This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2910826, IEEE Sensors Journal

Deep Sensor Fusion between 2D Laser Scanner and IMU for Mobile Robot Localization

Chi Li, Sen Wang, Yan Zhuang, *Member, IEEE,* Fei Yan, *Member, IEEE,*

***Abstract*—Multi-sensor fusion plays a key role in 2D laser based robot location and navigation. Although it has achieved great success, there are still some challenges, e.g., being fragile when having large angular rotation. In this paper, we present a deep learning based approach to localizing a mobile robot using a 2D laser and an Inertial Measurement Unit. A novel Recurrent Convolutional Neural Network (RCNN) based architecture is developed to fuse laser and inertial data for scan-to-scan pose estimation. A scan-to-submap optimization is also introduced to optimize the poses estimated by the RCNN for enhanced robustness and accuracy. Extensive experiments have been conducted in both simulation and practice with a real mobile robot, verifying the effectiveness of the proposed deep sensor fusion system.**

***Index Terms*—Data fusion, Pose estimation, 2D laser scanning, Inertial measurement unit (IMU).**

I. Introduction

Laser scanners are widely used for robot localization and mapping [1]. However, localization algorithms solely based on laser data may be unreliable when facing challenging scenarios [2]. One of the main techniques to improve the robustness of laser based robot localization and navigation is multi-sensor fusion.

Multi-sensor fusion plays a key role in the field of robot navigation and localization, especially when considering uncertainty from noisy sensor readings. Bayes filters, e.g., Kalman and particle filters, have been designed to fuse multiple sources of information for optimal estimation of robot poses [3], [4]. In [5], Kalman filter is used to fuse laser based Iterative Closest Point (ICP) registration algorithm with an Inertial Measurement Unit (IMU). Similar ideas are also introduced in [6], [7]. Other kinds of sensor combinations are also popular for sensor fusion, such as ultrasonic sensors with IMU [8], laser scan with stereo vision [9], and ultra-wideband system with IMU [10]. Although impressive performance has been achieved by these methods, they usually require rigorous time synchronization and extrinsic calibration between sensors [11]. Meanwhile, they lack a learning capability to automatically benefit from the large amount of sensor data collected over time. How to directly learn sensor fusion from laser and inertial data for robot localization remains an open question.

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 61503056 and Grant Nos. U1508208).

C.Li is with the School of Control Science and Engineering, Dalian Univ. of Technology, Dalian 116024, China. (e-mail: lichiduter@mail.dlut.edu.cn)

S.Wang is with Edinburgh Centre for Robotics at Heriot-Watt University,UK.(e-mail:s.wang@hw.ac.uk)

Y. Zhuang is with the School of Control Science and Engineering, Dalian Univ. of Technology, Dalian 116024, China (e-mail: zhuang@dlut.edu.cn)

F. Yan is with the School of Control Science and Engineering, Dalian Univ. of Technology, Dalian 116024, China (e-mail: fyan@dlut.edu.cn)

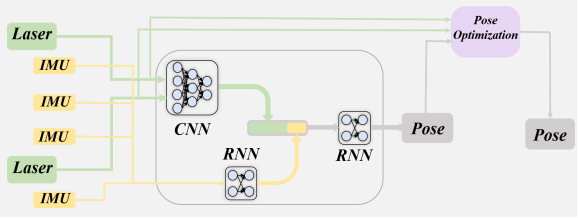


Fig. 1. System overview of the proposed sensor fusion algorithm for robot localization. The laser and inertial data are fused by a Recurrent Convolutional Neural Network (RCNN), and pose prediction of the RCNN is further optimized by a scan-to-submap scan matching.

One of the main elements of laser based robot localization systems is laser odometry, whose accuracy and robustness to some extent determine the performance of the whole system. Unfortunately, laser odometry tends to be unreliable when encountering a motion with significant angular changes. One of the possible solutions is to train deep learning models with sensor data collected from these challenging scenarios. However, there is very limited work on learning laser odometry using deep models.

In this paper, we propose a novel Recurrent Convolutional Neural Networks (RCNN) based sensor fusion framework for robot localization using a laser scanner and an IMU. The framework is formulated as a sequence-to-sequence pose regression problem. Our main contributions are as follows:

- A novel RCNN based sensor fusion algorithm is designed for laser inertial pose estimation. To the best of our knowledge, this is the first deep learning based sensor fusion of laser and inertial data.
- A hybrid framework incorporates optimization based scan matching with deep learning, which achieves high accuracy and robustness even in challenging scenarios.

The proposed deep learning based data fusion can fuse sensors with different sampling rates (in this work laser and IMU run at 40Hz and 100Hz, respectively), and does not require an accurate time synchronization or extrinsic calibration between sensors.

The rest of this paper is organized as follows. Section II reviews related work. Following an overview of the system in Section III, Section IV presents the proposed RCNN based laser and inertial sensor fusion method. In Section V, a pose optimization is introduced to refine poses predicted from the network. Experimental results are given in Section VI. The conclusions are drawn in Section VII.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2910826, IEEE Sensors Journal

II. Relate Works

Multi-sensor fusion aims for optimal state estimation by considering the characteristics of different sensors. Classical odometry measures often combine wheeled odometry with IMU. In this paper, we fucus on the laser and IMU data as the observation to estimate the robot pose.

A. Robot Localization

1) laser odometry: In recent years, there are more and more approaches in the geometric field of laser data, such as Point Cloud Registration, location, and laser odometry. Most of popular solutions of 2D laser Simultaneous Localization and Mapping (SLAM) such as Gmapping [12], Hector-SLAM [13], and Google Cartographer [14], use wheel odometer to improve accuracy. Hector-SLAM uses IMU data for the platform not exhibiting roll/pitch motion, cartographer and some 3D laser SLAM (LOAM[15]) used IMU data as an initial value before laser points registration which makes the registration more quickly and robust. Earlier, Zebedee [16] used 2D laser scanner and IMU to generate a 3D SLAM.

2) Deep Learning based methods: Most of existing deep learning based robot localization methods are vision based [17], [18], [19]. Other machine learning techniques are employed for the loop-closure problem, such as [20], [21]. Nicolai et al. [22] utilized deep learning for laser based odometry estimation. [23], [24] achieved the deep learning based laser registration, and Elbaz et al. [25] used the deep Auto-Encoder for point cloud localization. Deep learning based 2d scan matching method is proposed by Li et al.[26], in which using the Hector SLAM to achieve a good result. However, the benefits of deep learning for laser and inertial data have not been fully explored in the field of geometric matching of 2D laser data.

