High-Precision Depth Estimation Using Uncalibrated LiDAR and Stereo Fusion

Kihong Park, *Student Member, IEEE*, Seungryong Kim, *Member, IEEE*, and Kwanghoon Sohn, *Senior Member, IEEE*

***Abstract*— We address the problem of 3D reconstruction from uncalibrated LiDAR point cloud and stereo images. Since the usage of each sensor alone for 3D reconstruction has weaknesses in terms of density and accuracy, we propose a deep sensor fusion framework for high-precision depth estimation. The proposed architecture consists of calibration network and depth fusion network, where both networks are designed considering the trade-off between accuracy and efficiency for mobile devices. The calibration network first corrects an initial extrinsic parameter to align the input sensor coordinate systems. The accuracy of calibration is markedly improved by formulating the calibration in the depth domain. In the depth fusion network, complementary characteristics of sparse LiDAR and dense stereo depth are then encoded in a boosting manner. Since training data for the LiDAR and stereo depth fusion are rather limited, we introduce a simple but effective approach to generate pseudo ground truth labels from the raw KITTI dataset. The experimental evaluation verifies that the proposed method outperforms current state-of-the-art methods on the KITTI benchmark. We also collect data using our proprietary multi-sensor acquisition platform and verify that the proposed method generalizes across different sensor settings and scenes.**

***Index Terms*— Depth estimation, multi-modal sensor fusion, on-line calibration, real-time system, 3D reconstruction.**

I. Introduction

PERCEIVING the 3D geometric configuration of scenes is essential for numerous tasks in many robotics and computer vision applications, such as autonomous driving vehicles [2], mobile robots [3], localization and mapping [4], obstacle avoidance and path planning [5], and 3D reconstruction [6].

To estimate reliable depth information of a scene, two techniques are generally utilized, including time-of-flight (TOF), such as RGB-D sensors [7] or 3D LiDAR scanners [8]; and triangulation using passive matching algorithms on stereo images [1]. For challenging outdoor scenarios, 3D LiDAR scanners [8] have become practical solutions for 3D perception since RGB-D sensors, such as Kinect 2 [7], often fail in the presence of sunlight [9] and provide limited sensing range. 3D perception with LiDAR scanners can provide very accurate depth information with errors of the order of centimeters.

However, 3D reconstruction using LiDAR is somewhat limited in practice. One reason is that LiDAR density is sparse to enable covering all relevant objects in a scene. For example, the popular HDL-64 LiDAR model covers less than 6% of the total depths of image points. Although there have been various efforts to interpolate depth information from sparse 3D depth points [8], the performance remains limited. Another limitation is that LiDAR cannot acquire color information, which can provide useful cues to understand and perceive the scene.

Another alternative for 3D reconstruction is the triangulation based on stereo matching algorithms, which provide dense depth information with corresponding color information. However, 3D reconstructions from high resolution stereo images for many of the top ranked methods on the KITTI benchmark [10] are impractical due to their high computational complexity. Moreover, the sensing range of stereo depth estimation, even using state-of-the-art methods [1], [11], is substantially limited because its depth error grows quadratically with distance from the camera origin [12]. For this reason, the sensing ranges of commercial stereo cameras are generally short compared to that of the LiDAR sensor.

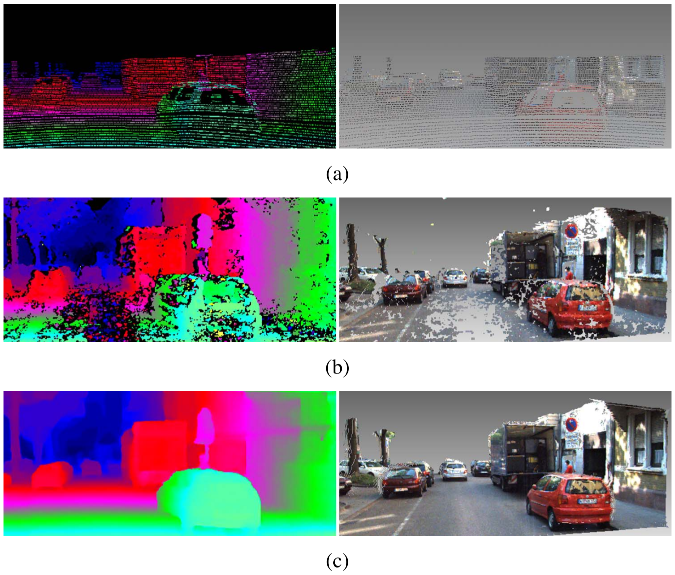
Deep convolutional neural networks (CNNs) have recently become popular in many robotics and computer vision applications [13], [14], and have been used to establish reliable dense disparity maps from stereo images, such as MC-CNN [11], with highly improved performance compared to conventional hand-crafted methods, such as SGM [1]. Several methods have interpolated depth information from sparse LiDAR point clouds by leveraging deep CNNs [15]. Compared to conventional methods [1], [8], deep CNN techniques [11], [15] achieve more accurate depth information under challenging outdoor environments. However, methods defined only on stereo images [11] or sparse depth [15] cannot simultaneously overcome both domain limitations [16]. Furthermore, CNN based methods require high computational complexity and memory usage, and hence are generally not practical for mobile systems.

Therefore, optimal fusion of LiDAR and stereo depth information could provide a practical solution to estimate high precision depth by leveraging complementary properties of each approach, as shown in Fig. 1. To construct a practical fusion system for LiDAR and stereo depth information, the current study focused on LiDAR-Stereo extrinsic calibration, and

Manuscript received June 1, 2018; revised November 16, 2018; accepted December 21, 2018. This work was supported by the Institute for Information and Communications Technology Promotion (IITP) funded by the Korean Government (MSIP) under Grant 2016-0-00197. The Associate Editor for this paper was H. G. Jung. *(Corresponding author: Kwanghoon Sohn.)*

The authors are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, South Korea (e-mail: khpark7727@yonsei.ac.kr; srkim89@yonsei.ac.kr; khsohn@yonsei.ac.kr).

Digital Object Identifier 10.1109/TITS.2019.2891788



(a)

(b)

(c)

Fig. 1. Complementary (a) sparse LiDAR and (b) (semi) dense stereo disparity characteristics, and (c) outcomes for the proposed method to fuse them within deep CNNs.

LiDAR-Stereo depth fusion. Online extrinsic calibration compensates for calibration quality degradation caused by practical driving conditions and enables the sensor fusion system to operate for longer periods. LiDAR-Stereo depth fusion solves inherent limitations of the individual sensors by selectively using complementary input sensor data, and obtains optimal depth information for outdoor 3D reconstruction. Since both approaches have shown limited performance with hand-crafted modules [16], [17], we improve their performance by leveraging high discriminative power of deep neural network. It is also essential to formulate the problems in a unified system and improve the efficiency of its architecture for practical applications.

This paper presents a CNN architecture to estimate high precision depth information by jointly utilizing uncalibrated LiDAR point clouds and stereo images. The proposed architecture consists of calibration and depth fusion networks, each addressing the two main problems discussed above while improving the performance of conventional approaches. By performing efficient stereo matching preprocessing, the calibration network reformulates the complex multi-modal calibration in the depth domain, simplifying the overall calibration process. Thus, accurate calibration parameters are estimated even with shallow network architecture. The depth fusion network fuses two complementary disparity maps by estimating reliable depth information sequentially. Incorporating the dilated convolution layer [18] maximizes the receptive field of the cascaded depth estimation modules and obtains improved depth information with compact network parameterization. Since limited training data is available for LiDAR and stereo fusion, we build a large dataset with pseudo ground truth labels, densified with raw LiDAR scans and the disparity map from an off-the-shelf stereo matching algorithm and its corresponding confidence, based on the raw KITTI benchmark [19]. Experimental results verified that the proposed approach outperformed current stereo disparity estimation [1], [11], LiDAR interpolation [20], LiDAR and camera calibration [21], and LiDAR and stereo depth fusion [16] methods on the KITTI benchmark [10]. We also constructed YONSEI datasets and evaluated stability and robustness for the proposed approach across different LiDAR channels and camera models.

This manuscript extends its preliminary conference version [22] with the following major differences: (1) Since daily changing calibration parameters may cause calibration errors between two input sensors [21], we incorporate an online calibration process into the depth fusion system. (2) In contrast to current LiDAR-camera calibration algorithms [16], we reformulate the calibration process in the depth domain by estimating depth information from stereo images, enhancing calibration accuracy even with a small number of network parameters. (3) We report qualitative and quantitative comparisons for the proposed method with current state-of-the-art approaches on various datasets.

The remainder of this paper is organized as follows. Section II describes related works in the field of depth estimation from LiDAR and stereo sensors. Section III presents the proposed uncalibrated LiDAR-Stereo depth fusion system, and Section IV describes the training method for the proposed networks. Section V provides experimental results and discussions, and Section VI summarizes and concludes the paper.

II. Related Work

This section reviews related studies for 3D reconstruction with color camera or LiDAR sensor, or both.

A. Depth Interpolation

3D LiDAR sensors are commonly employed to map outdoor scenes because of their high acquisition accuracy. However, since LiDAR data is sparse and incomplete, it is unsuitable for 3D reconstruction. Many approaches have been proposed to interpolate the sparse depth points and achieved reliable performance. These studies can be broadly divided into non-guided and guided interpolation approaches.

Early non-guided interpolation studies estimate high-resolution depth information by finding similar patches [23], [24]. CNN based methods [25] have recently been shown to outperform conventional interpolation techniques in terms of accuracy and efficiency.

Guided interpolation approaches leverage structure information from high resolution color images based on the assumption that color and depth are structurally similar [26]. The most popular approach is guided bilateral filtering [27], with many variants [20], [28], [29], due to its efficiency and its reliable performance. A color guided end-to-end model was recently proposed in [15], which also surpassed conventional algorithms. However irregular input patterns and LiDAR data sparsity compromises guided interpolation depth estimation performance. To overcome these limitations, various studies analyzed and formulated LiDAR data characteristics. Premebida *et al.* [8] penalized LiDAR points Euclidean distances and ranges to model positional relationships and uncertainties in sensor returns, respectively. To deal with various sparse data patterns, Drozdov *et al.* [30] proposed an interpolation scheme based on the total generalized variation (TGV) [31] and superpixel processing. Uhrig *et al.* [32] addressed the problem of sparsity and irregular data patterns in a deep learning formulation, and proposed a convolutional layer that calculated weights according to their locations.

