image rectification strategy aimed at enhancing the effectiveness of unsupervised loss constraints for depth estimation in such challenging scenes. These efforts represent initial strides towards extending the applicability of SLAM systems to a broader range of environments, acknowledging the need for further research in this area.

Supervised depth prediction and SLAM have seen significant advancements [31], [32], [33], [34], [35], primarily through innovative feature representation extraction from images. Eigen et al. [36] utilized a multi-scale convolutional neural network, inspired by AlexNet [37], to progressively refine depth maps. Liu et al. [32] combined deep neural networks with continuous conditional random fields (CRF) to enhance depth information learning. Expanding on this, Qi et al. [33] explored geometric relationships between depth maps and surface normals using a dual-stream CNN approach. Concurrently, Demon et al. [35] developed an end-to-end network that simultaneously estimates scene depth and camera motion from image sequences, incorporating known optical flow. Luo et al. [38] introduced a network for single image depth estimation that leverages depth labels and stereo pairs, simulating one stereo view from another and adopting a fully-supervised training regimen akin to [39]. While these methods have shown the ability to generate plausible depth estimations from single images, their performance is notably constrained in unfamiliar scenarios. Furthermore, a critical limitation is the requirement for extensive labeled datasets, often exceeding 10,000 pixel-level, aligned ground truth depth images, which poses a significant challenge for many applications due to the difficulty in obtaining such comprehensive datasets.

Fusion of camera and LiDAR enhances depth perception through integrating 3D LiDAR sensing and imagery. Badino et al. [40] enhanced stereo matching by incorporating LiDAR data as a priori disparity estimates to constrain search regions. Maddern et al. [41] devised an efficient probabilistic model that merges 3D LiDAR data with stereo images to produce dense depth maps in real-time. More recently, Kihong et al. [42] introduced a two-stage cascade deep neural network that integrates 3D LiDAR and stereo disparity maps to yield high-precision disparities. However, these methods face limitations in areas lacking sparse LiDAR points, and the reliance on 3D LiDAR data for network inputs [40], [41] and training labels [42] incurs substantial costs, limiting their accessibility for mass market applications. To address these challenges, we propose an unsupervised convolutional neural network (CNN) model that estimates high-precision depth from a single image, complemented by a low-cost single-beam LiDAR to produce reliable, high-precision depth maps. These maps can be integrated with estimated camera trajectories from sequential images to reconstruct comprehensive 3D maps. Our approach represents the first to leverage an online training algorithm combining a single-beam LiDAR with a single camera in a unified network, achieving full-resolution, high-precision depth estimation and 3D mapping.

III. SLAM Based on Camera-2D LIDAR Fusion

We will first calibrate the camera and 2D LiDAR to fuse the sensing outputs following [43]. The process involves capturing a 2D image and a 3D point arrangement from the combined camera-LiDAR module. Calibration involves selecting sets of vertical lines on three calibration boards at varying distances, identifying LiDAR response points on these lines, and forming LiDAR-camera correspondence pairs.

Eighteen correspondences are established from three plane distances to estimate the extrinsic transformation between the camera and LiDAR using the Direct Linear Transform (DLT) method within a RANSAC loop. This results in a 6-DOF transformation matrix that converts 3D points from the LiDAR's world coordinate system to the camera's coordinate system, expressed through a specific relationship involving rotation matrices and translation vectors.

We then propose a 3D mapping system that incorporates this LiDAR-camera fusion with monocular sequential cues, as depicted in Fig. 1. This system includes a self-supervised network that updates the depth estimation network, depth refinement network, and pose tracking network. Inputs include a monocular RGB video and corresponding 2D LiDAR line points, with outputs comprising a globally consistent 6-DoF camera trajectory and refined estimated scene depth for complete 3D map recovery.

Leveraging sequential images, the combination of single-beam LiDAR with a normal camera allows for more accurate training of depth and motion estimation CNNs from unlabeled videos. Specifically, the pose estimation network, trained on adjacent local frames, regresses a group of relative poses, while the depth estimation network generates corresponding depth maps. These maps, along with 2D LiDAR points, are input into the depth refinement MLP to correct the local scale across depth ranges, with geometric constraints between refined depth maps and estimated relative motions guiding self-training and complete 3D scene recovery.

