Visual Simultaneous Localization and Mapping using Filtered Depth Maps based on a Foundation Model

Jimin Song[1], HyungGi Jo[1], Yongsik Jin[2] and Sang Jun Lee[1]

***Abstract*— Depth estimation is a challenging task that predicts pixel-level distances from a single RGB image through a deep neural network. Depth estimation models can recognize 3D information by utilizing a monocular camera, and it can either replace or complement distance measurement sensors in mobile robots. In this paper, we propose a visual simultaneous localization and mapping (SLAM) pipeline which leverages 2D features and depth maps estimated from a foundation model. To improve the performance of the SLAM algorithm, we introduce a filtering strategy for eliminating depth values with low reliability. Experimental results on the KITTI dataset demonstrate that the filtering strategy is effective to improve the reliability of depth values in the SLAM pipeline. Moreover, experiments on our dataset demonstrates that the proposed method outperform previous visual SLAM methods with a significant margin.**

I. INTRODUCTION

SLAM is a key technology in robotics that enables a system to perceive its surroundings through sensors, build a global map, and estimate its current location within that map. It is primarily researched for implementing autonomous driving in mobile robots and unmanned aerial vehicles. SLAM system [1] that utilize expensive light detection and ranging (LiDAR) sensors for accurate distance measurement have already demonstrated high performance, particularly in indoor settings. However, LiDAR sensors have several difficulty as the light used for distance measurement can pass through glass and be absorbed by black materials, introducing noise in certain environments. Additionally, the high cost compared to other sensors makes it less suitable for mass production. The acquisition of monotonous point cloud data in environments like straight highways reduces the effectiveness of LiDAR-based SLAM, presenting challenges for its application. As a result, there has been growing interest in camera-based visual SLAM technologies as an alternative, which can be applied in more general environments. Nevertheless, RGB camera-based visual SLAM algorithms have the inherent difficulty of lacking depth information, making it difficult to determine the scale of translation. Although depth maps obtained from RGB-D cameras are denser than those generated by LiDAR, they have a shorter effective depth range, making them less suitable for visual SLAM in larger spaces beyond indoor environments. As illustrated in Fig. 1, we aim to improve the performance of RGB camera-based visual SLAM by employing deep neural networks to estimate depth maps from images. Additionally, standard deviation-based filtering is employed to eliminate regions in the depth estimation results that may adversely affect the performance of visual SLAM.

Recent advancements in deep learning have rapidly expanded its application across various fields, including robotics, for a wide range of tasks. One of the key tasks in robotics for spatial perception is depth estimation, which estimates pixel-level distances from camera images. Estimating a dense depth map, a three-dimensional representation, from a single two-dimensional RGB image is a highly challenging task from a deep learning perspective, as it involves pixel-wise regression with generalization difficulty. To address this challenging problem, the most intuitive deep learning training approach is supervised learning, where the model is trained to minimize the loss function between the estimated depth map and the ground truth data. The most widely used public dataset in the field of depth estimation, KITTI [2], includes ground truth depth maps generated by projecting point cloud data acquired via LiDAR onto images. This data is collected from vehicles equipped with both cameras and LiDAR sensors in various road-driving scenarios. In this paper, we also evaluate the performance of the proposed method for both depth estimation and odometry estimation using a underground parking lot dataset collected from an underground parking lot.

II. RELATED WORK

Estimating absolute depth in an image is inherently an ill-posed problem due to scale ambiguity. So Eigen et al. [3] emphasized the importance of estimating relative depth, and proposed a scale-invariant log (SIlog) loss function to address this issue. SIlog loss has been widely adopted in recent supervised depth estimation algorithms and is also utilized for fine-tuning the foundation depth model in this work. Recently, various methods have been proposed, proposing novel model architectures, approaching depth estimation from different perspectives and including the incorporation of techniques used in different fields. Lee et al. [4] proposed a deep learning model that estimates geometric parameters, used to generate feature maps for depth estimation. Bhat et