B. Stereo Matching

In the field of computer vision, methods to estimate depth information from a stereo camera have been a major focus. Local methods for patch-level comparison were initially employed [19], [33], but local stereo matching methods often fail in challenging scenarios, such as weakly textured or saturated regions. Therefore, recent studies [34] have concentrated on global methods, considering smoothness constraints between neighboring pixels. Among these algorithms, SGM [1] based approaches remain one of the most popular algorithms for practical applications, from self-driving cars to autonomous surveillance, due to its computational efficiency, accuracy, and simplicity. CNN models have been proposed recently for accurate depth estimation, and MC-CNN [11] showed excellent performance on the KITTI benchmark [10] using CNN based features at patch level. However, the computational complexity of this model is higher than that of SGM, so it is burdensome for commercial systems. In contrast, Kuzmin *et al.* [35] employed very fast classical matching scores in a CNN model and achieved real-time performance, but accuracy was unacceptable for 3D reconstruction.

C. LiDAR and Camera Calibration

Multi-modal sensor fusion approaches have been studied to leverage complementary properties of sensors in various applications [13]. Among them, LiDAR and camera fusion is one of the most frequently considered setups in outdoor situations [36], preceded by extrinsic calibration to align the coordinate systems. Unnikrishnan *et al.* [37] first addressed LiDAR and camera calibration, proposing an interactive solution based on manually marking corresponding points. To reduce the manual effort, Naroditsky *et al.* [38] proposed an automatic calibration approach exploiting reflectance measurements from the LiDAR scanner. Geiger *et al.* [39] also proposed an automatic single shot approach leveraging multiple chess boards and reduced the number of recordings. However, these offline calibration approaches are not practical solutions to deal with extrinsic parameters that change daily because they require restricted conditions, such as a calibration target, special room with blackout windows, or hand labeling.

Several recent studies have proposed general calibration processes for the online scenario to correct calibration errors based on concurrently viewed objects from LiDAR and camera sensors. Bileschi [17] detected contours on projected depth and image and aligned them. Pandey *et al.* [40] corrected calibration parameters based on mutual information between LiDAR reflectivity and camera intensity. RegNet [21] first introduced CNNs into extrinsic calibration and formulated three conventional calibration steps using a single CNN model.

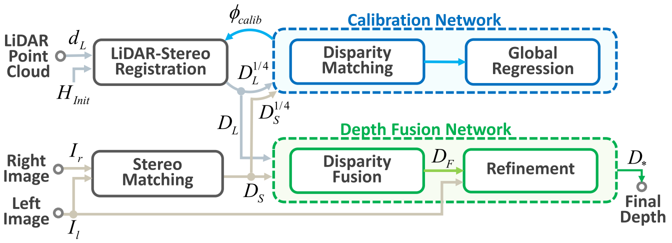


Fig. 2. Flowchart of proposed system. Our system takes the LiDAR point cloud and stereo image pair as inputs and produces the high-precision disparity as outputs.

These algorithms have the advantage of performing calibration in a variety of situations, but they do not always guarantee accuracy, particularly for highly textured surfaces and shadows, where it is difficult to establish correspondence between sensor inputs.

D. LiDAR and Stereo Fusion

In the field of robotics, data fusion techniques between 3D range sensing and stereo matching have been proposed to leverage complementary properties of their disparity maps [3]. Badino *et al.* [41] proposed an efficient framework based on dynamic programming, and Gandhi *et al.* [42] combined time-of-flight sensors and stereo cameras. However, these algorithms could not provide reliable depth information due to challenging outdoor circumstances. Maddern and Newman [16] proposed a probabilistic fusion approach for real-time applications, but performance was significantly reduced in areas without LiDAR information. To alleviate this problem, we introduce a CNN model for reliable 3D image reconstruction. Although being widely used in many computer vision and robotics applications, to the best of our knowledge CNNs have not been previously implemented in the context of LiDAR and stereo depth fusion.

III. UNCALIBRATED LiDAR-STEREO DEPTH FUSION

A. Problem Formulation and Overview

Let I_l and I_r be a pair of stereo images, and d_L be a sparse 3D point cloud represented in the world coordinate, estimated by an active 3D scanner, such as LiDAR. Given LiDAR point clouds and stereo images, the objective is to estimate a parametric model for high precision depth estimation that fuses unregistered LiDAR points with stereo disparity. Figure 2 shows the proposed system which consists of two main networks. The LiDAR projection module acts as a pre-processor, recovering the initial sparse disparity map, D_L, by projecting sparse 3D LiDAR point clouds onto the 2D image coordinates with respect to the left image, I_l, based on an initial calibration matrix, H_{init}. We leverage the dense disparity map, D_S, estimated from the stereo matching module on I_l and I_r, following [43] to compute D_S in real-time, but in principle any stereo matching algorithm could be used. The proposed system takes D_L and D_S as inputs and estimates optimal depth, D_*, for 3D reconstruction.

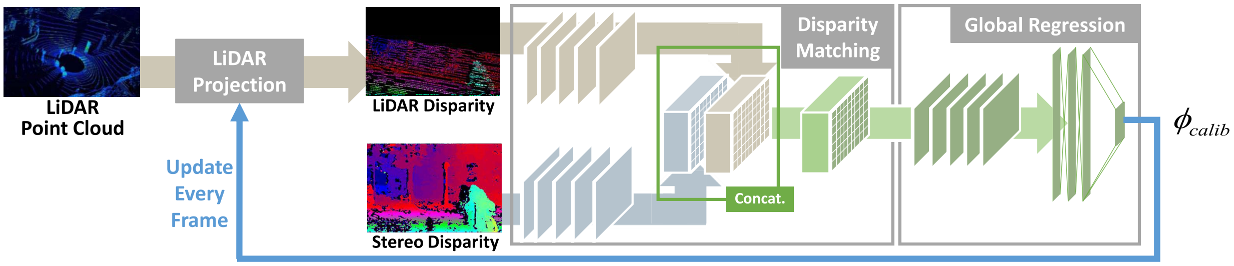


Fig. 3. Illustration of the *calibration* network. Calibration network takes LiDAR and stereo disparities as inputs and produces calibration parameters as outputs. The proposed network consists disparity matching and global regression modules. LiDAR and stereo disparities are registered based on calibration parameters updated every frame.

Therefore, we designed a two-stage fully convolutional CNN architecture to learn the parametric model for uncalibrated LiDAR and stereo depth fusion, combining *calibration* and *depth fusion network*. The calibration network estimates extrinsic calibration parameters using disparity matching and global regression modules. The disparity matching module establishes correspondences between D_L and D_S, and the global regression module updates the extrinsic calibration parameters. The network outputs well-aligned disparity maps that form the inputs for the depth fusion network. The depth fusion network fuses LiDAR and stereo disparities to obtain optimal disparity using disparity fusion and refinement modules. The disparity fusion module extracts features from each disparity and fuses them. The refinement module to estimates the residual for the initial disparity map, yielding a reliable disparity map.

We found three desirable criteria for the proposed system:

- **Accuracy**: The system should guarantee high quality 3D perception even under driving situations.
- **Speed**: The inference step should be fast, ideally achieving high processing speeds even for high resolution images.
- **Compactness**: The networks should be sufficiently compact to be deployed within mobile robots or autonomous vehicles.

The following section describes the proposed network architecture which achieves a balance between the presented criteria and demonstrates system performance compared with current state-of-the-art methods.

B. Calibration Network

The calibration network consists of two cascade sub- modules: disparity matching and global regression, as shown in Fig. 3. The architecture is based on an intuition that extrinsic calibration parameters can be calculated using disparity maps as calibration inputs. Conventional methods [21] take different modality data (i.e., color images and disparity maps) as inputs, which reduces calibration performance because their nonlinear relationships disturb the feature matching process [44]. In contrast, we simplify the problem by providing inputs in the same domain, i.e., disparity domain, hence improving extrinsic calibration accuracy and efficiency.

To estimate extrinsic calibration parameters, we define the 4×4 calibration matrix ϕ_{calib} as [21]

$$\phi_{calib} = \begin{bmatrix} \mathcal{R}(r_x, r_y, r_z) & [t_x, t_y, t_z]^T \\ 0 \quad 0 \quad 0 & 1 \end{bmatrix} \tag{1}$$

where each of r_x, r_y, and r_z represents the rotational error angle for each axis in the world coordinate, and $\mathcal{R}(r_x, r_y, r_z)$ is their corresponding rotation 3×3 matrix, and $[t_x, t_y, t_z]^T$ is a translational error vector. The calibration parameters can be represented as $\theta_{calib} = [r_x, r_y, r_z, t_x, t_y, t_z]^T$.

When the LiDAR points are projected as the input disparity D_L using H_{init} and intrinsic camera matrix, P, ϕ_{calib} corrects H_{init} error,

$$[u, v, 1]^T = P\, H_{init}\, \phi_{calib}^{-1}[x, y, z, 1], \tag{2}$$

where

$$P = \begin{bmatrix} f_u & 0 & c_u & -f_u b_s \\ 0 & f_v & c_v & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \tag{3}$$

(x, y, z) and (u, v) are a LiDAR point in the world coordinate and the corresponding pixel location in image coordinates, respectively; and $(f_u, f_v), (c_u, c_v)$, and b_s are the stereo camera focal length, principal point, and baseline, respectively.

Leveraging this updated transform function, d_L is re-projected onto D_L more accurately, and its disparity can be calculated as

$$D_L(u, v) = b_s f_u / x \tag{4}$$

Based on this formulation, we design the calibration network to estimate θ_{calib} as follows.

1) Disparity Matching Module: Intermediate features are extracted from $D_L^{1/4}$ and $D_S^{1/4}$, which are the down sampled from D_L and D_S, respectively, with scaling factor 4, and combined through concatenation and convolution layers to estimate their feature correspondence. There are two advantages to extract intermediate features from disparity inputs.