To address the challenge of scaling ambiguity in depth estimation, which varies with object distance, we integrate estimated dense disparity maps with 2D LiDAR points using a Multilayer Perceptron Neural Network (MLP). This approach, which contrasts with the direct application of a global scale factor common in existing methods, utilizes an MLP with a two hidden-layer structure to align predicted depth values with actual LiDAR measurements, effectively predicting local scale factors.

A. Multi-view Re-projection Loss

Given consecutive images I_t and I_{t+1}, alongside refined depth map D_t and camera motion $P_{t\to t+1}$, pixel correspondence is calculated to project pixels from target to reference images using the camera's intrinsic matrix K:

$$p_{t+1} = KP_{t\to t+1}\tilde{D}(p_t)K^{-1}p_t \tag{1}$$

Aiming for a seamless reconstruction, we determine the minimal photometric loss across frames by computing:

$$L_{photo} = \sum_{i=1}^{N} \min_{t'\in\{t-i,t+i\}} \rho(I_t, I_{t'\to t}) \tag{2}$$

where ρ combines L1 loss and SSIM for robust image reconstruction, defined as:

$$\rho(I_1, I_2) = \alpha\left(\frac{1-\text{SSIM}(I_1,I_2)}{2}\right)+(1-\alpha)\|I_1-I_2\|_1 \tag{3}$$

B. Moving Masking Strategy

To address scene dynamics, a masking strategy excludes regions with significant motion or occlusion, defined by depth inconsistency:

$$D_{diff}(p) = \frac{|D^t_{t+1}(p) - D'_{t+1}(p)|}{D^t_{t+1}(p) + D'_{t+1}(p)} \tag{4}$$

Here, D^t_{t+1} represents the depth map projected from I_{t+1} to I_t using estimated motion, and D'_{t+1} is the directly estimated depth for I_{t+1}. The moving mask, computed from depth inconsistency D_{diff}, reduces the influence of moving objects and occlusions on the learning process:

$$M_{moving} = 1 - D_{diff}(p) \tag{5}$$

This mask, M_{moving}, ranges from 0 to 1, assigning lower weights to pixels within moving or occluded regions to mitigate their impact on the estimation of camera poses and dense depth maps.

Given the occurrence of static frames in certain scenes, which could potentially affect the learning of camera motion, we implement an auto-masking technique to selectively compute photometric loss, thus filtering out points whose relative motion corresponds precisely to the camera's motion:

$$M_{auto} = \begin{cases} 1, & \text{if } \|I_t - I'_t\|_1 < \|I_t - I_{t+1}\|_1 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Here, M_{auto} acts as a binary mask, where I'_t is the image warped from I_{t+1} using the refined depth map and relative motion estimation. This auto-masking strategy effectively discriminates against static areas in the scene, ensuring that the training process focuses on areas with discernible relative motion, enhancing the accuracy of motion estimation.

Through these methodologies, our system proficiently fuses single-beam LiDAR data with monocular video inputs to train deep neural networks for accurate depth and motion estimation. This fusion, supported by innovative loss functions and masking strategies, allows for the effective reconstruction of detailed 3D scenes from relatively sparse and inexpensive LiDAR data, alongside commonly available video sequences. The synergistic use of geometric constraints and self-supervised learning paradigms facilitates the generation of globally consistent depth maps and camera trajectories, marking a significant advancement in the field of 3D scene reconstruction.

C. Refined Multi-view Re-projection

Implementing masking strategies enables us to refine the multi-view re-projection loss L by focusing on static scenes and minimizing attention to moving and occluded objects. The refined loss function $L_{refined}$ considers both moving and static masks to concentrate on relevant regions as:

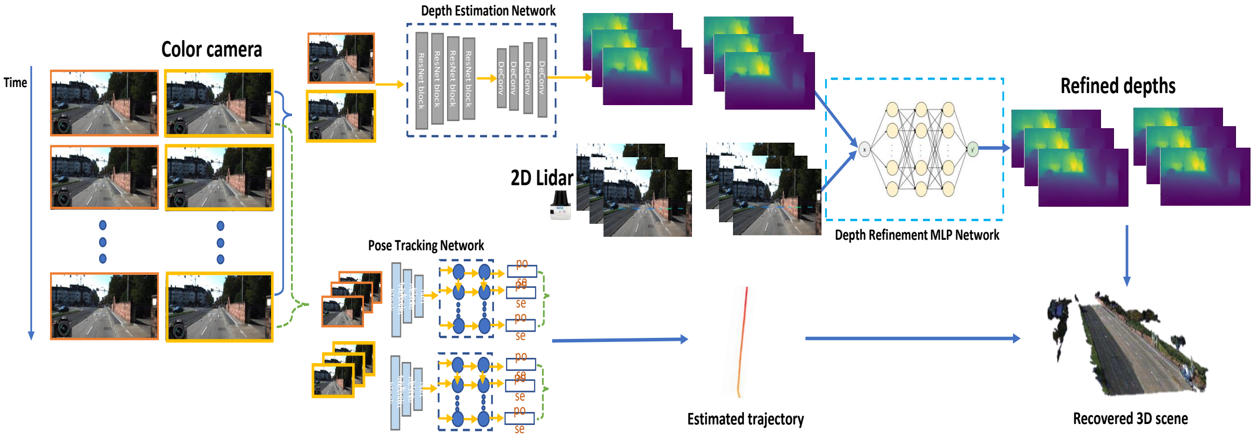

Fig. 2. Our training framework leverages 2D LiDAR point arrays and stereo images (left **L** and right **R**) as inputs, enabling the production of high-precision depth maps from just one image and a single LiDAR beam in real-world applications. This system integrates offline calibration between the 2D LIDAR and a color camera to establish pixel-point correspondences, utilizing a stereo camera setup for initial depth estimation training. In practice, depth maps are generated online from single images via a pre-trained system, with LiDAR points aligned using a pre-calibrated LiDAR-camera matrix. An MLP network, refined by aligning LiDAR depth points with image pixel depths, accurately adjusts depth scales within the image. This innovative approach effectively enhances 2D LiDAR's capabilities to full-resolution 3D mapping.

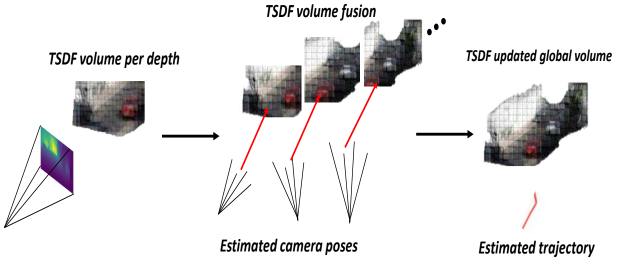

Fig. 3. Illustration of fusing the estimated depth maps into a global volume with the corresponding estimated camera poses.

$$L_{refined} = \sum_{i=1}^{N} M_{moving} \cdot M_{static} \cdot \min_{t' \in \{t-i, t+i\}} \rho(I_t, I_{t' \to t}) \tag{7}$$

This adjustment ensures that the network's focus is directed towards areas of the image less likely to be affected by dynamic changes or occlusions, thereby improving the robustness of depth estimation.

D. Dense Map Building from Multiple Depths

The construction of a dense map leverages the Truncated Signed Distance Function (TSDF) fusion method, integrating multiple estimated depths $\tilde{D}(t = 1, 2, ..., T)$, estimated poses $\tilde{P}_t$, and camera intrinsics K_t into a discretized signed distance function $S_t \in \mathbb{R}^{X \times Y \times Z}$ and corresponding weight function $W_t \in \mathbb{R}^{X \times Y \times Z}$. As an incremental process, each new depth map is assimilated using the update equations:

$$V_t(x) = \frac{W_{t-1}(x) \cdot V_{t-1}(x) + w_t(x) \cdot v_t(x)}{W_{t-1}(x) + w_t(x)} \tag{8}$$

$$W_t(x) = W_{t-1}(x) + w_t(x) \tag{9}$$

where V_t and W_t initiate from empty volumes V_0 and W_0. This process allows for the continuous updating of the signed distance v_t and the corresponding weight w_t, cumulatively integrating depth information over time into the final TSDF volume. Correspondingly, the input I_t to the depth fusion module is comprised of the estimated depth, the camera poses, and the camera intrinsics, mapped as:

$$I_t \longmapsto [\tilde{D}_t, \tilde{P}_t, C] \longmapsto [\tilde{D}_t, W_{t-1}, V_{t-1}] \tag{10}$$

The truncation depth distance is set to 80 meters for outdoor scenes and 10 meters for indoor scenarios, optimizing the process for different environments by mitigating depth noise and enhancing the efficiency of the fusion strategy. Fig. 3 illustrates the fusion process.