*This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2024-RS-2024-00439292) and a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2024-00346415)

[1]Jimin Song, HyungGi Jo and Sang Jun Lee are with Division of Electronic Engineering, Jeonbuk National University, 567 Baekje-daero, Deokjin-gu, Jeonju 54896, Republic of Korea jimin_song@jbnu.ac.kr, hygijo@jbnu.ac.kr and sj.lee@jbnu.ac.kr

[2]Yongsik Jin is with Daegu-Gyeongbuk Research Center, Electronics and Telecommunications Research Institute (ETRI), Daegu 42994, Republic of Korea yongsik@etri.re.kr

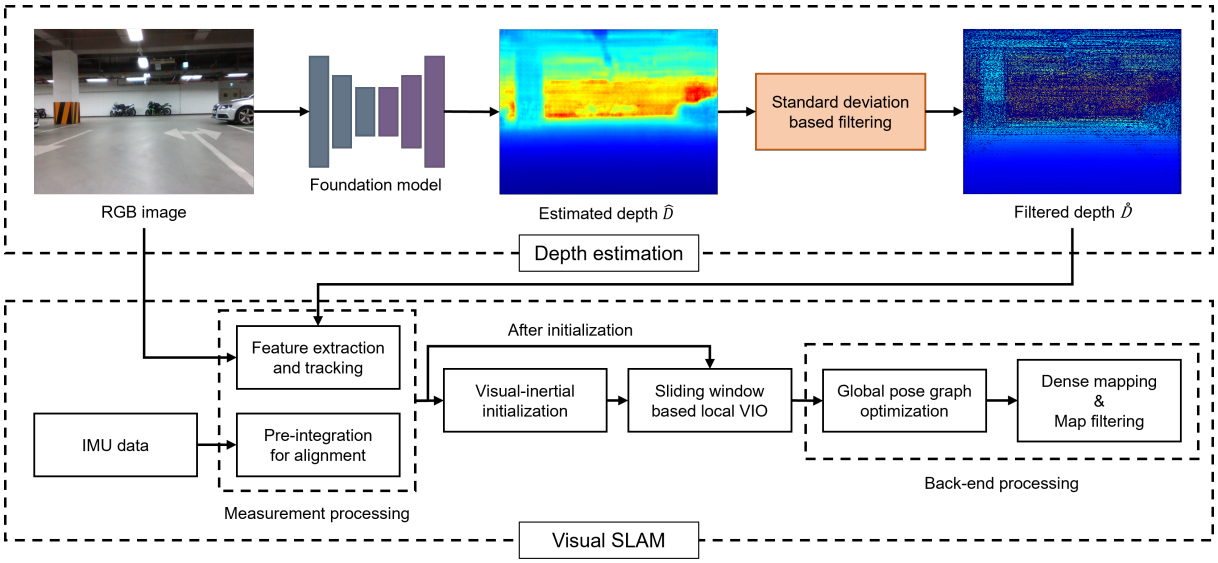

Fig. 1. Overview of proposed visual SLAM pipeline.

al. [5] approached depth estimation as an ordinal regression and proposed an adaptive binning method that adjusts to the varying distribution of depth values in each image. Yuan et al. [6] introduced the conditional random fields technique, originally used in natural language processing, into the field of computer vision for depth estimation. In this paper, we compare the performance of our proposed method with the three recent supervised learning-based methods [4]–[6].

Recently, studies utilizing foundation models, which are sophisticated deep neural networks trained on large-scale datasets, have demonstrated superior performance. Ranftl et al. [7] proposed a meta-dataset for depth estimation named MiDaS, considering factors such as data acquisition environment, density, accuracy, and diversity. They also proposed a novel loss function to estimate relative depth. Ranftl et al. [8] enhanced MiDaS [7] and introduced a network based on vision transformers [9] for dense prediction. Yang et al. [10] proposed a pipeline named Depth Anything that incorporates not only existing labeled datasets but also unlabeled datasets into the meta-dataset, achieving state-of-the-art performance in zero-shot relative depth estimation. In this study, we propose a method that fine-tunes pre-trained model [10] for absolute depth estimation on the target dataset. The proposed method demonstrates superior performance compared to existing supervised learning approaches [4]–[6] on both public dataset [2] and our underground parking lot dataset.