First, the network exhibits accurate calibration performance even with high sampling factors, because this approach is robust to highly textured surfaces and shadows which are common limitations of conventional multi-modal input based approaches [45]. Second, since the stereo color image is converted into disparity information, the network is robust to differences in camera models and can be generalized

TABLE I
SPECIFICATION OF THE CALIBRATION NETWORK

Layer	Kernel	Channels	*In*	*Out*	Input
		Disparity Matching Module			
conv1$_L$	11×11	1/48	1	1	$D_L^{1/4}$
conv1$_S$	11×11	1/48	1	1	$D_S^{1/4}$
conv1b$_\dagger$	1×1*	48/48	1	1	conv1$_\dagger$
pool1$_\dagger$	3×3	48/48	1	2	conv1b$_\dagger$
conv2$_\dagger$	5×5	48/128	2	2	pool1$_\dagger$
conv2b$_\dagger$	1×1*	128/128	2	2	conv2$_\dagger$
pool2$_\dagger$	3×3	128/128	2	4	conv2b$_\dagger$
conv3$_\dagger$	3×3	128/128	4	4	pool2$_\dagger$
conv3b$_\dagger$	1×1*	128/128	4	4	conv3$_\dagger$
conv3c$_\dagger$	1×1*	128/128	4	4	conv3b$_\dagger$
pool3$_\dagger$	3×3	128/128	4	8	conv3c$_\dagger$
concat.	–	128/256	8	8	pool3$_\dagger$
		Global Regression Module			
conv4	3×3	256/256	8	8	concat.
conv4b	1×1*	256/256	8	8	conv4
conv4c	1×1*	256/128	8	8	conv4b
pool4	3×3	128/128	8	16	conv4c
conv5	5×5	128/128	16	16	pool4
conv5b	1×1*	128/512	16	16	conv5
conv5c	1×1*	512/256	16	16	conv5b
pool5	3×3	256/256	16	32	conv5c
conv6	3×3	256/256	32	32	pool5
conv6b	1×1*	256/1024	32	32	conv6
conv6c	1×1*	1024/256	32	32	conv6b
conv7	1×8	256/256	32	32	conv6c
conv7b	1×1*	256/1024	32	32	conv7
T_C	1×1	1024/6	32	32	conv7b

Notes: *In* and *Out* denote input and output downscaling factors, respectively, for each layer relative to the input image. Subscripts '$_L$' and '$_S$' represent LiDAR and stereo layers, respectively. We add '*' to network in network blocks. Since feature extraction layers of LiDAR and stereo have the same network architecture, we denote them as '$_\dagger$' for clarity.

across datasets, whereas conventional CNN based methods show degraded performance when the dataset changes. This is experimentally demonstrated in Section V.

2) Global Regression Module: Since there is no need to analyze the non-linearity between data of different modality, the proposed calibration approach simplifies the network architecture. We modify the network-in-network block [46] to minimize network parameters and build the global regression module by connecting them. This simplification enhances the proposed calibration network efficiency and obtains real-time performance. The overall process of the calibration network Φ_C was constructed such that $\theta_{calib} = \Phi_C(D_L^{1/4}, D_S^{1/4})$. Table I shows the calibration network configuration.

The calibration network updates the calibration parameters every frame, and the input disparity maps are re-aligned in the LiDAR projection module to the original image resolution. Since our calibration process is performed online, the proposed fusion framework is robust to changes in calibration parameters due to external forces and timing errors. In this case, the external forces include car body torsion, temperature, and humidity changes, and the timing errors indicate errors in time stamp acquisition. By solving these problems, the proposed approach has high reliability in terms of system consistency in challenging outdoor situations.

C. Depth Fusion Network

Depth fusion network consists of two cascade sub-modules, including disparity fusion and refinement. The architecture design was inspired by two intuitions that: 1) 3D LiDAR disparity and stereo disparity encode different aspects of 3D geometric configuration, such that both information provide complementary cues that can assist to reconstruct high-precision disparity, and 2) color guidance can be utilized to boost disparity estimation performance.

To estimate a high precision disparity map efficiently, the key network design factor is incorporating the dilated convolution (DC) layer, originally developed for high level vision tasks, such as image classification and semantic segmentation [18]. A large receptive field is essential for a neural network [47], and deeper architecture [18] or larger filters [48] are easy methods to ensure a large receptive field. However, both schemes not only require more parameters, but also increase computational burden. In contrast, DC layers accomplish global information aggregation with very compact parameterization.

1) Disparity Fusion Module: The fusion module, Φ_F, consists of nine layers with three different blocks, i.e., 3×3 DC, batch normalization (BN), and rectified linear units (ReLU). The dilation factors of convolutions were set to $k = 1, 2, 4, 8, 16, 8, 4, 2,$ and 1, respectively. The 3×3 DC with factor k is a sparse filter of size $(2k + 1) \times (2k + 1)$, i.e., only 9 entries of fixed positions can be non-zero. The number of feature maps in each layer was set to 32. To encode complementary information from D_L and D_S, the fusion module takes them as inputs and extracts intermediate features through the first five layers. It is desirable that those intermediate features describe distinctive and complementary disparity cues of each channel. The intermediate features are then combined by concatenation at the 5th layer, and the final 4 layers produce the fusion module output, such that $D_F = \Phi_F(D_L, D_S)$.

2) Refinement Module: The refinement module, Φ_R has the same specification as Φ_F, i.e., 9 layers with three different 3×3 DC, BN, and ReLU blocks. In contrast to Φ_F, Φ_R is designed to enhance the D_F quality using color guidance. Another difference is that D_R does not directly compute D_*, but rather the residual $D_R = D_* - D_F$ to D_F. After adding D_F to the residual, the final disparity is $D_* = D_F + \Phi_R(D_F, I_l)$. The computation of this residual is particularly beneficial for Φ_R, since it does not need to carry the input information through the whole network [50]. Guided by I_l, Φ_R only estimates high frequency details, omitted in D_F. Table II shows the depth fusion network configuration.

Figure 5 shows example input disparity maps D_L and D_S and output disparity map D_* from the proposed network. Figure 5, row 2 shows that D_L provides sparse depth information, whereas D_S is dense but inaccurate (see Fig. 5, row 3). When LiDAR fails to acquire depth information for high reflectance material, especially in cars and windows of

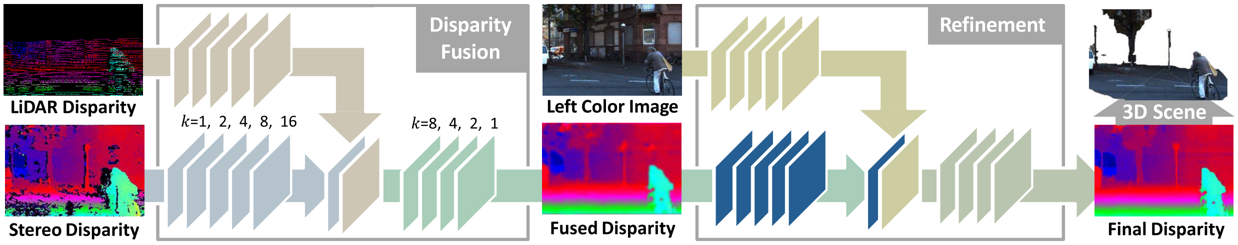


Fig. 4. The illustration of the *depth fusion network*. Depth fusion network takes LiDAR and stereo disparities as inputs and produces high precision disparity map as output. The proposed network consists disparity fusion and refinement modules.

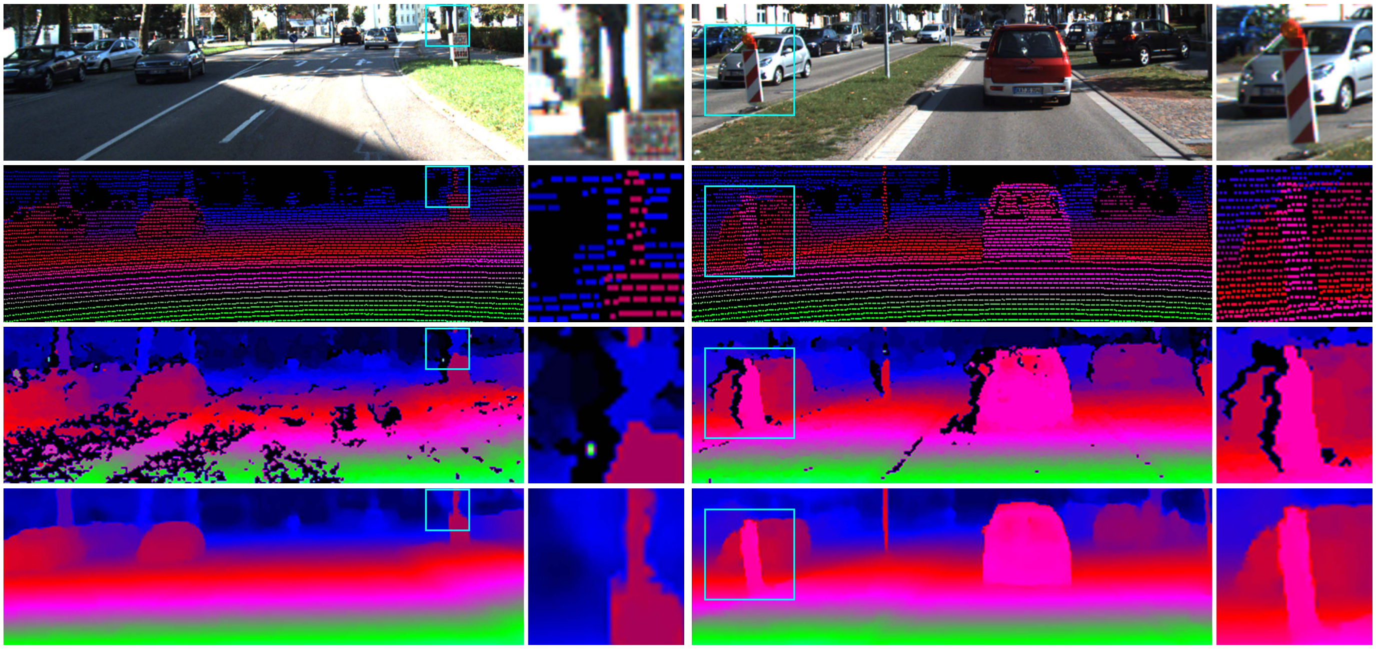


Fig. 5. Proposed system LiDAR and stereo fusion examples: (from top to bottom) Input color image, LiDAR disparity, SGM [1] results, and proposed system results.

buildings, the proposed algorithm succeeds in filling these information by fusing it with the dense stereo disparity map. When stereo matching fails to acquire depth information for thin objects, the proposed algorithm also succeeds in acquiring accurate depth information by fusing it with the sparse and accurate LiDAR disparity map. The fattening phenomenon frequently observed in stereo disparity maps [1] is also solved based on LiDAR information. By simultaneously using LiDAR and stereo disparities, the proposed network provides a dense and accurate disparity map that can be successfully employed for high precision 3D reconstruction.

IV. Training

A. Generating Training Data

Training the proposed networks requires a large dataset consisting of 3D LiDAR points, stereo images, extrinsic calibration parameters, and ground truth disparity maps. Unfortunately, there are few benchmarks including extrinsic calibration parameters and dense ground truth disparities, hence supervised learning for the proposed CNN models is problematic. Although training on indoor or synthetic datasets [51] is possible, it remains an open question if the level of accuracy obtained is sufficient for challenging outdoor situations. Therefore, we created three large training datasets based on the KITTI raw data [19] which is comprised of 42,382 stereo image frames with corresponding LiDAR point clouds and extrinsic calibration parameters for each frame.

