2018 IEEE International Conference on Robotics and Automation (ICRA)
May 21-25, 2018, Brisbane, Australia

High-precision Depth Estimation with the 3D LiDAR and Stereo Fusion

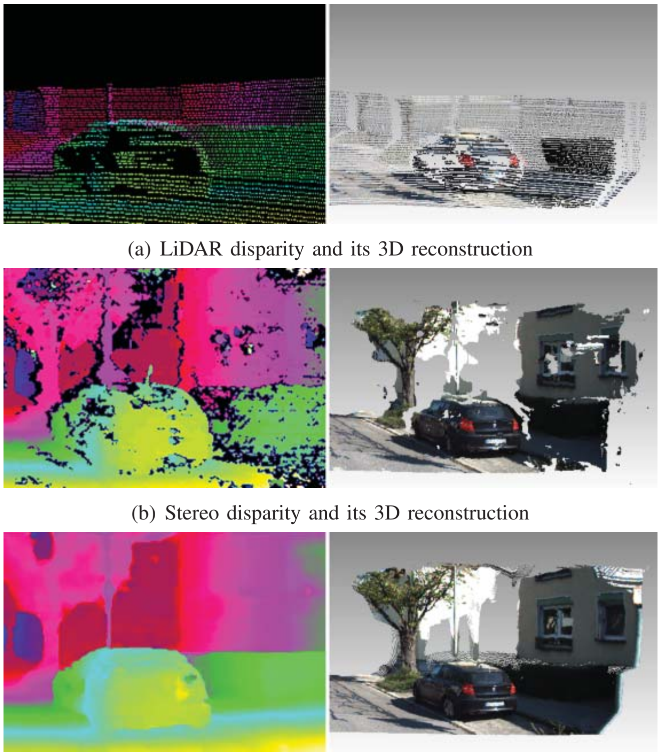
Kihong Park, Seungryong Kim, and Kwanghoon Sohn.

***Abstract*— We present a deep convolutional neural network (CNN) architecture for high-precision depth estimation by jointly utilizing sparse 3D LiDAR and dense stereo depth information. In this network, the complementary characteristics of sparse 3D LiDAR and dense stereo depth are simultaneously encoded in a boosting manner. Tailored to the LiDAR and stereo fusion problem, the proposed network differs from previous CNNs in the incorporation of a compact convolution module, which can be deployed with the constraints of mobile devices. As training data for the LiDAR and stereo fusion is rather limited, we introduce a simple yet effective approach for reproducing the raw KITTI dataset. The raw LiDAR scans are augmented by adapting an off-the-shelf stereo algorithm and a confidence measure. We evaluate the proposed network on the KITTI benchmark and data collected by our multi-sensor acquisition system. Experiments demonstrate that the proposed network generalizes across datasets and is significantly more accurate than various baseline approaches.**

I. Introduction

Perceiving 3D geometric configuration of scene or object is undoubtedly essential for numerous tasks in many robotics and computer vision applications, such as autonomous driving vehicle [19], mobile robots [15], localization and mapping [20], obstacle avoidance and path planning [21], and 3D reconstruction [22].

To estimate reliable depth information of scene, two kinds of techniques can be utilized, the use of active 3D scanners such as RGB-D sensors [24] or 3D LiDAR scanners [23] and the use of passive matching algorithms on stereo images [12]. For challenging outdoor scenarios, 3D LiDAR scanner [23] has been the most practical solution for 3D perception since the RGB-D sensor such as Kinect [24] frequently fails in the presence of sunlight [24] and provides a limited sensing range. The 3D perception with the LiDAR scanner can provide very accurate depth information with errors in terms of centimeter. Reconstructing 3D using the LiDAR would be limited in practice though. One reason is that its density is sparse to cover all salient objects in a scene since it offers fewer than 6 % of total image points. Even though there exist some efforts to interpolate depth information of sparse 3D depth points [23], its performance is also limited. Another reason is that it cannot achieve color information, which can be useful cue to understand and perceive the scene. Another alternative for 3D perception is to leverage the stereo disparity estimation from stereo images, which achieves dense depth information with corresponding color information. However, reconstructing 3D from high-resolution stereo images cannot be possible in practice due to their high computational complexity, with many of the top-ranked methods on the KITTI benchmark [14]. Moreover, the accuracy of stereo depth estimation, even using the state-of-the-art methods [12], [13], is substantially limited according to sensing ranges due to not only small baseline of stereo cameras but inherent limitations of stereo matching algorithms. Therefore, the optimal fusion technique of the 3D LiDAR and stereo depth information can be a solution to estimate high-precision depth by leveraging complementary properties of each information as in Fig. 1.