Recent SLAM algorithms often leverage inertial measurement unit (IMU) sensors that measure acceleration and angular velocity. While the integration of acceleration and angular velocity allows for the estimation of translation and rotation, these sensors are prone to noise. Consequently, algorithms that employ additional sensors have been proposed, notably LiDAR-inertial odometry (LIO) and visual-inertial odometry (VIO), which combine LiDAR or camera sensor with IMU. Bai et al. [1] offers an efficient LIO algorithm that processes LiDAR data in a sparse incremental voxel format, as opposed to the traditional tree-like structures. In indoor environments, LIO performance is notably high; thus, the ground truth odometry for the dataset collected in this study was generated using Faster-LIO [1]. VINS-Mono [11] is a tightly coupled visual SLAM that utilizes a single cost function to estimate the optimal state based on camera images and IMU data. VINS-RGBD [12] builds upon VINS-Mono [11] by leveraging depth map from RGB-D camera. In this study, we compare the performance of the proposed RGB-based visual SLAM method with existing RGB-based [11] and RGB-D-based [12] approaches.

III. METHODOLOGY

In this study, we utilize the architecture of the deep neural network proposed by Ranftl et al. [8] for depth estimation. The initial parameters of the model is set to the pre-trained parameters for estimating relative depth [10]. We propose a method to train the model to estimate absolute depth $\hat{d}$ while preserving the ability of model to estimate relative depth which estimates the disparity space value $\hat{t}$. During the pre-training process, regions that correspond to infinitely distant areas, such as the sky, are identified using a segmentation model [13], and the ground truth value for these areas is defined as zero. Consequently, the estimated disparity map $\widehat{T}$ could contain zeros, necessitating a process to constrain the minimum value for calculation of loss function as follows:

$$\tilde{t} = max(\epsilon, \hat{t}) \tag{1}$$

where $\hat{t}$ and $\tilde{t}$ denote the elements of the estimated disparity map $\widehat{T}$ and the revised disparity map $\widetilde{T}$, respectively. In experiments, ϵ is set to 10^{-6}. The inverse of the revised disparity $\tilde{t}$ is defined as estimated depth $\hat{d}$. We define the SIlog loss [3] as the final depth loss function for the n valid values in the ground truth depth map :

$$L = \frac{1}{n}\sum_i e_i^2 - \frac{\lambda}{n^2}(\sum_i e_i)^2, e_i = \log \hat{d}_i - \log d_i. \tag{2}$$

Here, $\hat{d}_i$ and d_i represent the i-th values in the estimated depth map $\widehat{D}$ and the ground truth depth map D, respectively.

Depth estimation tends to be more challenging around object boundaries. However, these regions are advantageous for feature point extraction, and the performance of depth estimation directly affects the overall performance of visual SLAM. To address this issue, we propose applying standard deviation-based filtering to the depth map. For the estimated depth map $\widehat{D}$, we compute the depth mean map $\overline{D}$ using a convolution operation with a $k \times k$ kernel consisting of the values of $1/k^2$. Based on this, the i-th pixel of the depth standard deviation map $\widetilde{D}$ is calculated as follows:

$$\tilde{d}_i = \sqrt{\frac{1}{k^2} \sum_{j \in N_i} (\hat{d}_j - \bar{d}_i)^2}, \tag{3}$$

where $\bar{d}_i$ represents the i-th value in the depth mean map $\overline{D}$, and N_i is a $k \times k$ neighborhood centered at the i-th pixel. In experiments, k is set to 7. Depth values are filtered out if their local standard deviation is larger than the median of the depth standard deviation map $\tilde{D}$ as follows.

$$\mathring{d}_i = \begin{cases} \hat{d}_i, & \text{if } \tilde{d}_i < median[\tilde{D}]. \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

In (4), $\mathring{d}_i$ is the i-th pixel value of the filtered depth map $\mathring{D}$. The filtered depth map is utilized in the following RGB-D SLAM pipeline.