1) LiDAR Data for Calibration: To calculate extrinsic calibration parameters between LiDAR and camera, most datasets incorporate calibration steps with offline calibration methods, e.g. [26] and [28]. These methods employ specific calibration target and hand labeling, respectively, and provide quite accurate results. However, their calibrations require high computational complexity for target setting and/or additional human modification, and are unsuitable for constructing large databases to train the proposed networks. Therefore, we constructed a calibration dataset, similar to [21]. It was done by applying the artificial calibration error θ_{gt} to the original calibration parameter of KITTI dataset. Initial LiDAR disparity maps D_L were built by projecting LiDAR point clouds d_L with the calibration parameters including θ_{gt}. This approach can generate various training data by applying randomized θ_{gt} to only a few offline calibration results. An infinite number

TABLE II
SPECIFICATION OF THE DEPTH FUSION NETWORK

Layer	Dilation	Receptive Field	Channels	Input
		Disparity Fusion Module		
conv1$_L$	1	3×3	1/32	D_L
conv1$_S$	1	3×3	1/32	D_S
conv2$_\dagger$	2	7×7	32/32	conv1$_\dagger$
conv3$_\dagger$	4	15×15	32/32	conv2$_\dagger$
conv4$_\dagger$	8	31×31	32/32	conv3$_\dagger$
conv5$_\dagger$	16	63×63	32/32	conv4$_\dagger$
concat.	–		32/64	conv5$_\dagger$
conv6	8	79×79	64/32	concat.
conv7	4	87×87	32/32	conv6
conv8	2	91×91	32/32	conv7
conv9	1	93×93	32/32	conv8
D_F		93×93	32/1	conv9
		Refinement Module		
conv10$_F$	1	3×3	1/32	D_F
conv10$_I$	1	3×3	3/32	I
conv11$_\ddagger$	2	5×5	32/32	conv10$_\ddagger$
conv12$_\ddagger$	4	15×15	32/32	conv11$_\ddagger$
conv13$_\ddagger$	8	31×31	32/32	conv12$_\ddagger$
conv14$_\ddagger$	16	63×63	32/32	conv13$_\ddagger$
concat.			32/64	conv14$_\ddagger$
conv15	8	79×79	64/32	concat.
conv16	4	87×87	32/32	conv15
conv17	2	91×91	32/32	conv16
conv18	1	93×93	32/32	conv17
conv19	1	93×93	32/1	conv18
sum				D_F, conv19

Notes: Subscripts '$_L$' and '$_S$' represent LiDAR layers and stereo layers, respectively. We denote I_l as I with a slight abuse of notation, and '$_F$' and '$_I$' represent fusion and color layers, respectively. Since LiDAR and stereo feature extraction layers have the same network architecture, we denote them as '$_\dagger$' for clarity. Similarly, fusion and color layers are denoted as '$_\ddagger$'.

of calibration training sets can be generated by varying θ_{gt}. Another advantage of the calibration approach is that we can determine a range of calibration errors. The proposed system was designed to be executed in driving situations, and we set calibration range as 20 cm and 2° similar to [21], considering the calibration quality degradation between the sensors due to external forces and timing errors.

2) LiDAR Data for Depth Fusion: Although the KITTI dataset provides depth information from raw Velodyne scans, the density of a 3D point cloud from a single frame is insufficient to train the CNN based depth fusion model. Furthermore, significant manual effort is required to remove noise due to occlusions and dynamic objects. To overcome these limitations, following [19], we accumulated the previous 11 frames of 3D point clouds to increase the density of the generated disparity map $\mathcal{D}_V$. When conflicting values occurred, we chose the disparity closest to the color capture time. The reference frame was independently interpolated using color guided interpolation [52]. Although color guided interpolation [52] leads to texture copying artifacts (Fig. 6(b)), it is robust to outliers from occlusions and dynamic objects. Therefore, we used the interpolated reference frame to determine outlier points and clean $\mathcal{D}_V$ by removing them. Fig. 6(d) shows that most outliers in $\mathcal{D}_V$ could be removed using this simple technique.

3) Stereo Data for Depth Fusion: Despite the accumulation, $\mathcal{D}_V$ contains disparity values for less than 35% of the pixels in I_l. Aside from this, disparity values were only provided for the bottom region of I_l due to inherent occlusion between the 3D LiDAR scanner and stereo camera (see Fig. 7(a)). We address these issues by leveraging a sophisticated stereo algorithm and confidence measure. Given a stereo pair, I_l and I_r, we first generate disparity maps using the state-of-the-art stereo algorithm [11], and then retain disparity values having confidence >0.95 using [49]. Figure 7(b) shows the resulting $\mathcal{D}_S$, where the density of $\mathcal{D}_S$ is higher than that of $\mathcal{D}_V$. This enables the proposed model to look at portions of the scene seldom included in $\mathcal{D}_V$.

B. Loss Function

This section describes the training procedure to find optimal network parameters for the proposed model given the training data. Although the proposed architecture consists of fully convolutional layers, training this in a single procedure from 3D LiDAR and stereo images as inputs to provide disparity map output cannot guarantee the optimal global solution due to gradient vanishing problems. To alleviate this, we employ separate loss functions for each sub-module, and formulate training schedules for each.

As described above, the loss function for the proposed networks includes three terms

$$\mathcal{L} = \mathcal{L}_{\Phi_C} + \mathcal{L}_{\Phi_F} + \mathcal{L}_{\Phi_R}. \tag{5}$$

For the loss related to the calibration network, $\mathcal{L}_{\Phi_C}$, we use the following L1-loss function

$$\mathcal{L}_{\Phi_C} = \left|\theta_{calib} - \theta_{gt}\right|_1, \tag{6}$$

which penalizes errors of estimated calibration parameters θ_{calib} from the ground truth, θ_{gt}.

The loss related to the depth fusion network, $\mathcal{L}_{\Phi_F}$, must balance $\mathcal{D}_V$ and $\mathcal{D}_S$, without over-fitting any specific scenario. First, we apply point-wise L1-loss directly to the fusion module

$$\begin{aligned}\mathcal{L}_{\Phi_F} = &\sum_{p \in \Omega(\mathcal{D}_V)} |D_F(p) - \mathcal{D}_V(p)|_1 \\ &+ \lambda \sum_{p \in \Omega(\mathcal{D}_S)} |D_F(p) - \mathcal{D}_S(p)|_1,\end{aligned} \tag{7}$$

where $\lambda > 0$ is a constant that balances the two terms: larger λ lets $\mathcal{D}_S$ contribute more to the learning parameters; p denotes spatial locations, and Ω is the set of spatial locations, including valid disparity values. During training, most $\mathcal{D}_V$ and $\mathcal{D}_S$ have some missing values, which we address by evaluating the loss only on valid points $p \in \Omega$

Secondly, since the residual learning strategy is adopted, we use the refinement module loss function as

$$\begin{aligned}\mathcal{L}_{\Phi_R} = &\sum_{p \in \Omega(\mathcal{D}_V)} |(D_R(p) + D_F(p)) - \mathcal{D}_V(p)|_1 \\ &+ \lambda \sum_{p \in \Omega(\mathcal{D}_S)} |(D_R(p) + D_F(p)) - \mathcal{D}_S(p)|_1.\end{aligned} \tag{8}$$

Note that the refinement module output is the residual, hence we need to add D_R to D_F for the final disparity.

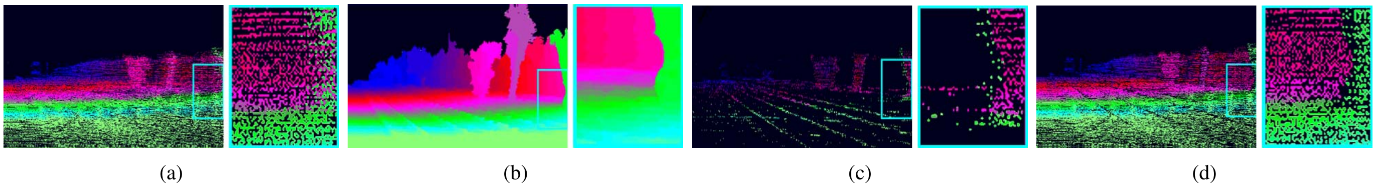


Fig. 6. Outlier removal on the raw KITTI dataset [19]. Most errors due to occlusions or reflecting surfaces can be removed using the proposed simple technique. (a) Accumulated LiDAR scans. (b) Interpolated single LiDAR scan. (c) Removed disparity points. (d) Our final $\mathcal{D}_V$.

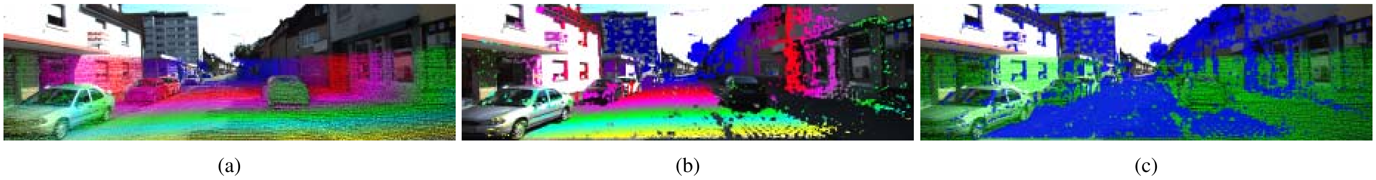


Fig. 7. Depth fusion training data examples. (a) $\mathcal{D}_V$ obtained by outlier removal and accumulation, (b) $\mathcal{D}_S$ obtained by the stereo algorithm [11] and confidence measure [49], and (c) support regions of $\mathcal{D}_V$ (green) and $\mathcal{D}_S$ (blue). The generated disparity is denser and has larger spread across the image compared to the sparse ground-truth data available in the raw KITTI dataset [19].

C. Implementation Details

The proposed model was trained from scratch with the Adam solver [53] using momentum = 0.9 and weight decay = 0.0005. The whole training procedure consists of four phases.[1] It took approximately 20 hours. First of all, we trained the *calibration network* for 50 epochs to register LiDAR and stereo sensor data. After that, the disparity fusion and the refinements module were trained sequentially for 50 epochs each with batch size of 32. When training the next module, we kept all the parameters from the previous one. Finally, the overall network was simultaneously trained. Learning rate was initialized at 1e-3 and then fixed at 1e-5 when training error stopped decreasing.

V. Experimental Results and Discussion

A. Experimental Settings

The proposed network was implemented using the VLFeat MatConvNet toolbox [55] and trained on a NVIDIA GeForce GTX TITAN X GPU. Network inputs were stereo images, and corresponding LiDAR and stereo disparity maps. Since any stereo matching algorithms could be applied in the proposed framework, considering the trade-off between efficiency and accuracy, we employed SGM [43], which is described in Section V-B2. We also employed MC-CNN [11] to build the ground truth disparity maps using the confidence estimation technique [49]. However, the outcomes are not restricted to specific algorithm choices.