(a) LiDAR disparity and its 3D reconstruction

(b) Stereo disparity and its 3D reconstruction

(c) Fusion disparity by our method and its 3D reconstruction

Fig. 1. Visualization of complementary characteristics of the sparse 3D LiDAR and (semi-) dense stereo disparity, and the results of our proposed method to fuse them.

Over the past few years, deep convolutional neural networks (CNNs) have been become increasingly popular in many robotics and computer vision applications [4]. In estimating depth information, CNN based techniques have been also popularly used to help establish reliable dense disparity maps from stereo images, such as MC-CNN [13] that has shown highly improved performance compared to conventional handcrafted methods such as SGM [12]. Fur-

The authors are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea, {khpark7727, srkim89, khsohn}@yonsei.ac.kr. This work was supported in part by the Institute for Information and communications Technology Promotion Grant through the Korea Government (MSIP) under Grant 2016-0-00197.

978-1-5386-3081-5/18/$31.00 ©2018 IEEE

thermore, CNN based methods also have tried to interpolate sparse depth information with respect to the structures of the color guidance image [8]. Compared to conventional methods [23], [12], the aforementioned techniques [13], [8] enabled achieving more accurate depth information under challenging outdoor environments. However, those methods, defined only on stereo images [13] or sparse depth [8], inherently cannot overcome the all limitations proposed from the both domains at the same time [18]. Furthermore, these CNN based methods need high computational complexity and memory usage, and it is not practical for mobile devices.

In this paper, we present a deep CNN architecture to jointly utilize 3D LiDAR and stereo depth information for high-precision disparity estimation. This architecture differs from previous CNNs in several ways tailored to our particular problem. One is that the two complementary disparity maps are formulated as inputs in a synergistic manner, which allows each depth information can be boostly used. Another is the compact convolution architecture that can be deployed with the constraints of mobile devices. Furthermore, since limited training data is available for LiDAR and stereo fusion problem, we build a large-scale dataset using the raw LiDAR scans densified with the disparity map from an off-the-shelf stereo matching algorithm and its correspondence confidence on the raw KITTI benchmark [25]. In the experimental results, we demonstrate that our proposed network outperforms existing stereo disparity estimation methods [12], [13], LiDAR interpolation methods [33], and LiDAR and stereo depth fusion methods [18] on various benchmarks such as the KITTI [14] and our own YONSEI datasets.

II. Related Work

For reliable disparity estimation under challenging outdoor circumstances, a number of approaches have been proposed to interpolate sparse depth information such as the LiDAR points, establish dense disparity map from stereo images, and estimate disparity from the LiDAR and stereo disparity fusion as follows.

A. Depth Upsampling

3D LiDAR sensors are common in outdoor scene understanding approaches because of their high acquisition accuracy. However, since LiDAR data is sparse and incomplete, it is not suitable for 3D reconstruction. To address this problem, many approaches tried to upsample the sparse 3D points and achieved reliable performance. These studies can be divided into two major categories: non-guided upsampling and guided upsampling. Early approaches of the non-guided upsampling have leveraged repetitive structures to identify similar patches across different scales [1], [2]. Recently, CNN based methods [3] have outperformed conventional upsampling techniques in terms of accuracy and efficiency. Guided upsampling approaches use structure information of high resolution color images based on the assumption that color and depth are structurally similar [5]. One of the famous approaches of this guided upsampling is guided bilateral filtering [36]. Due to its efficiency and reliable performance, there are many variants of this scheme [36], [35]. More recently, color guided end-to-end model was proposed [8], and also suppressed conventional algorithms.