The proposed visual SLAM algorithm is constructed based on the existing RGB-D camera-based visual SLAM algorithm [12], to leverage the filtered depth map. Sensor measurement processing refers to the preprocessing of sensor data prior to odometry estimation, utilizing the RGB images from the camera and the acceleration and angular velocity measurements from the IMU. In the RGB images, features extraction [14] and tracking [15] are conducted for perspective-n-points based visual odometry. Pre-integration is conducted on the IMU data, which is sampled a higher rate than the camera. An initialization process is necessary for aligning structure-from-motion based visual odometry with the integration-based inertial odometry. After the initialization, local VIO based on a sliding window is performed for triangulation [16]. To minimize accumulated errors, back-end processing is implemented, which includes loop closing-based global pose graph optimization, octree [17]-based dense mapping, and map filtering.

IV. EXPERIMENTS

A. Experiment setup and evaluation metrics

To evaluate the performance of both the proposed depth estimation and odometry estimation methods, we collected and processed data using a mobile robot equipped with the sensor configuration shown in Fig. 2 in an underground parking lot. The hardware used for the experiments included an AMD EPYC 7313P 16-core processor and two NVIDIA GeForce RTX 4090 GPUs. Existing supervised learning-based depth estimation algorithms were re-trained in the same environment. For fine-tuning the foundation depth model, the Adam optimizer [18] with weight decay values of 10^{-2} and 0 was employed, along with learning rates of 10^{-6} and 10^{-5} for the encoder and decoder, respectively. To prevent overfitting, several data preprocessing techniques were applied during the training of the depth network. Random rotations were applied to the KITTI dataset and the underground parking lot dataset, with maximum rotations of $1.0°$ and $2.5°$, respectively. Horizontal flipping and color adjustments to contrast, brightness, and color space, were applied with a 50 % probability.

To evaluate the performance of depth estimation, six error metrics and three accuracy metrics were used. The error metrics, which lower values indicate better performance, include SIlog, absolute relative error (AbsRel), squared relative error (SqRel), root mean squared error (RMSE), root mean squared error of inverse depth (RMSEi), and logarithmic error (log10). The accuracy metrics are defined by the proportion of pixels where $\delta = \max(\hat{d}/d, d/\hat{d})$ is less than thresholds of $[1.25, 1.25^2, 1.25^3]$. Detailed formulas and explanations for these depth estimation metrics can be found in previous work [3]. For odometry estimation, three error metrics [20] were employed. One is the RMSE of the absolute trajectory error (ATE), which evaluates the cumulative odometry error. The others compute the translation and rotation errors of the relative pose error (RPE) between two time steps. In the tables comparing quantitative performance, the best results are highlighted in bold, and the second-best results are underlined.

B. Results on the KITTI dataset

The KITTI dataset is a benchmark widely used for computer vision tasks essential to autonomous driving technologies, including depth estimation. For depth estimation, the dataset includes ground truth depth maps generated using two different methods. The original KITTI [2] ground truth is produced using iterative closest point algorithm to integrate point cloud data from adjacent time steps. The improved KITTI dataset [19] enhances depth map density through network-based interpolation. For training the network, we

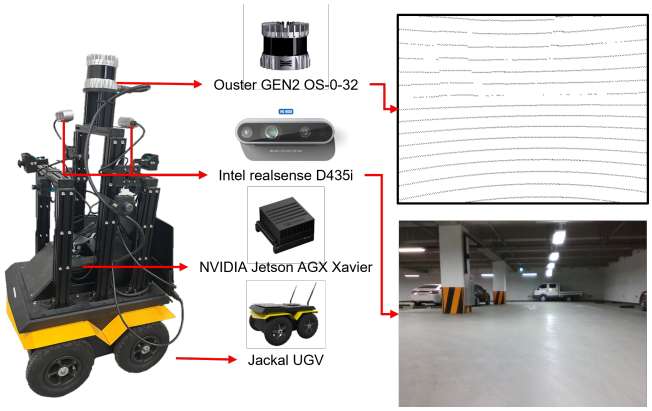

Fig. 2. Sensor configuration of the mobile robot. The top right visualizes the depth map obtained by projecting LiDAR points corresponding to the RGB image shown at the bottom right.