B. Deep Learning based Multi-Sensor Fusion

Because of limited benefits from data fusion, Hosseinyalamdary et al.[27] presented the deep Kalman filter which is based an RNN to fuse the IMU and Global Navigation Satellite System (GNSS). To increase the accuracy of the vision based depth estimation, Gianluca et al.[28] utilized deep learning for extracting confidence maps from a camera and a stereo vision system. Deep learning based fusion method also is used in other fields. For instance, a network is trained to fuse images and point cloud for 3D object detection [29], [30].

Rambach et al.[31]used an RNN to fuse the output of a standard marker-based visual pose system and IMU, thereby eliminating the need for statistical filtering. Clark et al.[32] presented visual-inertial odometry by using a deep-learning network to fuse video and IMU data.

III. System Overview

In this section, an RCNN based sequence-to-sequence data fusion pose estimation system is briefly described. Fig. 1 shows the proposed system framework. It consists of two parts: an RCNN based data fusion for scan-to-scan pose estimation and an ICP based scan-to-submap pose optimization.

The RCNN based sensor fusion is composed of a CNN based point cloud feature extraction, an RNN based IMU registration and an RNN based the data fusion. The CNN is designed to extract features from two laser scans for fusion. Meanwhile, the RNN for IMU registration processes sequences of IMU data between two laser scans. Since laser and IMU sensors usually run at quite different sampling rate, one of the challenges is the sensor data streams being multi-rate. To tackle this problem, we use an additional RNN to fuse features learnt from the laser CNN and IMU RNN. The RNN not only combines the information of laser and IMU sensors but also learns a motion model of a robot.

Scan-to-scan pose estimation tends to be accurate as it is an estimate for a short period between only two consequtive laser scans. However, the global pose which accumulates scan-to-scan errors suffers from drifts over time. In order to further optimize the pose prediction from the RCNN, an ICP based scan-to-submap algorithm is utilized. The ICP based pose optimization can reduce the accumulated error of scan-to-scan estimation and eliminate some outliers.

IV. Laser Inertial Fusion through Recurrent Convolutional Neural Networks

This section presents the proposed RCNN based sensor fusion for robot localization in detail. The cost functions designed for training the RCNN are also described.

A. CNN based Feature Extraction

Feature extraction can be formulated as a nonlinear mapping problem. An efficient approach to automatically learning useful feature representation is deep neural networks. It is well known that convolutional layers can be interpreted as transforming inputs into feature representations, which can be used for regression or classification problems. In order to learn useful features that are suitable for laser scans, a CNN is developed to perform feature extraction on the concatenation of a pair of two consecutive laser scans. Because odometry systems need to be deployed in unknown environments, the CNN in this work is to extract geometric features instead of semantic ones.

TABLE I
Configuration of the CNN

Layer	Kernel Size	Padding	Stride	Number of Channerls
Conv1	15	7	2	16
Conv2	15	7	2	16
Conv3	9	4	2	64
Conv4	9	4	2	64
Conv5	5	2	2	256
Conv6	5	2	2	256

The configuration of the CNN is outlined in TABLE I. The input of the network is a stack of two laser scans. There are six convolutional layers, each of which has a rectified linear unit (ReLU) activation. Since 2D laser sensors are adopted in work, the network uses 1D convolutional layers. Following the last convolutional layer, a global max pooling is used to extract a global feature which is part of the input for the RNN network.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2910826, IEEE Sensors Journal

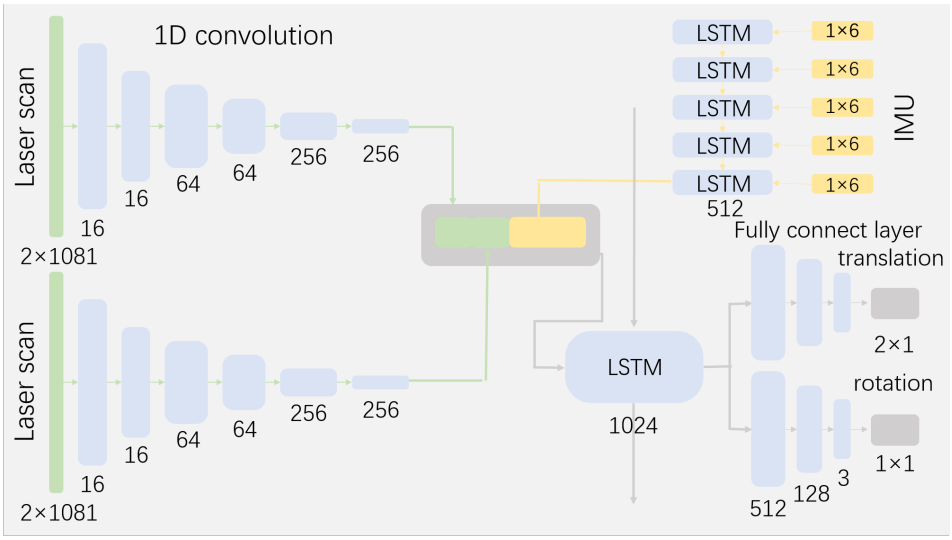


Fig. 2. Architecture of the network in proposed data fusion system. The network consists of 1D convolutional layer, two LSTM layers with two fully connected layers.

B. RNN based Sequence Learning and Fusion

Recurrent Neural Networks (RNNs) are one kind of neural networks which have memories for temporal learning. It is capable of processing sequential data since it can leverage historical information for current state prediction, modeling dependencies in a sequence. Therefore, RNN is well suited for the localization problem which involves both temporal model (motion model) and sequential data (laser scan sequence and IMU sequence). For instance, estimating the pose of the current frame can benefit from information encapsulated in previous frames.