We analyzed the proposed system performance by comparing with current state-of-the-art extrinsic calibration (RegNet [21]), disparity estimation from stereo images (SGM [43] and MCCNN [11]), depth interpolation from sparse disparity images (Bilateral upsampling [27], WMOF [20], and Premebida *et al.* [8]), and LiDAR-Stereo fusion (probabilistic fusion [16]) methods on the KITTI 2015 benchmark [11]. We also constructed an in-house system for outdoor 3D scene reconstruction and evaluated the proposed method on this dataset. Finally, we discuss the proposed system efficiency in terms of speed and compactness.

B. Evaluation on KITTI Dataset

1) Dataset: KITTI datasets were built by Velodyne HDL-64E LiDAR scanner and 1242 × 375 resolution stereo camera for outdoor environments, and provide ground truth calibration parameters and disparity maps. However, no raw LiDAR data was provided in the KITTI benchmark test sets. To evaluate LiDAR and stereo fusion, we used the KITTI 2015 benchmark training set [10] because its corresponding LiDAR point cloud data could be extracted from the raw KITTI dataset. Among 200 training images, 141 images were included in the raw KITTI dataset, covering 28 scenes in the raw KITTI dataset. Thus, we trained the proposed networks on the remaining 33 scenes, containing 30,159 images and corresponding LiDAR point clouds, following [56]. We used the raw data development kit [10] to project LiDAR point clouds onto the left image coordinates.

2) Calibration: Figure 8 and Table III show quantitative evaluations on the KITTI 2015 benchmark [10] using the mean squared error (MSE) metric. We first analyzed calibration performance for the proposed calibration network with two aspects: input modality and down-sampling factor (Fig. 8). When the network takes a disparity map as stereo input, average rotation error was reduced even for high down-sampling factors. This verifies that geometric relationships extracted from the two disparity maps provide improved performance because they are robust to highly textured surfaces and shadows, where non-linearity between multi-modal inputs is maximized. We also observed there was a tradeoff between accuracy and computational efficiency as the down-sampling factor changes. Average rotation error increased by a factor of 8. Thus, we fixed down-sampling factor =4 to achieve real-time speed.

[1] It is possible to train our model in an end-to-end manner. However, in practice we observed faster convergence and increased accuracy by four-phase learning similar to [54].

TABLE III
CALIBRATION ERROR ON THE KITTI 2015 BENCHMARK [10]

Architecture	Parameter (M)	Model Size (MB)	LiDAR & Stereo Domains		Rotation Error in °			Translation Error in m			Time (sec.)
			Depth - Color	Depth - Depth	r_x	r_y	r_z	T_x	T_y	T_z	
RegNet [21]	38.3	285	✓		0.54	0.23	0.16	0.05	0.07	0.02	0.019
RegNet*				✓	0.42	0.23	0.15	0.04	0.05	0.02	0.019
Ours	**4.2**	**32**	✓		0.61	0.25	0.16	0.06	0.09	0.03	**0.009**
				✓	0.46	0.23	0.15	0.04	0.06	0.02	**0.009**

Notes: Down-sampling factor for inputs = 4. RegNet* denotes where RegNet [21] architecture takes both LiDAR and stereo disparity map inputs. Although the proposed architecture shows better efficiency in terms of parameter, model size, and speed, there is only marginal accuracy differences between RegNet [21] and the proposed architecture when both LiDAR and stereo inputs are in the depth domain.

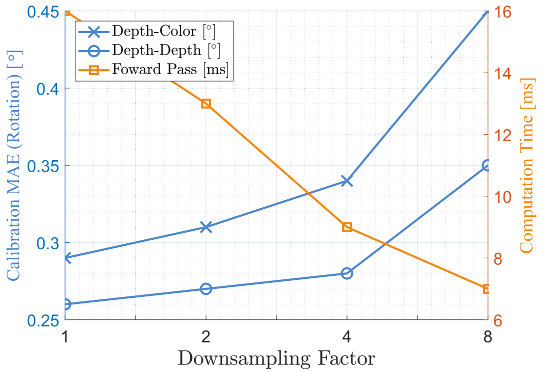


Fig. 8. Mean average error (MAE) of rotation and computation time with respect to down-sampling factors for inputs on the KITTI 2015 benchmark [10]. These results show the trade-off between accuracy and efficiency. Depth-Depth represents that LiDAR and stereo inputs are both disparity maps and shows better accuracy than Depth-Color, which uses a color image as the stereo input.

Table III compares the proposed network to RegNet [21]. Although RegNet provides reliable accuracy due to its CNN based formulation and calibration approaches, it can be improved by leveraging the proposed depth domain formulation and compact network parameterization. The proposed network achieved mean angle error =0.28°, compared to 0.31° for RegNet, and showed 8-fold less memory usage and 2-fold faster than RegNet.

Figure 9 shows the necessity to incorporate online calibration in a sensor fusion framework. Ground truths, θ_{gt}, for different days are not same (Fig. 9(b) and (c)). Thus, the positional relationship between LiDAR and stereo camera changes over time. To address this problem, we integrated the calibration and depth fusion networks into a reliable unified sensor fusion system. Section V-B3 evaluates depth estimation performance.

3) Disparity Estimation: We performed quantitative evaluations on the KITTI 2015 benchmark with bad-pixel error rate measured using the KITTI stereo development kit [10].

Stereo matching algorithms can provide high resolution depth information but have poor depth estimation accuracy for object boundaries or far away objects. Figure 10 and Table IV show that the proposed fusion technique can be applied to these stereo matching techniques to correct input disparity map errors by leveraging LiDAR information. Although MC-CNN [11] and the corresponding fusion result shows best accuracy among the stereo matching algorithms, they require high computational complexity. Since an important focus was

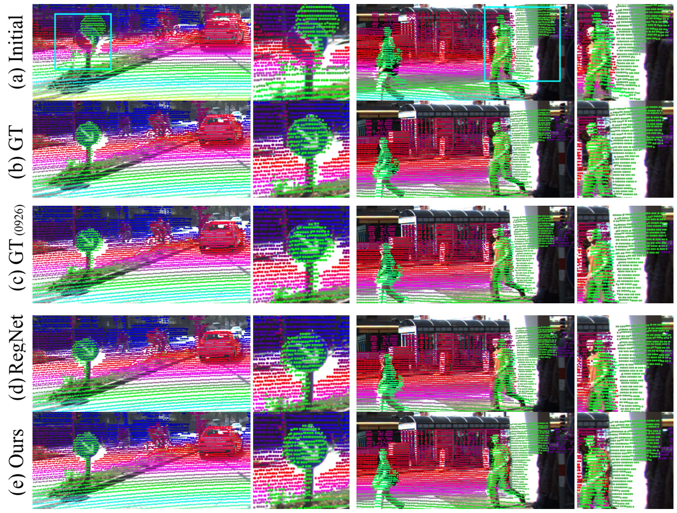


Fig. 9. Calibration result examples for an online scenario based on calibrations up to 0.2 m and 2°: (a) initial parameters, (b) ground truth, (c) ground truth for different day (2011/09/26), (d) RegNet [21], and (e) proposed method. Scenes were captured on 2011/09/28 and 2011/09/29, respectively. To deal with calibration parameters changing daily, we improved system reliability by constructing the calibration network for the online scenario.

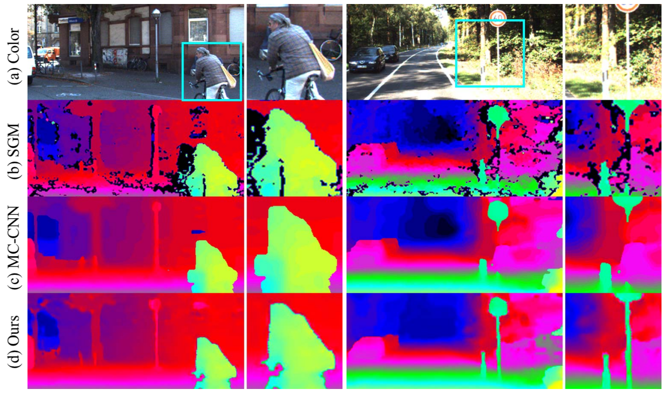


Fig. 10. Comparison with stereo matching algorithms on KITTI dataset [11]. (from up to down) Color image, the depth estimation results of SGM [1], MC-CNN [11], and proposed method (based on SGM disparity map). By leveraging LiDAR information, the proposed method solved the problems of stereo matching approaches, such as fattening effects and depth estimation failures in thin objects.

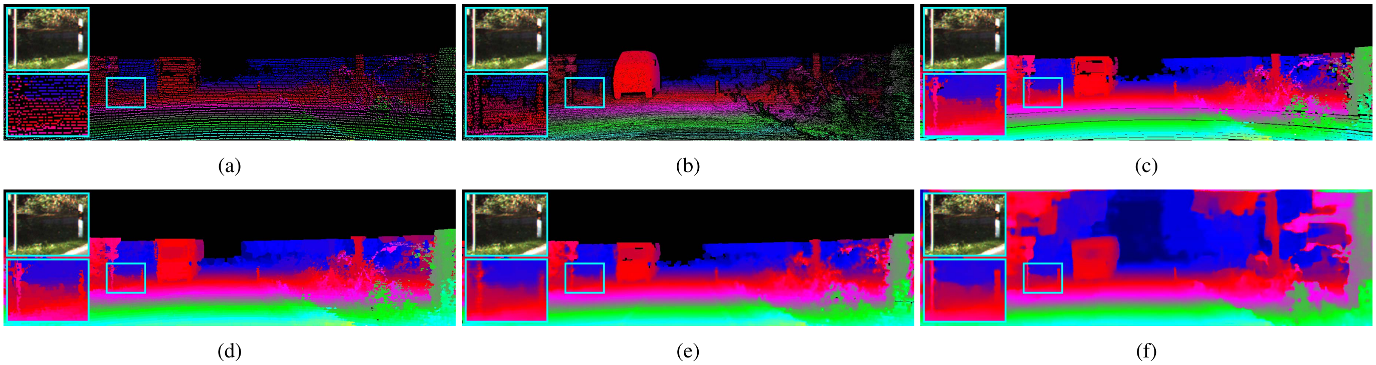


Fig. 11. Qualitative evaluation on KITTI dataset [11]. Compared to conventional interpolation methods, our network provides stable depth estimation performance. (a) Input LiDAR dispariy. (b) Ground truth disparity. (c) Bilateral upsamp. [27]. (d) WMOF [20]. (e) Premebida *etal.* [8]. (f) Ours.