TABLE I

QUANTITATIVE RESULTS OF DEPTH ESTIMATION ON THE KITTI DATASET. ERROR METRICS HIGHLIGHTED IN RED INDICATE THAT LOWER VALUES ARE BETTER, WHILE ACCURACY METRICS HIGHLIGHTED IN BLUE INDICATE THAT HIGHER VALUES ARE BETTER.

Method	Error metric ↓						Accuracy metric ↑		
	AbsRel	SqRel	RMSE	RMSEi	SIlog	log10	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Original KITTI [2]									
BTS [4]	0.085	0.565	4.102	0.176	16.644	0.041	0.903	0.965	0.983
NewCRFs [6]	0.101	0.642	4.324	0.189	17.602	0.048	0.875	0.959	0.981
AdaBins [5]	0.093	0.537	3.877	0.177	16.601	0.043	0.894	0.964	0.984
Depth anything [10]	0.071	0.434	3.568	0.156	14.806	0.034	0.932	0.973	0.986
Ours	**0.045**	**0.111**	**1.477**	**0.094**	**8.746**	**0.021**	**0.977**	**0.992**	**0.995**
Improved KITTI [19]									
BTS [4]	0.061	0.253	2.826	0.098	8.955	0.027	0.952	0.993	0.998
NewCRFs [6]	0.079	0.358	3.247	0.123	11.010	0.035	0.923	0.986	0.997
AdaBins [5]	0.069	0.256	2.671	0.106	9.551	0.031	0.945	0.990	0.998
Depth anything [10]	0.048	0.147	2.175	0.073	6.598	0.021	0.979	0.998	0.999
Ours	**0.037**	**0.045**	**0.910**	**0.052**	**4.360**	**0.016**	**0.993**	**0.999**	**1.000**

TABLE II

QUANTITATIVE RESULTS OF DEPTH ESTIMATION ON THE UNDERGROUND PARKING LOT DATASET. ERROR METRICS HIGHLIGHTED IN RED INDICATE THAT LOWER VALUES ARE BETTER, WHILE ACCURACY METRICS HIGHLIGHTED IN BLUE INDICATE THAT HIGHER VALUES ARE BETTER.

Method	Error metric ↓						Accuracy metric ↑		
	AbsRel	SqRel	RMSE	RMSEi	SIlog	log10	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
BTS [4]	0.101	0.444	1.974	0.154	14.419	0.038	0.909	0.973	0.991
NewCRFs [6]	0.106	0.462	1.823	0.153	14.771	0.039	0.904	0.975	0.992
AdaBins [5]	0.096	0.409	1.938	0.150	14.208	0.037	0.914	0.973	0.991
Depth anything [10]	0.104	0.369	1.918	0.150	14.644	0.041	0.916	0.978	0.993
Ours	**0.068**	**0.114**	**0.920**	**0.096**	**8.894**	**0.027**	**0.972**	**0.993**	**0.997**

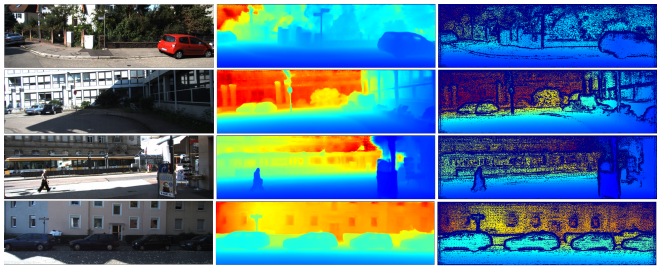

Fig. 3. Qualitative experimental results of depth estimation on the KITTI dataset.

utilize the data split proposed by Eigen et al. [3], which consists of 39,810 images for training, 4,424 images for validation, and 697 images for evaluation. Depending on the network architecture, network input images at a resolution of 1216×352, while the foundation depth model processes input images at a resolution of 1204×350.