E. Stereo SLAM Based on Camera-2D LIDAR Fusion

Incorporating a monocular training strategy, while advantageous, introduces challenges such as motion confusion that can lead to erroneous infinite depth estimations for objects moving at the camera's speed. This issue, however, is mitigated in single-image estimation frameworks. To enhance scene reconstruction accuracy, we integrate stereo image pairs and single-beam LiDAR data. This approach, described in Fig. 2, inputs left image sequences from adjacent time points into the depth estimation and pose estimation networks to initially generate left disparity maps estimated from stereo pairs and relative motion estimations. Subsequently, these depth maps are refined using single-beam 2D LiDAR points to dynamically adjust the local scale.

The refinement process ensures that the disparity conforms to geometric relationships derived from both monocular video sequences and stereo image pairs, forming spatial-temporal optimization constraints. This dual constraint allows for the disparity to satisfy not only the refined multi-view re-projection loss but also the left-right appearance matching loss, thereby resolving the motion confusion and improving depth estimation accuracy across the scene.

IV. Experimental Results

A. Data Description

To validate our approach on a widely recognized dataset, we utilize the public KITTI dataset [52] for training our disparity estimation network on outdoor scenes. Specifically, the KITTI Eigen split serves as our training and evaluation ground, encompassing 22,600 images across 29 scenes for training, alongside 697 images for benchmarking against contemporary methodologies. To mirror our single-beam LiDAR-camera setup, we commence with stereo color images for initial training, subsequently narrowing down

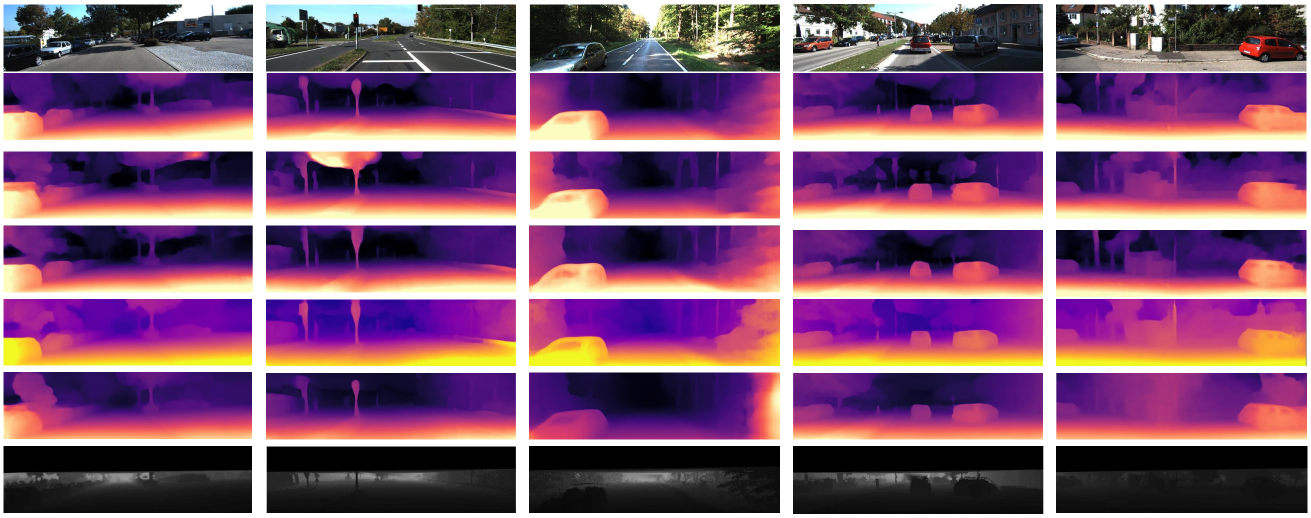

Fig. 4. Our dataset's visual comparison showcases our method's performance against leading alternatives. The presentation sequence is as follows: First to seventh row: Input color images; Depth maps generated by our approach; Watson et al. [44]; Godard et al. [17]; Guizilini et al. [45]; Bian et al. [46]; Corresponding ground truth depth maps.