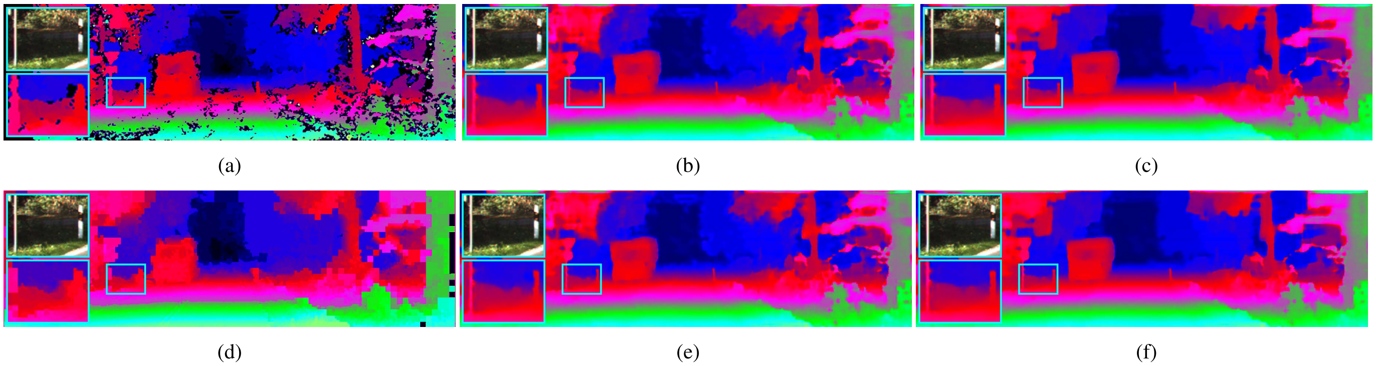


Fig. 12. Qualitative evaluation on KITTI dataset [11]. Compared to conventional LiDAR-Stereo fusion method [16], our network provides stable depth estimation performance. (a) SGM [1]. (b) D_F w/Ground truth θ_{gt}. (c) Ours w/Ground truth θ_{gt}. (d) Probabilistic fusion [16] w/Ground truth θ_{gt}. (e) D_F w/Estimated θ_{calib}. (f) Ours w/Estimated θ_{calib}.

TABLE IV
DISPARITY ERROR FOR STEREO MATCHING ALGORITHMS ON THE KITTI 2015 BENCHMARK BENCHMARK [10]

Methods	D1-bg (%)	D1-fg (%)	D1-all (%)	Time (sec.)
SGM [43]	6.77	10.12	8.11	0.003
MC-CNN [11]	5.68	6.86	6.34	0.822
Ours w/SGM	2.85	9.80	4.92	0.056
Ours w/MC-CNN	**2.24**	**9.72**	**3.94**	0.875

TABLE V
DISPARITY ERROR COMPARISON WITH DEPTH INTERPOLATION ALGORITHMS ON THE KITTI 2015 BENCHMARK [10]

Methods	D1-bg (%)	D1-fg (%)	D1-all (%)
Bilateral upsamp. [27]	3.49	19.46	6.55
WMOF [20]	3.37	15.32	6.19
Premebida et al. [8]	4.10	26.56	7.99
Ours w/SGM	**2.85**	**9.80**	**4.92**

real-time 3D reconstruction, SGM [43] is the most suitable algorithm considering the trade-off between efficiency and accuracy. By using SGM [43], the proposed fusion technique achieved 1.42% lower disparity error and 14 times faster computation time than MC-CNN [11]. Therefore, we used SGM disparity as the stereo input for all subsequent quantitative evaluations.

The proposed depth fusion network was further evaluated by comparing with current state-of-the-art depth interpolation methods, including Bilateral upsampling [27], WMOF [20], and Premebida *et al.* [8]. Table V and Fig. 12 show that the proposed approach accurately estimates depth information even in areas where LiDAR cannot provide range information, such as outside the viewing angle or high reflectance objects. In these cases, the stereo depth information supplemented the sparse LiDAR data, improving depth quality.

TABLE VI
ABLATION TEST FOR THE PROPOSED DEPTH FUSION SYSTEM ON THE KITTI 2015 BENCHMARK [10]

Methods	Calib.	D1-all (%)	Time (sec.)
SGM [43]	-	8.11	**0.003**
Probabilistic fusion [16]	Ground truth θ_{gt}	5.75	0.024
D_F	Ground truth θ_{gt}	5.24	0.025
	Estimated θ_{calib}	5.37	0.034
Ours	Ground truth θ_{gt}	**4.78**	0.047
	Estimated θ_{calib}	4.92	0.056

To evaluate each module of our depth fusion network, we measured D_F and D_* error rates, as shown in Table VI and Fig. 12. Compared to SGM [43], which is the stereo input for the network, the error rates of D_F and D_* were significantly

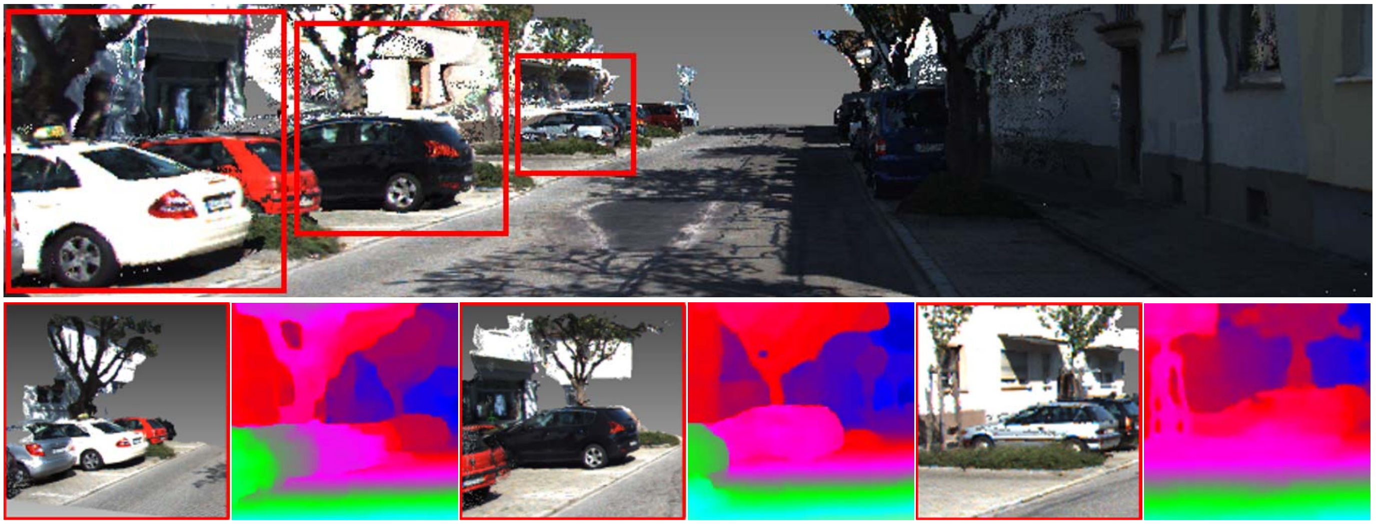
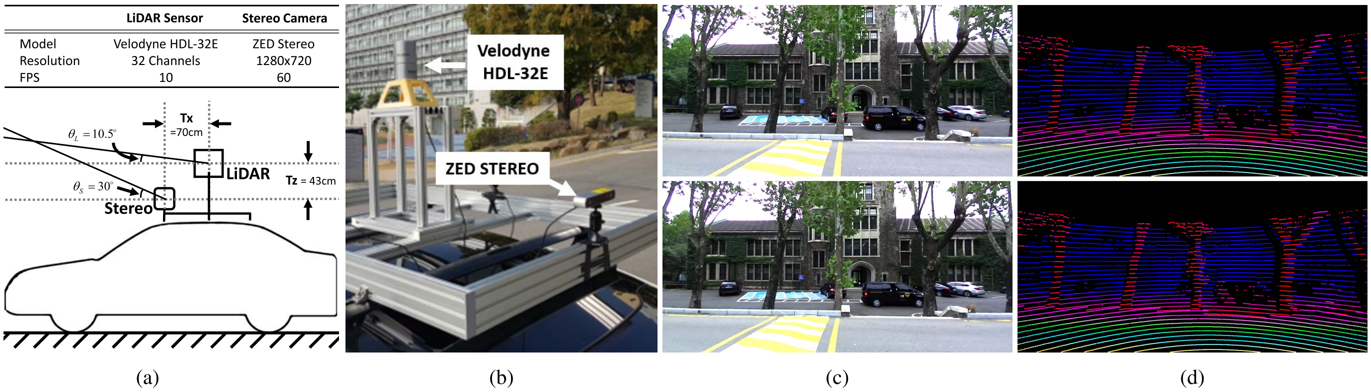


Fig. 13. Examples of outdoor 3D scene reconstruction using proposed method on KITTI dataset [19]. By leveraging complementary properties of LiDAR and stereo data, we can successfully reconstruct 3D model even in challenging outdoor conditions.



	LiDAR Sensor	Stereo Camera
Model	Velodyne HDL-32E	ZED Stereo
Resolution	32 Channels	1280x720
FPS	10	60

Fig. 14. Illustration of our system. (a) Schematic diagram, (b) our system equipped with low-channel LiDAR and stereo camera, (c) example of stereo color image acquisition, and (d) example of LiDAR disparity acquisition.

reduced by leveraging complementary information of the two input sensors and estimation performance was boosted by the color guided refinement process. The proposed network achieved error rate = 4.92%, compared with probabilistic fusion [16] error rate = 5.75%, by taking advantage of the high capacity CNN model. The overall computational time of the proposed method is two times slower than that of Probabilistic fusion [16] but is also available for real-time applications.

Thus, the experiments verify that LiDAR and stereo depth information complement each other, and the high receptive DC field boosts CNN based depth estimation performance by providing context information. We also observed positive effects for online calibration error on the disparity estimation process. Since the performance gap between disparity estimation using θ_{gt} and θ_{calib} was marginal, we adopted the online calibration process to deal with daily calibration parameter changes.

4) 3D Reconstruction: To evaluate the proposed method for practical applications, we reconstructed the 3D model using estimated depth information, as shown in Fig. 13. Since stereo disparity input accuracy reduces with the square of distance, only significant areas up to 15 m distance were visualized. The 3D reconstruction results verify that the proposed method successfully reconstructed 3D maps even for challenging outdoor environments.