1) LSTM based Sequence Learning: Long Short-Term Memory (LSTM) is one of the most popular types of RNNs because it is capable of learning long-term dependencies by introducing memory gates and units. Its gates automatically decide when to store or discard data and memory while training. Assure the LSTM's input is x_k at time k and its hidden state is h_{k-1} and the memory cell is c_{k-1} of the previous LSTM unit. LSTM updates at time step k according to

$$\begin{aligned} i_k &= \sigma(W_{xi}x_k + W_{hi}h_{k-1} + b_i) \\ f_k &= \sigma(W_{xf}x_k + W_{hf}h_{k-1} + b_f) \\ g_k &= \tanh(W_{xg}x_k + W_{hg}h_{k-1} + b_g) \\ c_k &= f_k \odot c_{k-1} + i_k \odot g_k \\ o_k &= \sigma(W_{xo}x_k + W_{ho}h_{k-1} + b_o) \\ h_k &= o_k \odot \tanh(c_k) \end{aligned} \tag{1}$$

where $\odot$ is element-wise product of two vectors, σ is sigmoid non-linearity, $\tanh$ is hyperbolic tangent non-linearity, W terms denote corresponding weight matrices, b terms denote bias vectors, i_k, f_k, g_k, c_k and o_k are input gate, forget gate, input modulation gate, memory cell and output gate at time k, respectively,

The LSTM network takes joint feature representation of laser and IMU as input and learns the sequential model for localization. We are now in the position of discussing how to register and fuse laser and IMU data.

2) IMU Sequence Registration: The output of the system is a scan-to-scan pose estimation, which takes laser frame as a reference in our work. Since laser and IMU sensors run at very different sampling rates (laser at about 40 Hz while IMU at 100 Hz), their sensor data needs to be temporally registered. Therefore, we use an additional LSTM network to learn a global feature of a sequence of IMU data which is collected between two consecutive laser scans. In practice, the sequence length of IMU data corresponding to two frames of laser scans may not be fixed. Thanks to the flexibility of the RNN structure, IMU data with different lengths can be natively processed by the LSTM network.

3) Laser and IMU Fusion: We use a structure similar to IMU RNNs to achieve fusion between laser and IMU data. As shown in the Fig. 2, the input of the main LSTM network is the concatenation of the two global features extracted from the CNN of laser scans and the LSTM of the IMU data, respectively. The main LSTM network takes laser and IMU sequences as input and realizes data fusion through sequential learning. It can learn a complex motion dynamics of a platform and also sequential dependencies which are difficult to be modelled manually.

C. Cost Function

The proposed RCNN based network can be considered as a scan-to-scan pose regression function by learning suitable hyperparameters. Therefore, the cost function is designed to train the RCNN so that its predicted poses are close to the ground truth. Specifically, the cost function contains three parts: a scan-to-scan pose error, a sequence pose error and a reconstruction error.

1) Scan-to-Scan Pose Error: Scan-to-scan pose error is formulated as Mean Square Error (MSE) of all positions and orientations. The function aims to minimize the Euclidean

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2910826, IEEE Sensors Journal

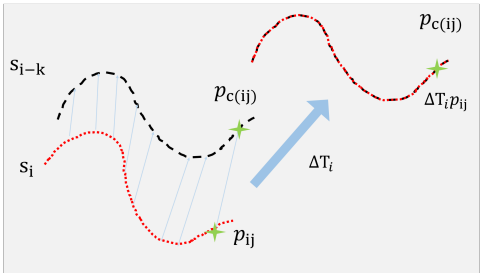


Fig. 3. Reconstruction error. p_{ij} is a point in scan s_i, and $p_{c(ij)}$ is its corresponding points in scan s_{i-k}. The reconstruction error is the sum of the distances between the transformed points $\Delta T p_{ij}, j = 1, 2, 3...N$ and their corresponding points. The more accurate the transformation ΔT_i is, the smaller the reconstruction error would be.

distance between the ground truth poses and the estimated ones from the network. The error function is as follows:

$$error_{s2s} = \frac{1}{n}\sum_{i=0}^{n} \|\Delta x_i - \Delta\hat{x}_i\|_2^2 + \|\Delta y_i - \Delta\hat{y}_i\|_2^2 + \lambda\|\Delta\theta_i - \Delta\hat{\theta}_i\|_2^2 \tag{2}$$

where $\|\cdot\|_2^2$ is the 2-norm, and λ is a factor to balance the weight of positions and orientations, and n is the batch size.

2) Global Pose Error of Sequence: Since robot localization is to obtain global pose estimates of the robot concerning the coordinate defined by an initial pose, the scan-to-scan pose estimation error will accumulate errors on poses over time. The global pose error of sequence is designed to reduce the accumulated pose errors in a sequence, which further reduces scan-to-scan errors. The experiment results suggest the sequence error can drive the network training to converge quickly. The sequence global errors can be calculated as

$$error_{sg} = \frac{1}{Q}\sum_{q=0}^{Q} \|T_q - \hat{T}_q\|_2^2 \tag{3}$$

where Q is the number of sequences, T_q is the pose with respect to the end pose of the sequence. T_q can be derived from $T_q = T_0 \prod_{i=0}^{k} \Delta T_i$ where k is the number of laser scans in each sequence. $\hat{T}_q$ is the ground truth of this sequence. The three degrees of freedom pose between two scan frames is conventionally represented as an element of the special Euclidean group SE(2) of transformations. SE(2) is a differentiable manifold with elements consisting of rotation from the special orthogonal group SO(2) and a translation vector:

$$T = \{\begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0^T} & 1 \end{bmatrix} \in \mathbb{R}^{3\times3} \mid \mathbf{R} \in \mathrm{SO}(2), \mathbf{t} \in \mathbb{R}^{\mathbf{2}}\} \tag{4}$$

The relationship between T and pose vector $[x, y, \theta]$ is

$$T = \begin{bmatrix} cos\theta & -sin\theta & x \\ sin\theta & cos\theta & y \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

3) Reconstruction Error: Reconstruction error considers the geometric matching errors of ICP based scan matching. It performs a geometric consistency check to reduce the errors on pose estimation. As shown in Fig. 3, the reconstruction error minimizes the distances between the transformed laser points in scan s_i and its corresponding points in the scan s_{i-k}:

$$error_{rs} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{N} \|\Delta T_i p_{ij} - p_{c(ij)}\|_2^2 \tag{6}$$

where $p_{ij} \in s_i$ and $p_{c(ij)} \in s_{i-k}$ are laser scan points, N is the number of points in each scan. ΔT_i is the relative transformation predicted from the network.

In summary, the cost function considering these three different error functions for training is

$$F_c = \alpha \times error_{s2s} + \beta \times error_{sg} + \gamma \times error_{rs} \tag{7}$$

where α, β and γ are the weights to balance the error functions.