Method	Supervision	Resolution	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
SfMLearner [14]	M	416 ×128	0.198	1.836	6.565	0.275	0.718	0.901	0.960
DDVO [47]	M	416 ×128	0.151	1.257	5.583	0.228	0.810	0.936	0.974
EPC++ [48]	M	256 ×832	1.414	1.029	5.350	0.216	0.816	0.941	0.976
Vid2Depth [13]	M	416 ×128	0.163	1.240	6.220	0.250	0.762	0.916	0.968
SC-SfMLearner [49]	M	832 ×256	0.137	1.089	5.439	0.217	0.830	0.942	0.975
Monodepth2 [17]	M	640 ×192	0.115	0.903	4.863	0.193	0.877	0.959	0.981
SuperDepth [50]	M	1024 ×382	0.112	0.875	4.958	0.207	0.852	0.947	0.977
PackNet-SfM [45]	M	640 ×192	0.111	0.785	4.601	0.189	0.878	0.960	0.982
HR-Depth [20]	M	640 ×192	0.109	0.792	4.632	0.185	0.884	0.962	0.983
Lite-Mono [21]	M	640 ×192	0.110	0.802	4.671	0.186	0.879	0.961	0.982
Ours	M	640 ×192	0.102	0.793	4.612	0.187	0.884	0.962	0.983
Depth-vo-feap [51]	MS	608 ×160	0.144	1.391	5.869	0.241	0.803	0.928	0.969
EPC++ [48]	MS	256 ×832	0.128	0.935	5.011	0.209	0.831	0.945	0.979
Monodepth2 [17]	MS	640 ×192	0.106	0.818	4.750	0.196	0.874	0.957	0.979
HR-Depth [20]	MS	640 ×192	0.107	0.785	4.612	0.185	0.887	0.962	0.982
Ours	MS	640 ×192	0.101	0.719	4.561	0.177	0.889	0.965	0.984

TABLE I

QUANTITATIVE DEPTH ESTIMATION RESULTS ON KITTI EIGEN TEST SPLIT. METHODS TRAINED WITH ONLY MONOCULAR IMAGE SEQUENCES ARE PRESENTED IN THE UPPER PART AND THOSE ALSO COMBINING STEREO IMAGE PAIRS DURING THE TRAINING PROCESS ARE PROVIDED IN THE BOTTOM PART.

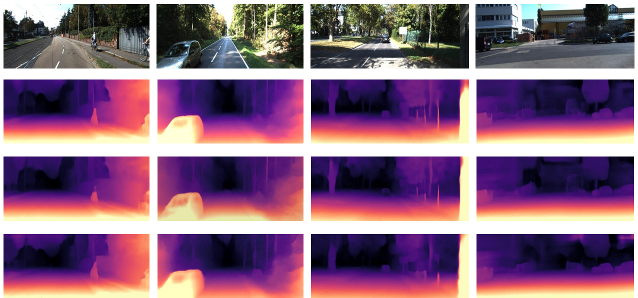

Fig. 5. The ablation study outcomes for our configurations are concisely depicted as follows, arranged top to bottom for clarity: input images (with 2D LiDAR); Depth map from stereo sequence; Depth estimation from mono sequence; Depth map only based on stereo images.

LiDAR inputs to single-line points from the provided multi-beam data. This process, coupled with our MLP neural network and local scaling strategy, yields refined depth maps for comparative analysis with recent advancements.

B. Training Configuration

Our depth estimation network, developed in PyTorch, leverages dual GeForce GTX 4090 GPUs for processing. Each dataset split undergoes 50 epochs of training with a batch size of 8. Input images undergo random cropping and resizing to dimensions of 640 × 192. Optimization is facilitated through the Adam optimizer, configured with $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and a learning rate that halves every 20 epochs. The depth-fitting MLP network features two hidden layers, each with 10 units, to mitigate overfitting risks. Data augmentation strategies, including random adjustments to contrast, brightness, and color within a 0.8 to 1.2 range, enhance the robustness of our training process.