As shown in Table I, the fine-tuned foundation model significantly outperforms existing supervised methods [4]–[6] across all metrics. Furthermore, our method, which incorporates standard deviation filtering, demonstrates substantial performance improvements over the fine-tuned model alone, effectively eliminating high-error predictions. Fig. 3, specifically the third column, illustrates that the proposed depth estimation method successfully filters out region at object boundaries. While the second column presents a reasonable depth map, the application of filtering clearly enhances the distinction of object boundaries, making the content of the image easier to interpret. In summary, by applying both fine-tuning and filtering through the proposed approach, our method achieves notably superior depth estimation performance in road-driving scenarios, both quantitatively and qualitatively.

C. Results on the underground parking lot dataset

The underground parking lot, where vehicles and pedestrians coexist, represents an intermediate space between outdoor and indoor environments. To generate the ground truth depth map, we conducted offline calibration [21] to estimate the extrinsic parameters between the camera and LiDAR. The LiDAR points were then projected onto the image plane using these extrinsic parameters. The dataset, comprising RGB images with a resolution of 640×480 and

TABLE III

QUANTITATIVE RESULTS OF ODOMETRY ESTIMATION ON THE UNDERGROUND PARKING LOT DATASET.

Driving scenario		Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Average
Driving distance [m]		225.35	225.35	122.26	44.32	44.27	26.02	
Method	Depth	RMSE of ATE [m]						
VINS-Mono [11]	None	7.6614	2.0864	2.7397	0.8252	2.9859	4.6489	4.1034
VINS-RGBD [12]	Sensor	5.8164	7.8056	0.9733	1.5514	0.7652	0.7833	4.8166
Ours	Depth anything [10]	3.3402	**0.9557**	0.6888	**0.3065**	0.4287	0.3768	1.5921
	Ours	**3.2658**	0.9608	**0.6427**	0.3389	**0.2785**	**0.2319**	**1.5481**
Method	Depth	Translation error of RPE [m]						
VINS-Mono [11]	None	0.1518	0.0539	0.0563	0.0857	0.3136	0.2527	0.1127
VINS-RGBD [12]	Sensor	0.1217	0.5044	0.0753	0.2152	0.0632	0.1252	0.2413
Ours	Depth anything [10]	0.0608	0.0471	0.0440	0.0446	0.0452	**0.0604**	0.0512
	Ours	**0.0470**	**0.0460**	**0.0294**	**0.0316**	**0.0353**	0.0610	**0.0424**
Method	Depth	Rotation error of RPE [deg]						
VINS-Mono [11]	None	**0.1379**	**0.1442**	**0.1844**	**0.4872**	0.5463	**0.6791**	**0.2175**
VINS-RGBD [12]	Sensor	0.3601	0.4537	0.3722	1.3643	1.2669	2.4969	0.5969
Ours	Depth anything [10]	0.2513	0.2519	0.2847	0.8516	0.7982	1.4372	0.3762
	Ours	0.1411	0.1782	0.1882	0.4885	**0.5074**	0.7028	0.2288