V. Pose Optimization using Scan-to-Map Matching

To further refine the pose estimation, a geometric method is employed to fine tune the poses predicted by the network. Since the localization errors are mainly produced by the accumulated errors from the scan-to-scan matching, we propose to maintain a local map which can be used to locally match a laser scan for geometric constraints. A basic ICP is designed as the optimization for the scan-to-map matching.

1) Matching Laser Scan to A Local Map: An ICP fine-tuning is initialized by the relative transformation estimated by the RCNN. ICP takes two laser scans and a local map as input, and outputs a best-fit transformation ΔT which minimizes the distance between the transformed points in s_i and their corresponding points in the local map:

$$\Delta T = \arg\min_{\Delta T} \frac{1}{2}\sum_{j=1}^{N} \|\Delta T p_j - p_{c(j)}\|_2^2 \tag{8}$$

Combining the relative transformation ΔT and the pose of the local map T_{local} can get the global pose T_{global} of the current frame.

$$T_{global} = T_{local}\Delta T \tag{9}$$

2) Local Map Updating: The local map is a set of feature points which are used to match laser scans. In the process of updating the local map, new points are firstly added to the local map. The points which are never or hardly used in ICP fine-tuning are selected as new points.

Second, the pose of the local map is updated by using the transformation computed from ICP. Points outside the field of view are culled, thereby reducing the computational complexity of searching for corresponding points.

Map points culling can improve computational efficiency while ensuring that local maps do not excessively increase in a low-speed or static state.

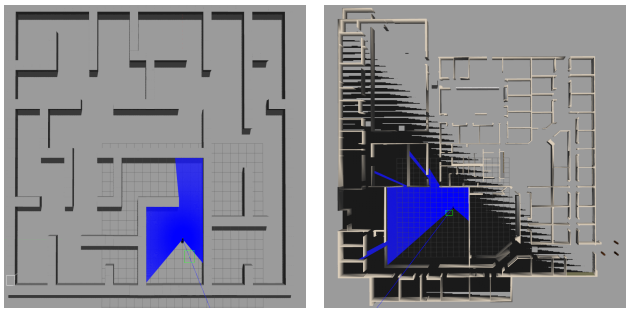
VI. Experimental Results

In this section, we evaluate the performance of the proposed system. The RCNN model is trained by using data collected from a simulator, and tested in simulation and on a real Pioneer 3-DX robot running in indoor environments. Here we use the Hector SLAM and an IMU enhanced Hector SLAM for comparison.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2910826, IEEE Sensors Journal

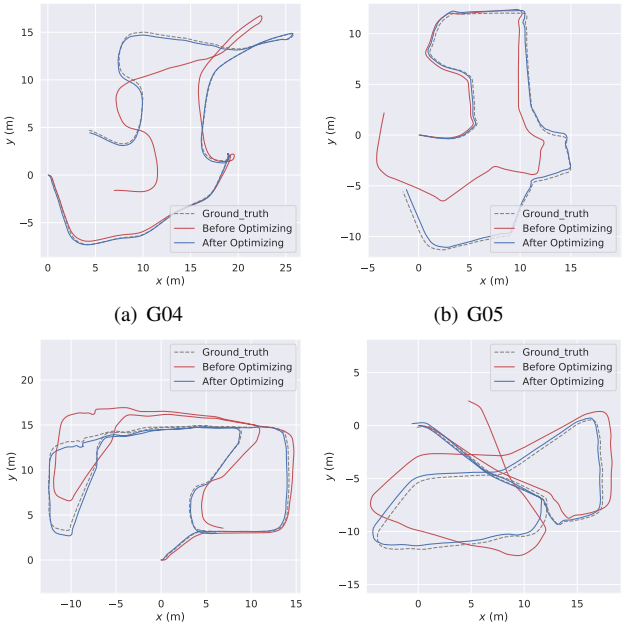
TABLE II
THE DATA COLLECTED FROM SIMULATION ENVIRONMENT.

Sequence	g01	g02	g03	g04	g05	w01	w02	w03	w04	w05	w06	w07	w08
Length (m)	100.650	72.233	99.634	85.355	69.190	96.243	129.226	129.031	150.574	105.137	90.045	88.179	95.016
Laser Scan	8496	7437	8483	7541	5707	7737	9313	10488	12345	10482	8921	8544	9164
IMU	21233	18591	21204	18849	14267	19329	23246	26215	30854	26194	22290	21346	22904



(a) Scene G (b) Scene W

Fig. 4. Simulation environments. (a). Scene G is a maze-like scene. (b). Scene W is an office-like scene.



(a) G04 (b) G05

(c) W05 (d) W07

Fig. 5. Simulation Results of the Pose Estimation Without and With Scan-to-map Optimization.

A. Dataset and Training

In order to obtain enough data for training, we generate 2D laser point clouds with poses in simulation.

1) Simulation Data: The simulated indoor environments, as shown in Fig.4, are built on Robot Operating System (ROS) Gazebo. There is a simulated Pioneer 3-DX robot mounted with a Hokuyo UTM-30LX 2D laser scanner and an IMU. The maximum linear and angular velocities of the robot are $0.6m/s$ and $1.0rad/s$, respectively. The maximum measurement range of the laser scanner is $30.0m$, and wide angle is 270, in addition, Gaussian noise $N(0, 0.1^2)$ is augmented into the simulated laser scanner. The ground truth positions $[x, y, \theta]^T$ and the range measurements are recorded at the speed of 40Hz. A training sample can be acquired according to Equation (10),

$$\Delta T = T_{i-k}^{-1} T_i \tag{10}$$

where T_i is the position at time i, and k is an integer to define the interval between two range measurements.

In total, 13 sequences are collected from scene G (G01-G05) and scene W (W01-W08) as shown in Table II. The Table II shows the length and the numbers of laser scan and IMU sensor data of each sequence. The training and validation data is generated from G01, G02, G03, W01, W02, W04 and W06, while G04, G05, W03, W05, W07 and W08 are used for testing. In order to train the RCNN network, the training sequences are generated at different sampling frequencies, for instance $k = [4, 8, 12]$. The network is implemented based on the Pytorch framework and trained by using an NVIDIA GTX1080 GPU. The Adam optimizer is employed to train the network for up to 50 epochs with learning rate 0.001.