C. Comparison with the state-of-the-art methods

Qualitative assessments of our methodologies based only on mono image sequence, stereo images, and stereo sequence—are illustrated in Fig. 4. When compared with state-of-the-art (SOTA) techniques [44], [17], [45], [46], our approaches excel in retaining intricate scene details such as traffic signs, trees, and vehicles.

Table I presents a comprehensive comparison of our proposed method against state-of-the-art techniques on the

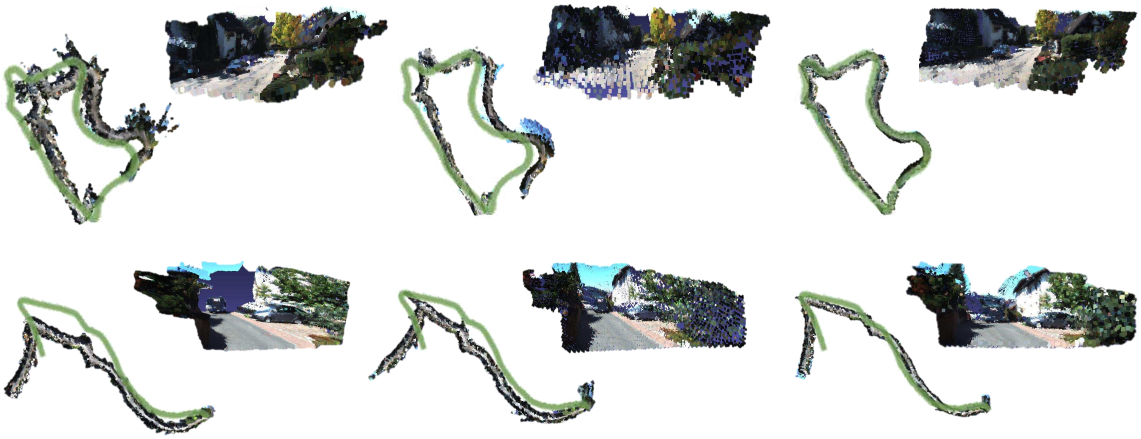

Fig. 6. We present a side-by-side comparison of 3D map reconstructions and estimated trajectory paths. From left to right, the sequence showcases: the red 3D map and a detailed zoom-in on the point cloud for a single frame, as produced by Monodepth2 [17], SC-SfMLearner [46], and our proposed method. The ground truth trajectory, depicted by a green line, is superimposed on each reconstructed map to facilitate direct comparison of the estimated paths with actual movement.

KITTI Eigen test split. Our approach, when trained solely on monocular video sequences, significantly surpasses competing methods across accuracy metrics (δ_1, δ_2, δ_3) and demonstrates substantial improvements in error metrics, particularly in Abs Rel and RMSE log. Incorporating stereo information into our single-beam LiDAR-camera fusion model, as discussed in Sec. III-E, yields further enhancements in accuracy and reductions in error, outperforming networks trained only with monocular inputs. Remarkably, our monocularly trained depth estimation already exceeds the performance of methods utilizing both monocular and stereo images, attributing to the efficacy of our single-beam LiDAR fusion module.

D. Ablation Studies and Analysis

A series of ablation studies detailed in Table II explore various configurations of our model. The inclusion of the refinement MLP network significantly betters both error and accuracy metrics across the mono sequence, stereo pairs, and stereo sequence setups, with notable improvements in RMSE and δ_1 accuracy. Integrating sequential data with spatial stereo data for training slightly enhances depth estimation, attributable to the additional geometric constraints and optimization opportunities afforded by combining stereo and adjacent view information.

	Without Refinement MLP network		With Refinement MLP network	
Settings	Abs Rel / Sq Rel / RMSE	$\delta < 1.25/\delta < 1.25^2/\delta < 1.25^3$	Abs Rel / Sq Rel / RMSE	$\delta < 1.25/\delta < 1.25^2/\delta < 1.25^3$
Mono sequence	0.124 / 0.915 / 4.985	0.861 / 0.956/ 0.981	0.102 / 0.793 / 4.612	0.884 / 0.962 / 0.983
Stereo pairs	0.115 / 0.897 / 5.047	0.854 / 0.949 / 0.975	0.106 / 0.786 / 4.582	0.871 / 0.951 / 0.982
Stereo sequence	0.113 / 0.833 / 4.901	0.861 / 0.956 / 0.981	0.101 / 0.719 / 4.561	0.889 / 0.965 / 0.984

TABLE II

ABLATION RESULTS ON DEPTH ESTIMATION FOR DIFFERENT SETTINGS AND COMPONENTS OF THE PROPOSED MODEL.