B. Stereo Matching

In the fields of computer vision, a method for estimating depth information from a stereo camera has been an another main stream. In the early stage, the local method to perform the patch unit comparison was mainly used [25], [10]. However, these local stereo matching methods often fail in challenging scenarios, such as weakly-textured or saturated regions. To solve these problems, recent researches [11] concentrated on global methods by considering smoothness constraints between neighboring pixels. Among these algorithms, the pixel-level approaches based on the SGM [12] are still one of the popular stereo matching algorithms that can be applied to practical applications, spanning from self-driving cars to autonomous surveillance, thanks to its computational efficiency, accuracy, and simplicity. Recently, CNN models have been proposed for accurate depth estimation, and MC-CNN [13] showed an excellent performance on the KITTI benchmark [14] based on CNN based features in a patch-level. However, their huge complexity are not suitable for commercial systems.

C. LiDAR and Stereo Depth Fusion

In the field of robotics, the data fusion technique between 3D range sensing and stereo matching have been proposed to leveraging complementary properties of their disparity maps [15]. Badino et al. [16] proposed efficient framework with dynamic programming, and Gandhi et al. [17] tried to fuse time-of-flight sensor and stereo camera. However, these algorithms could not provide reliable depth information due to challenging outdoor circumstances. Recently, Maddern [18] proposed a probabilistic fusion approach for real-time applications but their performance significantly decreased in the area without LiDAR information. To alleviate this problem, we introduce a CNN model for reliable 3D image reconstruction. Note that while being widely used in many computer vision and robotics applications, CNNs have never been implemented before in the context of LiDAR and stereo depth fusion to the best of our knowledge.

III. Method

A. Problem Formulation and Overview

Let I_l and I_r be a pair of stereo images. d_L is sparse 3D point clouds represented in the world coordinate, estimated by active 3D scanner such as the LiDAR. Given 3D LiDAR point clouds and stereo images, our objective is to estimate a parametric model for high-precision depth estimation, which fuses sparse 3D LiDAR points with dense stereo disparity. To end this, we first recover the sparse disparity map D_L from sparse 3D LiDAR point clouds, and then pre-process D_L with a bicubic interpolation. Moreover, we leverage dense disparity map D_S estimated from stereo matching algorithm on stereo images I_l and I_r. We follow the approach in [12] to compute D_S.

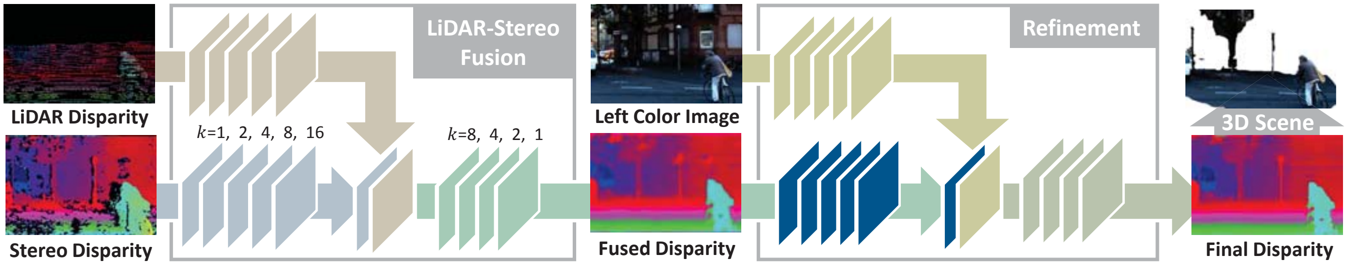
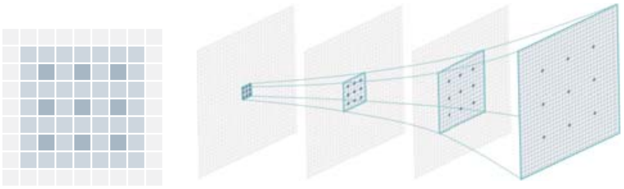


Fig. 2. Network configuration of our overall framework. Our proposed network takes LiDAR and stereo disparities as inputs and produces the high-precision disparity as outputs.