C. Evaluation on YONSEI Dataset

1) YONSEI Acquisition Platform: Figure 14 (a) and (b) show the acquisition sensors mounted on top of a vehicle. Considering the vertical viewing angles of the two sensors, LiDAR was positioned above the stereo camera to obtain as much valid LiDAR data as possible. The sensor setup consisted of the following sensors.

The sensor setup consisted of the following sensors:

- **ZED stereo camera** [57]: 60 fps, 16:9 format, 1280 × 720 pixel resolution, 90° (horizontal) and 60° (vertical) field of view, ~20m depth range, 120 mm baseline
- **Velodyne HDL-32E**: 10 Hz, 0.7 million points/second, 32 channels, 0.16° angular resolution, 2cm distance accuracy, 360° (horizontal) and 41° (vertical) field of view, ~70m depth range

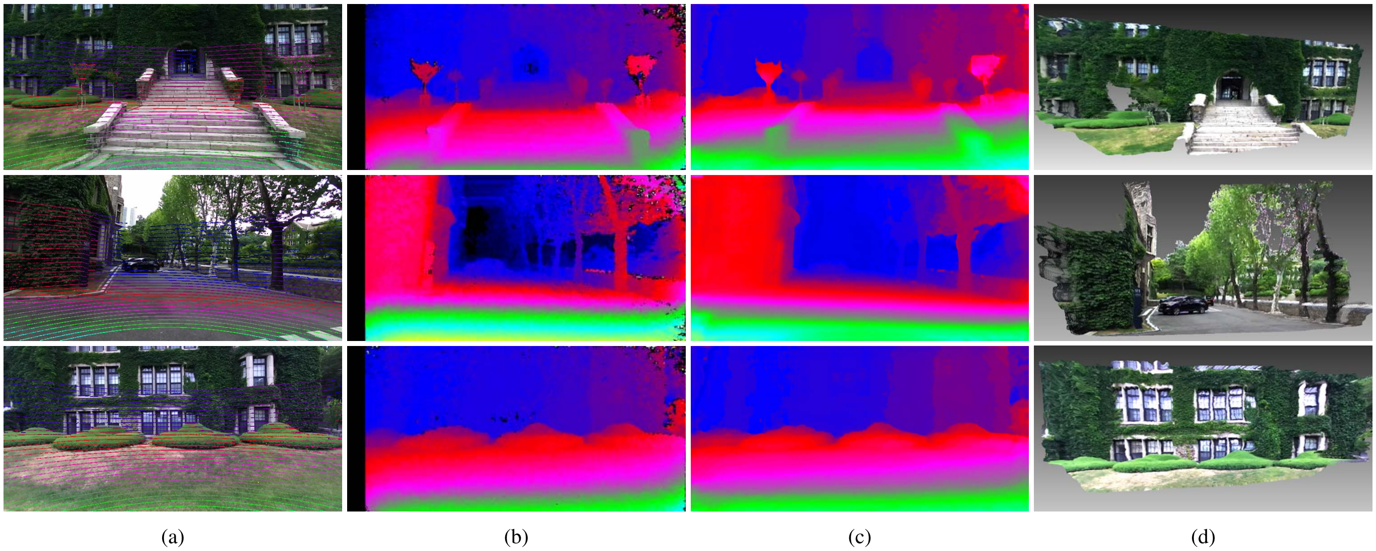


Fig. 15. Examples of outdoor 3D scene reconstructions using the proposed method on the YONSEI dataset. Despite various modality differences between the YONSEI and KITTI datasets, including color camera characteristics and the number of LiDAR channels, the proposed system trained on the KITTI dataset robustly estimated calibration parameters and depth information, and provided successful 3D reconstructions on the YONSEI dataset. (a) Calibration result. (b) Stereo disparity of [1]. (c) Fusion disparity. (d) 3D reconstruction.

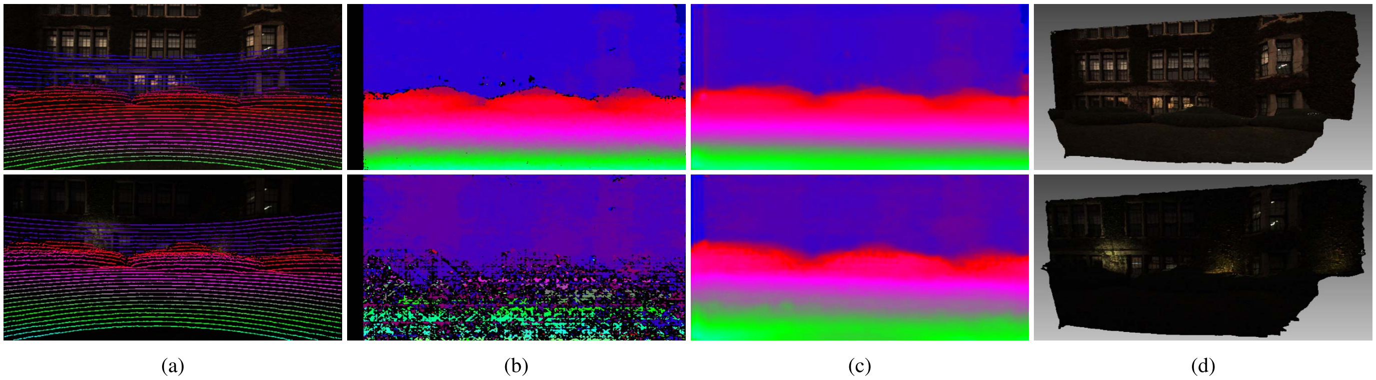


Fig. 16. Examples of outdoor 3D scene reconstructions using the proposed method on the YONSEI dataset. Despite low light conditions, the proposed system robustly estimated calibration parameters and depth information, and provided successful 3D reconstructions. (a) Calibration result. (b) Stereo disparity of [1]. (c) Fusion disparity. (d) 3D reconstruction.

The vehicle trunk housed a PC with Intel Core i7- 3770 CPU and Samsung pro-850 SSD to capture sensor data. We performed stereo camera extrinsic calibration using the Matlab toolbox [58]. The proposed extrinsic calibration method between LiDAR and stereo sensors (Section III-B) was performed by using the VLFeat MatConvNet toolbox [55]. In the LiDAR and stereo camera fusion system, the data from the two sensors must be acquired at the same time in order to fuse them. Thus, we removed the rolling shutter effect of LiDAR sensor, and performed synchronization between the two sensors by using camera motion and timestamps, similar to [10] and [59].

2) Dataset: We evaluated the proposed approach on the above multi-sensor data acquisition system for outdoor 3D scene reconstruction, acquiring various scene data under challenging outdoor environments. The resulting YONSEI dataset [60] contained 32,549 LiDAR-stereo sequential frame sets, recorded at 10 Hz. In comparison to experiments on the KITTI benchmark (Section V-C1), this dataset allows investigation of the proposed systems stability and robustness with a lower channel LiDAR sensor, which is a recent trend in low-cost 3D LiDAR scanners.

3) 3D Reconstruction: Figure 15 evaluates the proposed system, trained on the KITTI benchmark, on the YONSEI dataset. Since the KITTI and YONSEI dataset image coordinates are different, the calibration network cannot be directly applied to the YONSEI dataset. Therefore, we estimated calibration parameters by projecting YONSEI LiDAR and stereo disparity maps onto the KITTI benchmark image coordinates, obtaining YONSEI D_L and D_S as shown in Fig. 17. The transformation function between the two image coordinates can be expressed as

$$w\,[u, v, 1]^T = P\,H_{init}\;\hat{H}_{init}^{-1}\hat{P}^{-1}\hat{w}\left[\hat{u}, \hat{v}, 1\right]^T, \qquad (9)$$

where $\hat{}$ represents YONSEI pixel locations and calibration parameters.

Since the stereo input images were converted into the disparity map, the proposed calibration process can be robust

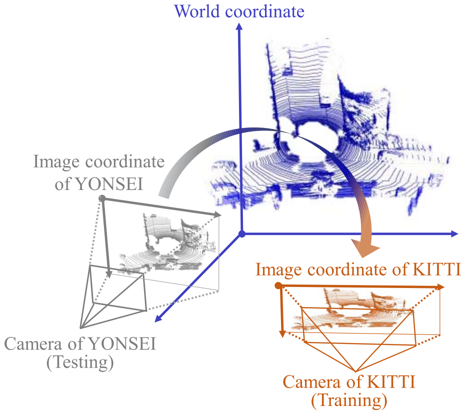

World coordinate
Image coordinate of YONSEI
Image coordinate of KITTI
Camera of YONSEI (Testing)
Camera of KITTI (Training)

Fig. 17. The transformation process from (9). To test the proposed calibration network trained on the KITTI dataset, we projected input disparity maps from the YONSEI dataset onto KITTI image coordinates.

to characteristic differences between the datasets. Despite the lower number of LiDAR channels in the YONSEI database, the proposed networks derived accurate depth information and satisfactory 3D scene reconstructions even under challenging outdoor conditions.

To evaluate the proposed system under various light environments, we also acquired data at night time as shown in Figure 16. As it becomes darker, the accuracy of input stereo disparity dramatically decreased. On the other hand, the proposed fusion system always showed reliable disparity estimation results and 3D reconstruction performance regardless of the light conditions, thanks to accurate depth information of input LiDAR disparity. This experiment demonstrated the reliability of the proposed fusion system in that it can perceive 3D information in more diverse environments than conventional single sensor systems.

D. Discussion

1) Speed: Table III and Table V compare computational complexity for the proposed system with current state-of-the-art algorithms for 1242 × 375 pixel stereo images and 64 channel LiDAR data. The proposed algorithm is significantly faster than the other algorithms, such as RegNet [21] for calibration and MC-CNN [11] for stereo matching. This verifies our main contribution that accurate depth information can be obtained efficiently from LiDAR point clouds and stereo images by adopting the proposed system. The overall process takes approximately 56 ms for one forward pass, hence the networks system is suitable for real-time applications.

2) Compactness: The proposed system showed higher efficiency in terms of memory usage compared with current state-of-the-art algorithms, requiring only 37 MB whereas RegNet [21] required 423 MB for only calibration process. This verifies the other contribution: the significantly reduced memory usage will allow other applications to be operated simultaneously. This is especially advantageous for systems with low total capacity, hence the proposed approach will be suitable for mobile devices.

3) System Reliability: The proposed system shows robust depth estimation performance than single sensor alone in various situations including the high reflectance material and low light conditions. However, even if the hardware synchronization is performed, perfect registration between two sensor data is practically difficult due to the timing error and movement of dynamic objects. Our calibration network solves the timing error by compensating displacement between two sensors caused by it. The error caused by the dynamic objects is hard to remove but can be minimized in depth fusion network by selectively using two sensor data as shown in Section V-B3.