The ablation analysis for our method encompasses both quantitative and qualitative dimensions, with the outcomes illustrated in Fig. 5. The comparison reveals that the integrated stereo sequence (displayed in the second row) markedly outperforms the results obtained from solely monocular sequence (third row) or stereo pairs (fourth row) settings. Specifically, the combined method excels in accurately delineating clear object contours and edges, such as those of bicyclists, cars, and trees, within the test images, showcasing its superior scene reconstruction capability.

E. Comparisons of 3D Mapping and Estimated Trajectories.

Figure 6 illustrates the textured 3D maps and per-frame point clouds produced by various methods for Sequences 09 and 10 of the KITTI dataset. To adapt to the extensive depth range encountered in outdoor scenes, we optimized depth fusion by reducing voxel size. This adjustment enables our method to generate comprehensive 3D maps that are visually coherent. Unlike the outcomes observed with the approaches by [17] and [46], our 3D maps exhibit consistent 3D mapping and motion trajectory accuracy over long operation time, underpinning the robustness of our technique.

The superiority of our method is underscored in the estimated trajectories depicted in Fig. 6. When juxtaposed with the ground truth (green line), our trajectory estimation closely aligns, demonstrating remarkable accuracy. Specifically, for Sequence 09, our method precisely captures the closed loop characteristic of the trajectory, a critical attribute not as accurately replicated by competing methods, which displays discernible discrepancies at the loop's start and end points.

V. CONCLUSION

In this study, we introduce a fusion-based lightweight SLAM framework aimed at achieving high-precision 3D mapping and camera motion estimation economically. Utilizing a novel approach that aligns 2D single-beam LiDAR data with image-derived depth maps, our system extends LiDAR's capabilities to generate detailed, full-resolution 3D maps. Starting with disparity estimation from single images, our network harnesses the precision of sparse LiDAR points and the broader coverage of estimated depth maps to produce refined, accurate depth information. This process facilitates the creation of comprehensive 3D maps and camera motion estimation, with continuous online learning enhancing real-world adaptability across varied environments. Our method represents a pioneering SLAM system development effort to integrate affordable 2D LiDAR and camera data for detailed motion estimation and full-resolution 3D mapping.

Ack: This publication is based upon work supported by NSF under Awards No. 2334624, No. 2334690, and No. 2334246.

References

[1] Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2722–2730.

[2] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2147–2156.

[3] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8445–8453.

[4] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn, "A survey of socially interactive robots," *Robotics and autonomous systems*, vol. 42, no. 3-4, pp. 143–166, 2003.

[5] Andreas Geiger, Julius Ziegler, and Christoph Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *2011 IEEE intelligent vehicles symposium (IV)*. Ieee, 2011, pp. 963–968.

[6] Tristan Laidlow, Jan Czarnowski, and Stefan Leutenegger, "Deepfusion: Real-time dense 3d reconstruction for monocular slam using single-view depth and gradient predictions," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4068–4074.

[7] Friedrich Fraundorfer, Christopher Engels, and David Nistér, "Topological mapping, localization and navigation using image collections," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2007, pp. 3872–3877.

[8] Ji Zhang and Sanjiv Singh, "Loam: Lidar odometry and mapping in real-time.," in *Robotics: Science and Systems*, 2014, vol. 2.

[9] Jakob Engel, Thomas Schöps, and Daniel Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[10] Yuliang Zou, Zelun Luo, and Jia-Bin Huang, "Df-net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 36–53.

[11] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *ECCV*, 2016.

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017.

[13] Reza Mahjourian, Martin Wicke, and Anelia Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5667–5675.

[14] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017.

[15] Zhichao Yin and Jianping Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *CVPR*, 2018.

[16] Jia-Wang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," *arXiv preprint arXiv:1908.10553*, 2019.

[17] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow, "Digging into self-supervised monocular depth estimation," *ICCV*, 2019.

[18] J. Zhou, Y. Wang, K. Qin, and W. Zeng, "Moving indoor: Unsupervised video depth learning in challenging environments," in *ICCV*, 2019.