B. Experiments in Simulation

1) Simulation Results: The testing results in the simulation are shown in Table III. We use averaged Root Mean Square Errors (RMSEs) of the pose errors to evaluate the performance of our system. We compare several algorithms:

- Fusion Net is our proposed whole system;
- Laser Net is the Fusion Net without IMU fusion;
- Hector SLAM is an open source implementation of geometric laser SLAM [13];
- Hector SLAM + IMU is Hector SLAM with EKF based IMU fusion.

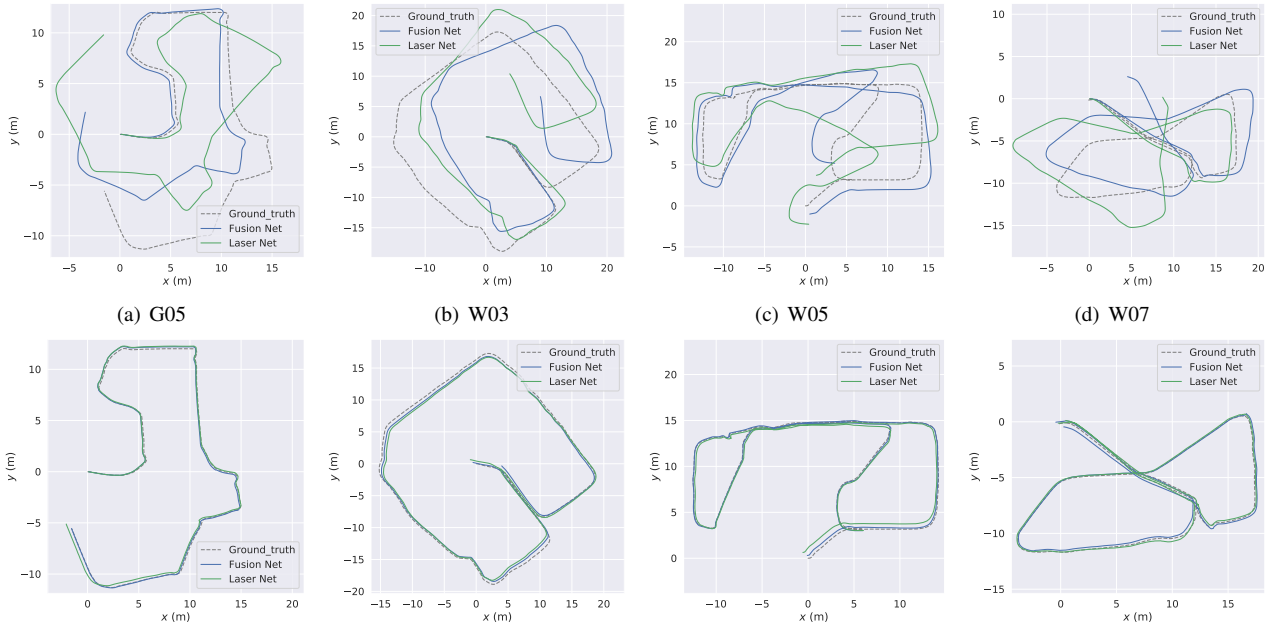
Table III shows the testing results of the 4 algorithms in the simulation. It can be seen Fusion Net and Hector SLAM + IMU produce better results than ones without fusion. Meanwhile, Fusion Net outperforms Hector SLAM + IMU for the average accuracy, which verifies the effectiveness of the proposed deep learning based fusion method.

The predicted trajectories of the simulation experiments are given in Fig. 5, where the black dashed line is ground truth, the red line is the RCNN predicted poses without the scan-to-map optimization, and the blue line is the RCNN predicted poses with the scan-to-map optimization. It can be seen that the proposed RCNN network can predict poses accurately using laser and IMU data, which validates the design of the deep learning based sensor fusion algorithm. The frame-to-submap optimization can achieve higher accuracy by reducing the accumulate errors and eliminating the outliers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2910826, IEEE Sensors Journal

TABLE III
ABSOLUTE TRANSLATION ERRORS (RMSE) OF THE TEST DATA COLLECTED FROM SIMULATION ENVIRONMENT.

Sequence	G04	G05	W03	W05	W07	W08	Average
Fusion Net	0.153344	**0.162791**	0.687987	**0.180551**	**0.234858**	0.321351	**0.290147**
Laser Net	0.272754	0.180043	0.661636	0.417849	0.248305	0.889022	0.444934833
Hector SLAM	0.217349	10.495914	0.260873	7.390917	2.423515	0.561514	3.558347
Hector SLAM + IMU	**0.134898**	0.510529	**0.173633**	0.350534	0.301404	**0.1128**	0.3392715



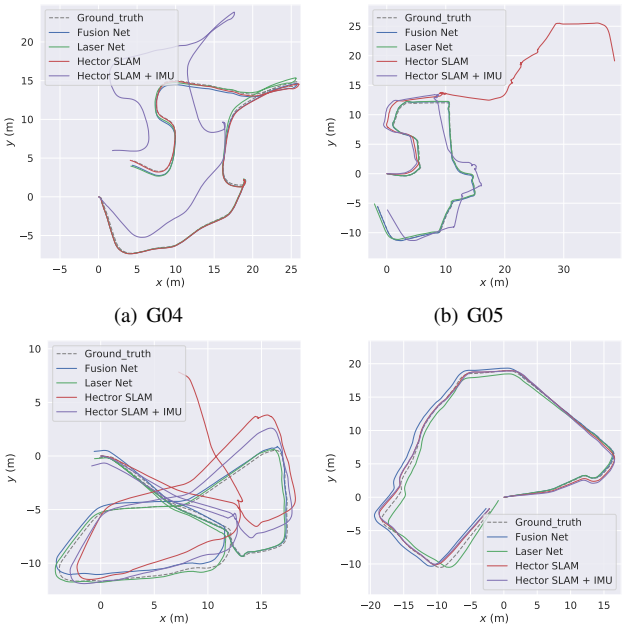
(a) G05 (b) W03 (c) W05 (d) W07

(e) G05 (f) W03 (g) W05 (h) W07

Fig. 6. Predicted trajectories of Fusion Net and Laser Net in simulation. Top row: Prediction without scan-to-map optimization. Bottom row: The corresponding prediction with scan-to-map optimization.