VI. Conclusion

We presented a LiDAR-Stereo fusion system for high precision depth estimation. In contrast to previous methods, we formulated the problem of uncalibrated sensor fusion within a unified deep learning framework. To reduce complexity in the multi-modal calibration process, the proposed calibration network estimated extrinsic parameters from disparity inputs. By incorporating a dilated convolution layer in the depth fusion network, we efficiently fused depth information from the input sensors. Based on these compact parameterizations, the proposed system is suitable for various real time applications. To the best of our knowledge, this system is the first CNN model specifically designed for uncalibrated LiDAR and stereo depth fusion. We constructed a large dataset using KITTI raw LiDAR data, and removed outliers in the accumulated LiDAR. This was further augmented by adapting an off-the-shelf stereo algorithm and confidence measure. We also collected data using an in-house multi-sensor acquisition platform and verified that the proposed networks outperformed current state-of-the-art algorithms. In further work, we will extend our approach to 3D point cloud domain to avoid loss of information due to 2D projection process and improve 3D reconstruction accuracy.

References

[1] H. Hirschmuller, “Stereo processing by semiglobal matching and mutual information,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[2] T. Cao, Z.-Y. Xiang, and J.-L. Liu, “Perception in disparity: An efficient navigation framework for autonomous vehicles with stereo cameras,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2935–2948, Oct. 2015.

[3] K. Nickels, A. Castano, and C. Cianci, “Fusion of LIDAR and stereo range for mobile robots,” in *Proc. IEEE Int. Conf. Adv. Robot.*, 2003, pp. 65–70.

[4] J. Engel, T. Schöps, and D. Cremers, “LSD-SLAM: Large-scale direct monocular SLAM,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.

[5] H. Yoo, J. Son, B. Ham, and K. Sohn, “Real-time rear obstacle detection using reliable disparity for driver assistance,” *Expert Syst. Appl.*, vol. 56, pp. 186–196, Sep. 2016.

[6] A. Geiger, J. Ziegler, and C. Stiller, “StereoScan: Dense 3D reconstruction in real-time,” in *Proc. IEEE Conf. Intell. Vehicles Symp.*, Jun. 2011, pp. 963–968.

[7] Y. Kim, B. Ham, C. Oh, and K. Sohn, “Structure selective depth superresolution for RGB-D cameras,” *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5227–5238, Nov. 2016.

[8] C. Premebida, J. Carreira, J. Batista, and U. Nunes, “Pedestrian detection combining RGB and dense LIDAR data,” in *Proc. IEEE Conf. Intell. Robots Syst.*, Sep. 2014, pp. 4112–4117.

[9] M. Sharma, S. Chaudhury, and B. Lall, "Kinect-variety fusion: A novel hybrid approach for artifacts-free 3DTV content generation," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 2275–2280.

[10] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3061–3070.

[11] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, pp. 1–32, Apr. 2016.

[12] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys, "Variable baseline/resolution stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[13] X. Z. Chen, K. Kundu, Y. Zhu, S. Fidle, R. Urtasun, and H. Ma, "3D object proposals using stereo imagery for accurate object class detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1259–1272, May 2018.

[14] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[15] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 353–369.

[16] W. Maddern and P. Newman, "Real-time probabilistic fusion of sparse 3D LIDAR and dense stereo," in *Proc. IEEE Conf. Intell. Robots Syst.*, Oct. 2016, pp. 2181–2188.

[17] S. Bileschi, "Fully automatic calibration of LIDAR and video streams from a vehicle," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops (ICCV)*, Sep./Oct. 2009, pp. 1457–1464.

[18] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Dec. 2015, pp. 1520–1528.

[19] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 25–38.

[20] D. Min, J. Lu, and M. N. Do, "Depth video enhancement based on weighted mode filtering," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1176–1190, Mar. 2012.

[21] N. Schneider, F. Piewak, C. Stiller, and U. Franke, "RegNet: Multimodal sensor registration using deep neural networks," in *Proc. IEEE Conf. Intell. Vehicles Symp.*, Jun. 2017, pp. 1803–1810.

[22] K. Park, S. Kim, and K. Sohn, "High-precision depth estimation with the 3d LIDAR and stereo fusion," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2018, pp. 2156–2163.

[23] O. M. Aodha, N. D. F. Campbell, A. Nair, and G. J. Brostow, "Patch based synthesis for single depth image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 71–84.

[24] M. Hornácek, C. Rhemann, M. Gelautz, and C. Rother, "Depth super resolution by rigid body self-similarity in 3D," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1123–1130.

[25] G. Riegler, M. Rüther, and H. Bischof, "ATGV-Net: Accurate depth super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 268–284.

[26] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A noise-aware filter for real-time depth upsampling," in *Proc. Workshop Multi-Camera Multi-Modal Sensor Fusion*, 2008, pp. 1–12.

[27] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, Jul. 2007.

[28] J. Dolson, J. Baek, C. Plagemann, and S. Thrun, "Upsampling range data in dynamic environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1141–1148.

[29] M.-Y. Liu, O. Tuzel, and Y. Taguchi, "Joint geodesic upsampling of depth images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 169–176.

[30] G. Drozdov, Y. Shapiro, and G. Gilboa, "Robust recovery of heavily degraded depth measurements," in *Proc. 4th Int. Conf. 3D Vis.*, Oct. 2016, pp. 56–65.

[31] K. Bredies, K. Kunisch, and T. Pock, "Total generalized variation," *SIAM J. Imag. Sci.*, vol. 3, no. 3, pp. 492–526, 2010.

[32] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant CNNs," in *Proc. Int. Conf. 3D Vis.*, 2017, pp. 11–20.

[33] T. Kanade and M. Okutomi, "A stereo matching algorithm with an adaptive window: Theory and experiment," in *Proc. IEEE Int. Conf. Robot. Automat.*, Apr. 1994, pp. 1088–1095.

[34] P. Heise, S. Klose, B. Jensen, and A. Knoll, "PM-Huber: Patchmatch with huber regularization for stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2360–2367.

[35] A. Kuzmin, D. Mikushin, and V. Lempitsky, "End-to-End learning of cost-volume aggregation for real-time dense stereo," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.

[36] R. O. Chavez-Garcia and O. Aycard, "Multiple sensor fusion and classification for moving object detection and tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 525–534, Feb. 2016.

[37] R. Unnikrishnan and M. Hebert, "Fast extrinsic calibration of a laser rangefinder to a camera," Robotics Institute, Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-05-09, 2005.

[38] O. Naroditsky, A. Patterson, and K. Daniilidis, "Automatic alignment of a camera with a line scan LIDAR system," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2011, pp. 3429–3434.

[39] A. Geiger, F. Moosmann, Ö. Car, and B. Schuster, "Automatic camera and range sensor calibration using a single shot," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2012, pp. 3936–3943.

[40] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice, "Automatic extrinsic calibration of vision and LIDAR by maximizing mutual information," *J. Field Robot.*, vol. 32, no. 5, pp. 696–722, 2015.

[41] H. Badino, D. Huber, and T. Kanade, "Integrating LIDAR into stereo for fast and improved disparity computation," in *Proc. Int. Conf. 3D Imag., Modeling, Process., Vis. Transmiss.*, May 2011, pp. 405–412.

[42] V. Gandhi, J. Čech, and R. Horaud, "High-resolution depth maps based on TOF-stereo fusion," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2012, pp. 4742–4749.

[43] D. Hernandez-Juarez, A. Chacón, A. Espinosa, D. Vázquez, J. C. Moure, and A. M. López, "Embedded real-time stereo estimation via semi-global matching on the GPU," in *Proc. Int. Conf. Comput. Sci.*, 2016, pp. 143–153.

[44] S. Kim, D. Min, B. Ham, M. N. Do, and K. Sohn, "DASC: Robust dense descriptor for multi-modal and multi-spectral correspondence estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1712–1729, Sep. 2017.

[45] H.-J. Chien, R. Klette, N. Schneider, and U. Franke, "Visual odometry driven online calibration for monocular LIDAR-camera systems," in *Proc. Int. Conf. Pattern Recognit.*, Dec. 2016, pp. 2848–2853.

[46] M. Lin, Q. Chen, and S. Yan. (2013). "Network in network." [Online]. Available: https://arxiv.org/abs/1312.4400

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[48] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4040–4048.

[49] M. Poggi and S. Mattoccia, "Learning from scratch a confidence measure," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 1–13.

[50] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[51] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

[52] Y. Li, D. Min, M. N. Do, and J. Lu, "Fast guided global interpolation for depth and motion," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 717–733.

[53] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 286–301.

[54] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[55] (2014). *MatConvNet: CNNs for MATLAB*. Accessed: 2014. [Online]. Available: http://www.vlfeat.org/matconvnet/

[56] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 270–279.

[57] S. Kim, D. Min, S. Kim, and K. Sohn, "Feature augmentation for learning confidence measure in stereo matching," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 6019–6033, Dec. 2017.

[58] Caltech, Pasadena, CA, USA. (2015). *Camera Calibration Toolbox for MATLAB*. Accessed: Oct. 14, 2015. [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

[59] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[60] (2018). *YONSEI LiDAR-Stereo Dataset*. Accessed: May 25, 2018. [Online]. Available: http://diml.yonsei.ac.kr/lsyonsei/

Kihong Park received the B.S. degree in electronic engineering from Sogang University, Seoul, South Korea, in 2014. He is currently pursuing the joint M.S. and Ph.D. degree in electrical and electronic engineering with Yonsei University. His current research interests include computer vision and machine learning, in particular, depth estimation and multi-modal detection.

Seungryong Kim received the B.S. and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2012 and 2018, respectively. He is currently a Post-Doctoral Researcher in electrical and electronic engineering with Yonsei University. His current research interests include 2D/3D computer vision, computational photography, and machine learning, in particular, sparse/dense feature descriptor and continuous/discrete optimization.

Kwanghoon Sohn received the B.E. degree in electronic engineering from Yonsei University, Seoul, South Korea, in 1983, the M.S.E.E. degree in electrical engineering from The University of Minnesota, Minneapolis, MN, USA, in 1985, and the Ph.D. degree in electrical and computer engineering from North Carolina State University, Raleigh, NC, USA, in 1992. He was a Senior Research Engineer with the Satellite Communication Division, Electronics and Telecommunications Research Institute, Daejeon, South Korea, from 1992 to 1993, and a Post-Doctoral Fellow with the MRI Center, Medical School, Georgetown University, Washington, DC, USA, in 1994. He was a Visiting Professor with Nanyang Technological University, Singapore, from 2002 to 2003. He is currently an Underwood Distinguished Professor with the School of Electrical and Electronic Engineering, Yonsei University. His research interests include 3D image processing and computer vision.