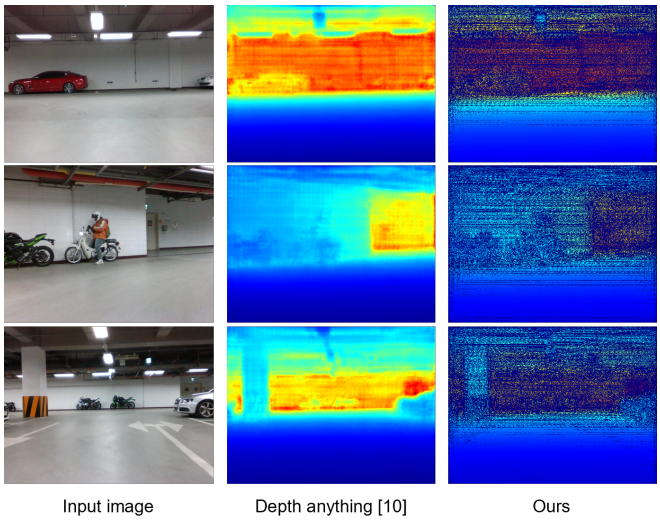

Fig. 4. Qualitative experimental results of depth estimation on the underground parking lot dataset.

corresponding ground truth depth maps, was split into 14,424 pairs for training and 5,846 pairs for testing.

As shown in Table II, unlike the results from the KITTI experiments, the performance of depth estimation on the custom dataset is similar among the existing methods, including Depth Anything [10]. It occurred due to the lower quality of the custom dataset for training depth networks compared to the well constructed KITTI dataset. However, the results applying filtering showed even greater performance improvement than in the KITTI experiments. The qualitative results in Fig. 4 demonstrate that not only were object boundaries, but horizontal noise patterns in the estimated depth map were also successfully removed. This noise, as shown in Fig. 2, likely results from the sparse LiDAR projection. Overall, in the custom dataset, the proposed method exhibited significant quantitative performance improvements and effectively filtered out noise in depth estimation.

The dataset for odometry estimation was acquired separately from the depth dataset and consists of typical driving scenarios (cases 1–3) and special driving scenarios (cases 4–6). The quantitative performance of the proposed RGB-based odometry estimation is compared in Table III with the existing RGB-based method [11] and the RGBD-based method [12]. An ablation study was conducted to evaluate the effect of standard deviation-based filtering in the proposed visual SLAM pipeline. Since VINS-Mono [11] does not have direct depth information, it suffers from scale ambiguity, resulting in lower performance in terms of ATE and the translation error of RPE. Although VINS-Mono [12] achieved the best performance in the rotation error of RPE, the proposed method closely followed with the second-highest performance. The proposed method consistently showed the highest or second-highest performance across all cases, demonstrating that it is the most effective algorithm quantitatively. Additionally, as shown in Fig. 5, the trajectory estimated by our method closely aligns with the ground truth trajectory, illustrating the high qualitative performance of the proposed algorithm.

V. CONCLUSIONS

In this paper, we propose a novel approach to SLAM, a crucial technology for autonomous driving. Our method leverages neural network to estimate high-accuracy depth maps from camera images, followed by standard deviation-based filtering into a visual SLAM pipeline. To validate