2) Pose Estimation of Data Fusion Results: Fig. 6 shows the results of Fusion Net and Laser Net. The blue lines are the results of the Fusion Net, while the green ones are from the Laser Net. It can be seen that the Laser Net can estimate the poses of the robot. However, its accuracy is lower than the Fusion Net, which uses laser and IMU data. This means the deep learning based multi-sensor fusion is capable of improving accuracy.

Fig.7 shows the comparison between the network-based method, Hector SLAM (red line) and IMU enhanced Hector SLAM (purple line). Although Hector SLAM provides nice results in general, it fails in some scenarios where a large angular velocity (such as 1 rad/s) presents. IMU enhanced Hector SLAM performs more robustly than the Hector SLAM, which means the data fusion is useful for pose estimation. In contrast, the proposed RCNN based method can estimate the poses robustly in extreme cases. This proves the proposed RCNN architecture is capable of learning to localize, even in some challenging environments.



(a) G04 (b) G05

(c) W07 (d) W08

Fig. 7. Trajectories of comparison in simulation. Both Fusion Net and Laser Net use the scan-to-submap optimization.

C. Experiments on Real Robot Platform

1) Robot Setup: To evaluate the performance in a practical environment, a real Pioneer 3-DX robot is set for the experiments. As shown in Fig.8, it is mounted with a Hokuyo UTM-30LX 2D laser scanner and an IMU, similar to the con-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2910826, IEEE Sensors Journal

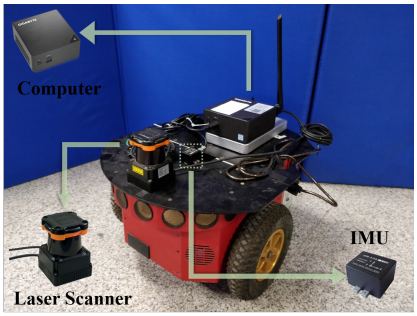


Fig. 8. The setup of the mobile robot equipped with a Hokuyo laser scanner and an IMU sensor.

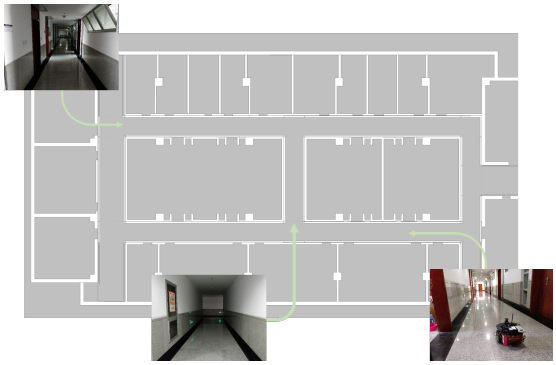


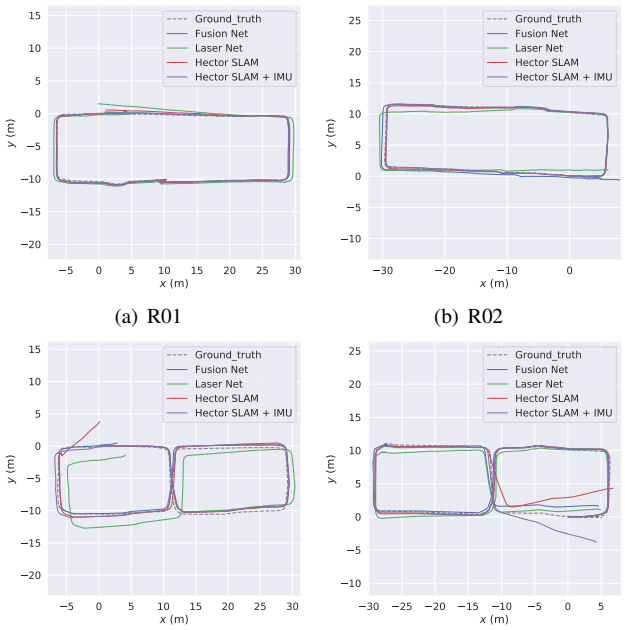
Fig. 9. Floor plan of the real environment used for testing.

figuration in the simulation. A Mini-PC (Intel© Core™Dual Core i7-8550U and 8GB RAM) is installed to run a ROS based navigation software which provides the ground truth for evaluation.

2) Results in Real Environments: The network is fine-tuning with the real data before testing in practice. The data gathered for fine tuning is summarized in Table IV. In order to verify the impact of different motions, the sequences R01, R03 and sequences R02, R04 are collected with various robot tuning orientations in an indoor environment shown in Fig. 9. The experimental results of the proposed system in real environments are shown in Fig 10. It can be seen the results of the Fusion Net outperform the ones of the other three algorithms. Since corners usually cause sudden geometry changes, traditional geometry based methods, e.g., Hector SLAM, may suffer from large errors at corners. This is demonstrated in Fig. 10(c). Therefore, the reliability of the prediction has been improved through the RCNN based sensor fusion architecture.

TABLE IV
ABSOLUTE TRANSLATION ERRORS (RMSE) OF DATA COLLECTED FROM REAL ENVIRONMENTS FOR TESTING.

Sequence	R01	R02	R03	R04
length(m)	95.129	95.213	107.408	109.919
Fusion net	**0.173728**	1.607354	**0.248956**	**0.480199**
Laser net	0.812354	0.823306	0.850727	0.793432
Hector SLAM	0.276757	0.321314	0.768335	1.091287
Hector SLAM + IMU	0.184775	**0.178936**	0.375193	0.823869



(a) R01 (b) R02

(c) R03 (d) R04

Fig. 10. Trajectories in real environmental. We use the data in Table IV evaluate our system.