(a) 3×3 DC, k=2 (b) Expenential expension of receptive field

Fig. 3. Illustration of the dilate convolution (DC) [34]. The shaded blue pixels show the receptive field of filters. Using the DC, we can accomplish global information aggregation with a very compact parameterization.

By leveraging CNNs, we design two-stage cascade deep architecture to learn a parametric model for the 3D LiDAR and stereo depth fusion in a fully convolutional and end-to-end manner, consisting of LiDAR-Stereo fusion module and disparity refinement module. The disparity fusion network is designed to extract the features from each disparity and fuse them. Furthermore, the disparity refinement network is formulated to estimate the residual of initial disparity map, which yields an edge-preserved accurate disparity map. The whole schematic of our network is illustrated in Fig. 2.

We have found three desirable criteria for our network:

- **Accuracy**: The network should guarantee high quality 3D perception by taking advantage of the complementary characteristics of LiDAR sensing and stereo matching.
- **Speed**: The inference step should be fast, ideally achieving interactive rates for high-resolution images.
- **Compactness**: The network should be compact to be deployed within mobile robots or autonomous vehicles.

In principle, any CNN architectures with large receptive fields [26], [27] can be used for the first criterion. Most of these architectures are, however, slow and not compact [27]. In the following, we will describe our network architecture that strikes the best balance between other criteria, and demonstrate outstanding performance of our architecture in an extensive comparative study with existing methods.

B. Network Configuration

Our overall network consists of two cascade sub-networks, including LiDAR-Stereo fusion and refinement. Our architecture design is inspired by two intuitions that: 1) the 3D LiDAR disparity and stereo disparity encode different aspects of 3D geometric configuration, such that information about one provides complementary cues that can assist to reconstruct high-precision disparity information, and 2) the guidance of color information can be utilized to boost the disparity estimation performance.

To estimate a high-precision disparity map efficiently, the key design factor of our network is the incorporation of the dilated convolution (DC) module, originally developed for high-level vision tasks such as image classification and semantic segmentation [27]. It has been widely acknowledged that the large receptive field is essential for a neural network [26]. Using a deeper architecture [27] or larger filters [28] is an easy way to ensure large receptive field. However, both schemes not only require more parameters but increase the computational burden. Unlike these, the DC module enable us to accomplish global information aggregation with a very compact parameterization.

1) LiDAR-Stereo Fusion Network: he 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 k of convolutions are set to 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-zeros, exemplified in Fig. 3. The number of feature maps in each layer is set to 32. To encode complementary information from D_L and D_S, the fusion module first takes them as inputs and extracts intermediate features through first five layers. It is desirable that those intermediate features describe distinctive and complementary disparity cues of each channel. Thus, the intermediate features are then combined through element-wise summation at 5th layer, followed by last four layers to produce the output of fusion module, denoted by D_F:

$$D_F = \Phi_F(D_L, D_S). \tag{1}$$

2) Refinement Network: The refinement module Φ_R has the same specification as the fusion one, which consists of nine layers with three different blocks of 3×3 DC/BN/ReLU. Unlike the fusion module, the refinement module is designed to enhance the quality of initial disparity D_F with the help of the color guidance. Furthermore, another difference of the refinement module is that it does not directly compute the high-quality disparity D_*, but the residual $D_R = D_* - D_F$ to the input D_F. After addition of D_F to the residual, the final disparity is given by:

$$D_* = D_F + \Phi_R(D_F, I_l). \tag{2}$$

The computation of a residual in (2) is especially beneficial for the refinement module, since it does not need to carry the