[19] Zehao Yu, Lei Jin, and Shenghua Gao, "P 2 net: patch-match and plane-regularization for unsupervised indoor depth estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 206–222.

[20] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan, "Hr-depth: High resolution self-supervised monocular depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 2294–2301.

[21] Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle, "Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18537–18546.

[22] Mu He, Le Hui, Yikai Bian, Jian Ren, Jin Xie, and Jian Yang, "Ra-depth: Resolution adaptive self-supervised monocular depth estimation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*. Springer, 2022, pp. 565–581.

[23] Antyanta Bangunharcana, Ahmed Magd, and Kyung-Soo Kim, "Dualrefine: Self-supervised depth and pose estimation through iterative epipolar sampling and refinement toward equilibrium," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 726–738.

[24] Ruoyu Wang, Zehao Yu, and Shenghua Gao, "Planedepth: Self-supervised depth estimation via orthogonal planes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21425–21434.

[25] Andrea Pilzer, Dan Xu, Mihai Puscas, Elisa Ricci, and Nicu Sebe, "Unsupervised adversarial depth estimation using cycled generative networks," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 587–595.

[26] Yasin Almalioglu, Muhamad Risqi U Saputra, Pedro PB de Gusmao, Andrew Markham, and Niki Trigoni, "Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5474–5480.

[27] T. Feng and D. Gu, "Sganvo: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks," *RA-L*, 2019.

[28] L. Sheng, D. Xu, W. Ouyang, and X. Wang, "Unsupervised collaborative learning of keyframe detection and visual odometry towards monocular deep slam," in *ICCV*, 2019.

[29] N. Yang, L. Stumberg, R. Wang, and D. Cremers, "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *CVPR*, 2020.

[30] H. Bian, J.and Zhan, N. Y. Wang, C. Chin, T.and Shen, and I. Reid, "Unsupervised depth learning in challenging indoor video: Weak rectification to rescue," *arXiv:2006.02708*, 2020.

[31] David Eigen and Rob Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2650–2658.

[32] Fayao Liu, Chunhua Shen, and Guosheng Lin, "Deep convolutional neural fields for depth estimation from a single image," in *CVPR*, 2015.

[33] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia, "Geonet: Geometric neural network for joint depth and surface normal estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 283–291.

[34] Hyungjoo Jung, Youngjung Kim, Dongbo Min, Changjae Oh, and Kwanghoon Sohn, "Depth prediction from a single image with conditional adversarial networks," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 1717–1721.

[35] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox, "Demon: Depth and motion network for learning monocular stereo," in *CVPR*, 2017.

[36] David Eigen, Christian Puhrsch, and Rob Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NIPS*, 2014.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[38] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin, "Single view stereo matching," in *CVPR*, 2018, pp. 155–163.

[39] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048.

[40] Daniel Huber, Takeo Kanade, et al., "Integrating lidar into stereo for fast and improved disparity computation," in *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*. IEEE, 2011, pp. 405–412.

[41] Will Maddern and Paul Newman, "Real-time probabilistic fusion of sparse 3d lidar and dense stereo," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 2181–2188.

[42] Kihong Park, Seungryong Kim, and Kwanghoon Sohn, "High-precision depth estimation with the 3d lidar and stereo fusion," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2156–2163.

[43] Yawen Lu, Yuxing Wang, Devarth Parikh, Yuan Xin, and Guoyu Lu, "Extending single beam lidar to full resolution by fusing with single image depth estimation," in *International Conference on Pattern Recognition (ICPR)*, 2021, pp. 6343–6350.

[44] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov, "Self-supervised monocular depth hints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2162–2171.

[45] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2485–2494.

[46] 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*, pp. 1–17, 2021.

[47] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey, "Learning depth from monocular videos using direct methods," in *CVPR*, 2018, pp. 2022–2030.

[48] Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille, "Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2624–2641, 2019.

[49] 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 (IJCV)*, 2021.

[50] Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon, "Superdepth: Self-supervised, super-resolved monocular depth estimation," in *International Conference on Robotics and Automation (ICRA)*, 2019, pp. 9250–9256.

[51] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *CVPR*, 2018.

[52] Andreas Geiger, Philip Lenz, and Raquel Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.