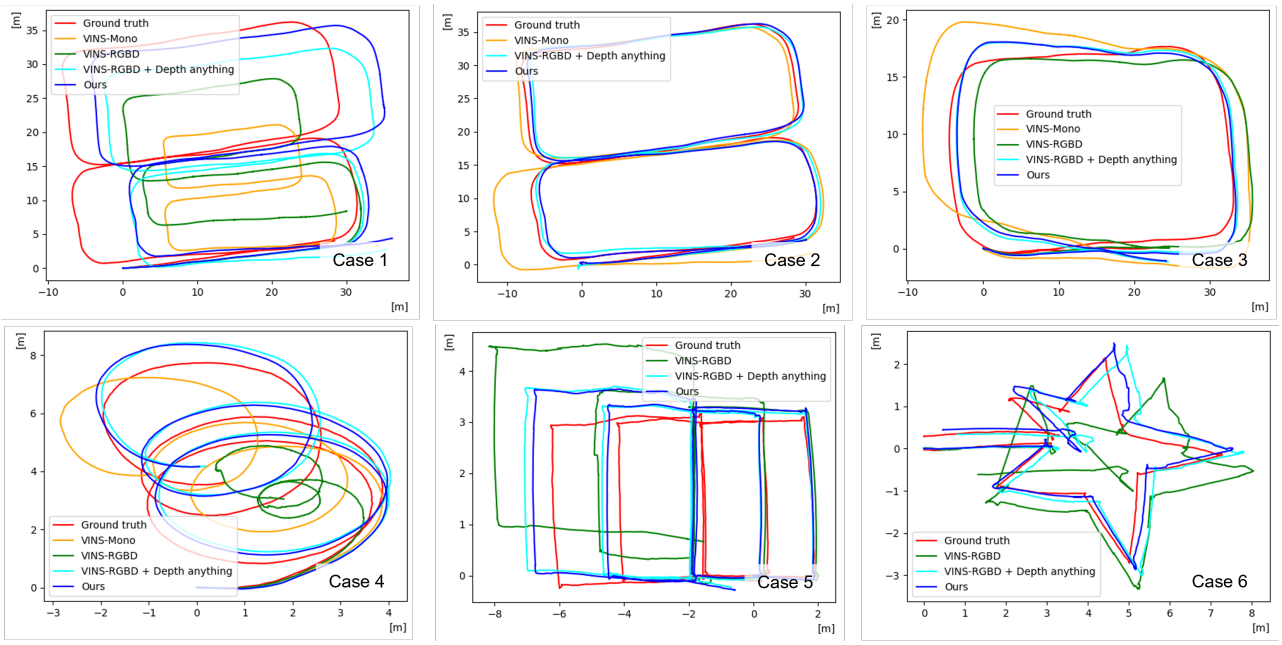

Fig. 5. Qualitative experimental results of odometry estimation on the underground parking lot dataset.

the effectiveness of the proposed method, we conducted both quantitative and qualitative comparisons with existing approaches using public and custom datasets, demonstrating significant performance improvements. Through these experiments, we confirmed that utilizing foundation models can significantly enhance performance in the depth estimation. Additionally, we demonstrated that improving depth estimation accuracy directly contributes to better odometry estimation performance. We hope that this study will contribute to the development of more effective and efficient visual SLAM algorithms.

REFERENCES

[1] C. Bai, T. Xiao, Y. Chen, H. Wang, F. Zhang, and X. Gao, "Faster-lio: Lightweight tightly coupled lidar-inertial odometry using parallel sparse incremental voxels," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4861–4868, 2022.

[2] A. Geiger, P. Lenz, and R. 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.

[3] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[4] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *arXiv preprint arXiv:1907.10326*, 2019.

[5] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4009–4018.

[6] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan, "New crfs: Neural window fully-connected crfs for monocular depth estimation. arxiv 2022," *arXiv preprint arXiv:2203.01502*, 2022.

[7] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.

[8] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[9] D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929*, 2020.

[10] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 371–10 381.

[11] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE transactions on robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[12] A. Tyagi, Y. Liang, S. Wang, and D. Bai, "Dvio: Depth-aided visual inertial odometry for rgbd sensors," in *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2021, pp. 193–201.

[13] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[14] J. Shi *et al.*, "Good features to track," in *1994 Proceedings of IEEE conference on computer vision and pattern recognition*. IEEE, 1994, pp. 593–600.

[15] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, 1981, pp. 674–679.

[16] J. Civera, A. J. Davison, and J. M. Montiel, "Inverse depth parametrization for monocular slam," *IEEE transactions on robotics*, vol. 24, no. 5, pp. 932–945, 2008.

[17] D. Meagher, "Geometric modeling using octree encoding," *Computer graphics and image processing*, vol. 19, no. 2, pp. 129–147, 1982.

[18] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *2017 international conference on 3D Vision (3DV)*. IEEE, 2017, pp. 11–20.

[20] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 573–580.

[21] D. Tsai, S. Worrall, M. Shan, A. Lohr, and E. Nebot, "Optimising the selection of samples for robust lidar camera calibration," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 2631–2638.