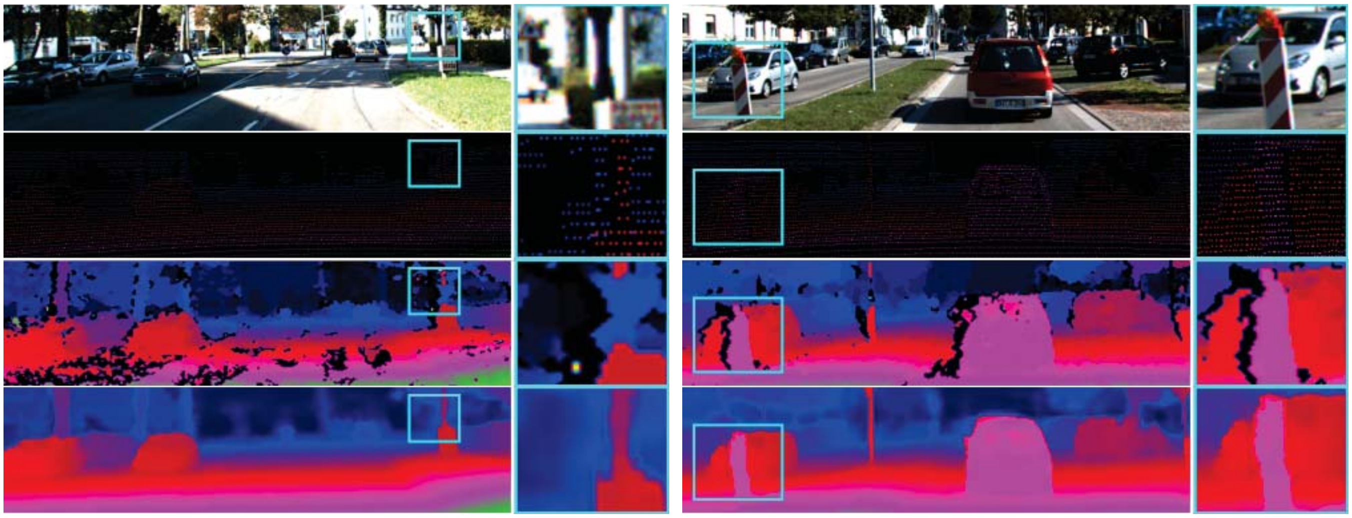


Fig. 4. Examples of our LiDAR and stereo depth fusion: (from top to down) Input color image, LiDAR disparity, the result of SGM [12], and proposed algorithm.

input information through the whole network [29]. Guided by the left color image I_l, the refinement module estimates high-frequency details only, omitted in D_F.

Fig. 4 exemplifies the input disparity maps of the LiDAR D_L and stereo D_S and the output disparity map D_* of the proposed network. As shown in second rows of Fig. 4, LiDAR disparity map D_L provides sparse depth information while stereo disparity map D_S is dense but inaccurate as in third rows of Fig. 4. When stereo matching fails to acquire depth information on a thin object, our algorithm 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 [12] is also solved based on LiDAR information. By simultaneously using the LiDAR and stereo disparity, our proposed network can provide dense and accurate disparity map that can be successfully used for high-precision 3D reconstruction.

C. Generating Training Data

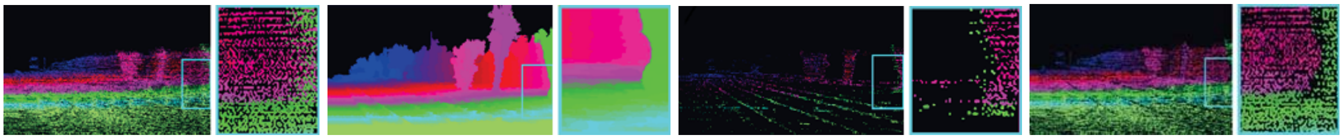
Training the proposed network requires access to a large-scale dataset, consisting of 3D LiDAR points, stereo images, and ground-truth disparity maps. However, there lacks a benchmark with dense ground truth disparities, making supervised learning of our CNN model less feasible. Although training on indoor or synthetic datasets [32] is possible, it remains an open question if the level of accuracy obtained by such datasets is sufficient to challenging outdoor situations. We therefore created two large-scale training dataset based on the KITTI raw data [25] which comprises 42,382 stereo frames with corresponding LiDAR point clouds.