D. Real-Time Performance

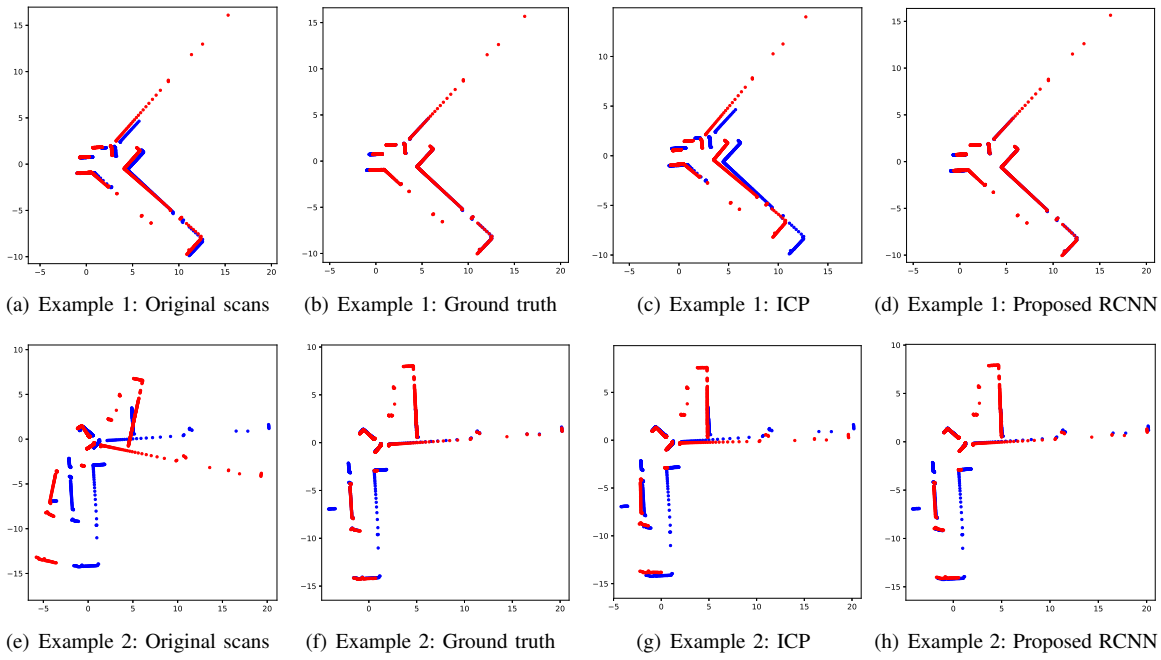
We test our system on a laptop (Intel© Core™Dual Core i7-4720HQ and 12GB RAM) for the real-time performance. Our system can run at about 10 Hz without the GPU. Compared with geometric methods, it is roughly similar speed with a laptop. However, with the aid of GPU, the deep learning based methods can be more effective. With an NVIDIA GTX1080 GPU, the network of our system can process more than 4000 laser scan per second, which is significantly faster than coventional methods.

E. Discussion

1) Better Frame to Frame Matching from RCNN: It is well known that geometry based ICP methods are sensitive to initial values. It may suffer from local minimums with large errors when the initialization is bad. In contrast, the neural network based methods match a pair of scans by extracting their global feature of scans, which may improve the robustness in some cases. Fig.11 gives two examples which demonstrate situations that the network performs more robust than traditional ICP algorithms.

The network-based method is data-driven which does not require the exact geometric correspondence. The laser scanner obtains discrete point cloud data in which there is no consistent one-to-one matching between each other. The geometry based ICP may fail to match points accurately when the consistent points cannot be found. It is also well known that dramatic rotation motion could cause problems for ICP algorithms because of the search function of the ICP. The disadvantage of the proposed deep learning based method is that the model needs a large amount of data for training.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2019.2910826, IEEE Sensors Journal

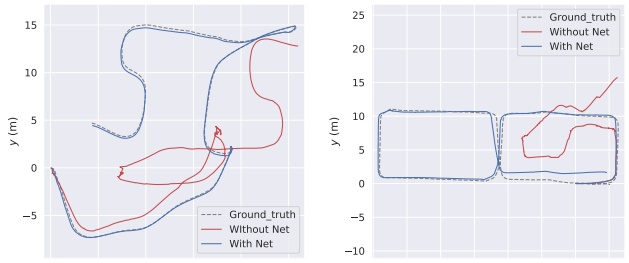


(a) Example 1: Original scans (b) Example 1: Ground truth (c) Example 1: ICP (d) Example 1: Proposed RCNN

(e) Example 2: Original scans (f) Example 2: Ground truth (g) Example 2: ICP (h) Example 2: Proposed RCNN

Fig. 11. Two exemplars of frame to frame matching. Note the proposed RCNN algorithm works better than the traditional ICP.

2) Essential to Have the RCNN Fusion.: Fig.12 shows the experimental results without and with the proposed deep neural network. It is obvious that the results of the scan-to-submap method degenerates without the initialization provided by the RCNN fusion. On the other hand, an excellent initialization value can facilitate faster ICP convergence with better results. This means it is essential for the whole system to have the proposed RCNN architecture.



(a) (b)

Fig. 12. Experimental results without and with the proposed RCNN. Note how the results degenerate without the RCNN.

VII. Conclusion and Future Work

In this paper, we proposed a deep learning based fusion framework to estimate robot pose. It fuses 2D laser scanner with IMU to implement an odometry system. According to the experiments in simulation and on a real mobile robot, our system can effectively estimate the poses of the robot by using sequence data from the 2D laser scanner and the IMU. It also performs more robustly than traditional geometry based methods in some challenging situations, such as moving with a large angular velocity.

In the field of 3D laser odometry, the geometric method also has problems with the large rotation angle. Therefore, the deep learning based method are still relevant. In order to improve the accuracy of the odometry, learning the motion dynamics of the robot by combining RNNs and CNN should be a promising direction. In further, we will continue to explore the deep learning based 3D laser odometry combined with the geometric method.

References

[1] K. Lingemann, A. Nüchter, J. Hertzberg, and H. Surmann, "High-speed laser localization for mobile robots," *Robotics and Autonomous Systems*, vol. 51, pp. 275–296, 2005.

[2] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, pp. 1309–1332, 2016.

[3] M. A. Zamora-Izquierdo, D. Bétaille, F. Peyret, and C. Joly, "Comparative study of extended kalman filter, linearised kalman filter and particle filter applied to low-cost gps-based hybrid positioning system for land vehicles," *IJIIDS*, vol. 2, pp. 149–166, 2008.

[4] S. A. Berrabah, Y. Baudoin, and H. Shali, "Mutli-sensor slam approach for robot navigation," *Sensors Transducers Journal*, vol. 9, no. Special Issue, pp. 200–213, 2010.

[5] F. Aghili and C. Y. Su, "Robust relative navigation by integration of icp and adaptive kalman filter using laser scanner and imu," *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 4, pp. 2015–2026, 2016.

[6] R. Li, J. Liu, L. Zhang, and Y. Hang, "Lidar/mems imu integrated navigation (slam) method for a small uav in indoor environments," *2014 DGON Inertial Sensors and Systems (ISS)*, pp. 1–15, 2014.

[7] J. Tang, Y. Chen, X. Niu, L. Wang, L. Chen, J. Liu, C. Shi, and J. Hyyppä, "Lidar scan matching aided inertial navigation system in gnss-denied environments," *Sensors*, vol. 15, no. 7, pp. 16 710–16 728, 2015.

[8] H. Zhao and Z. Wang, "Motion measurement using inertial sensors, ultrasonic sensors, and magnetometers with extended kalman filter for data fusion," *IEEE Sensors Journal*, vol. 12, pp. 943–953, 2012.

[9] J. Peng, W. Xu, B. Liang, and A. Wu, "Pose measurement and motion estimation of space non-cooperative targets based on laser radar and stereo-vision fusion," *IEEE Sensors Journal*, vol. 19, no. 8, pp. 3008–3019, April 2019.

[10] P. K. Yoon, S. Zihajehzadeh, B.-S. Kang, and E. J. Park, "Robust biomechanical model-based 3-d indoor localization and tracking method using uwb and imu," *IEEE Sensors Journal*, vol. 17, pp. 1084–1096, 2017.

[11] W. Liu, "Lidar-imu time delay calibration based on iterative closest point and iterated sigma point kalman filter," *Sensors*, vol. 17, no. 3, p. 539, 2017.

[12] G. Grisetti, C. Stachniss, and W. Burgard, "Improved techniques for grid mapping with rao-blackwellized particle filters," *IEEE Transactions on Robotics*, vol. 23, pp. 34–46, 2007.

[13] S. Kohlbrecher, O. von Stryk, J. Meyer, and U. Klingauf, "A flexible and scalable slam system with full 3d motion estimation," *2011 IEEE International Symposium on Safety, Security, and Rescue Robotics*, pp. 155–160, 2011.

[14] W. Hess, D. Kohler, H. Rapp, and D. Andor, "Real-time loop closure in 2d lidar slam," *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1271–1278, 2016.

[15] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time." in *Robotics: Science and Systems*, vol. 2, 2014, p. 9.

[16] M. Bosse, R. Zlot, and P. Flick, "Zebedee: Design of a spring-mounted 3-d range sensor with application to mobile mapping," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1104–1119, 2012.

[17] S. Wang, R. Clark, H. Wen, and A. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2043–2050, 2017.

[18] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6619, 2017.

[19] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7286–7291, 2018.

[20] P. Yin, Y. He, L. Xu, Y. Peng, J. Han, and W. Xu, "Synchronous adversarial feature learning for lidar based loop closure detection," *2018 Annual American Control Conference (ACC)*, pp. 234–239, 2018.

[21] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. D. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3223–3230, 2017.

[22] A. Nicolai, R. Skeele, C. Eriksen, and G. A. Hollinger, "Deep learning for laser based odometry estimation," in *RSS workshop Limits and Potentials of Deep Learning in Robotics*, 2016.

[23] H. Deng, T. Birdal, and S. Ilic, "Ppfnet: Global context aware local features for robust 3d point matching," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 195–205, 2018.

[24] J. Vongkulbhisal, B. I. Ugalde, F. D. la Torre, and J. P. Costeira, "Inverse composition discriminative optimization for point cloud registration," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2993–3001, 2018.

[25] G. Elbaz, T. Avraham, and A. Fischer, "3d point cloud registration for localization using a deep neural network auto-encoder," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2472–2481, 2017.

[26] J. Li, H. Zhan, B. M. Chen, I. D. Reid, and G. H. Lee, "Deep learning for 2d scan matching and loop closure," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 763–768, 2017.

[27] S. Hosseinyalamdary, "Deep kalman filter: Simultaneous multi-sensor integration and modelling; a gnss/imu case study," *Sensors (Basel, Switzerland)*, vol. 18, no. 5, 2018.

[28] G. Agresti, L. Minto, G. Marin, and P. Zanuttigh, "Deep learning for confidence information in stereo and tof data fusion," *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 697–705, 2017.

[29] D. Xu, D. Anguelov, and A. Jain, "Pointfusion: Deep sensor fusion for 3d bounding box estimation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 244–253, 2018.

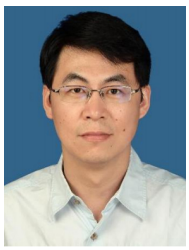
[30] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 641–656.

[31] J. R. Rambach, A. Tewari, A. Pagani, and D. Stricker, "Learning to fuse: A deep learning approach to visual-inertial camera pose estimation," *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 71–76, 2016.

[32] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem." in *AAAI*, 2017, pp. 3995–4001.

Chi Li received the bachelors degrees in control theory and engineering from Dalian University of Technology, P. R. China, in 2013. He is a Ph.D. candidate in the School of Control Science and Engineering, Dalian University of Technology, P. R. China. His research interests are in laser SLAM and deep learning in robotics.

Sen Wang is an Assistant Professor in Robotics and Autonomous Systems at Heriot-Watt University and a faculty member of the Edinburgh Centre for Robotics. Previously, he was a post-doctoral researcher at the University of Oxford. His research focuses on robot perception and autonomy using probabilistic and learning approaches, especially autonomous navigation, robotic vision, SLAM and robot learning. His research has been published in a number of flagship venues and been awarded a Best Paper Award and an Outstanding Paper Award.



Yan Zhuang (M11) received the bachelors and masters degrees from Northeastern University, Shenyang, China, in 1997 and 2000, respectively, and the Ph.D. degree from the Dalian University of Technology, Dalian, China, in 2004, all in control theory and engineering. He joined the Dalian University of Technology in 2005, as a Lecturer and became an Associate Professor in 2007. He is currently a Professor with the School of Control Science and Engineering, Dalian University of Technology. His current research interests include mobile robot 3-D environment perception and mapping, outdoor scene understanding and machine learning in robotics.

Fei Yan (M15) received the bachelors and Ph.D. degrees in control theory and engineering from the Dalian University of Technology, Dalian, China, in 2004 and 2011, respectively. In 2013, he joined the Dalian University of Technology, as a Post-Doctoral and became a Lecturer in 2015, where he is currently an Associate Professor with the School of Control Science and Engineering. His current research interests include 3-D mapping, path planning, and semantic scene understanding.