1) Velodyne HDL-64E LiDAR: While the KITTI dataset provides depth information from raw Velodyne scans, the density of 3D point clouds in single frame is not sufficient to learn the CNN model. Furthermore, significant manual efforts is required to remove noise due to occlusions and dynamic objects. To overcome these limitations, we first follow [25] in manner that we accumulated 11 frames of 3D point clouds to increase the density of the generated disparity maps $\mathcal{D}_V$. When there exist conflicting values, we chose the disparity recorded closest to the color capture time. Independently, the reference frame is interpolated by using color-guided upsampling [33]. While the color-guided upsampling [33] leads to texture-copying artifacts (Fig. 5. (b)), it is robust to outliers caused by occlusions and dynamic objects. Therefore, we use the interpolated reference frame to determine outlier points and clean the accumulated disparity $\mathcal{D}_V$ by removing these outlier points. In Fig. 5. (d), it can be seen that most outliers in $\mathcal{D}_V$ can be removed with our simple technique.

2) Point Grey Flea2 Stereo Camera: Despite the accumulation, $\mathcal{D}_V$ contains disparity values for less than 35% of the pixels in left color image. Aside from this, the disparity values are provided only at the bottom part of the left color image (see Fig. 6 (a)) due to inherent occlusion problems between 3D LiDAR scanner and stereo camera. We tackle these issues by leveraging on a sophisticated stereo algorithm and a confidence measure. Given a stereo pair I_l and I_r, we first generate disparity maps using the state-of-the-art stereo algorithm [13], and then retain disparity values having confidence higher than 0.95 using [30]. The resulting disparity $\mathcal{D}_S$ is shown in Fig. 6 (b), where $\mathcal{D}_S$ is the larger spread across the whole image. This enables our model to look at portions of the scene seldom included in $\mathcal{D}_V$.

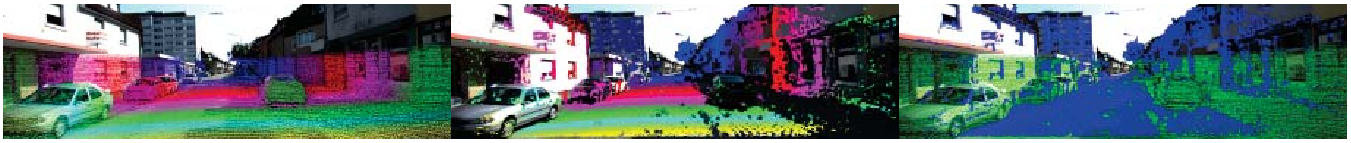
D. Training

In this section, we describe the training procedure in detail to find optimal network parameters of our model given the set of training data. Even though our architecture consists of fully-convolutional layers, training this in a single procedure from 3D LiDAR and stereo images as inputs to disparity as outputs cannot guarantee an optimal global solution due to the gradient vanishing problems. To alleviate this problem, we employ separate loss functions for each sub-network, and formulate training schedules of each sub-network.



(a) Accumulated LiDAR scans (b) Interpolated single LiDAR scan (c) Removed disparity points (d) Our final $\mathcal{D}_V$

Fig. 5. Outlier removal on the raw KITTI dataset [25]. We notice that most errors due to occlusions or reflecting surfaces can be removed with our simple technique.



(a) $\mathcal{D}_V$ from Velodyne LiDAR (b) $\mathcal{D}_S$ from stereo pair (c) Support regions of $\mathcal{D}_V$ (green) and $\mathcal{D}_S$ (blue)

Fig. 6. Examples of our training data: (a) $\mathcal{D}_V$ obtained by outlier removal and accumulation, (b) $\mathcal{D}_S$ obtained by the stereo algorithm [13] and confidence measure [30], and (c) support regions of $\mathcal{D}_V$ (green) and $\mathcal{D}_S$ (blue). Our generated disparity is more dense and is the larger spread across the whole image, compared to the sparse ground-truth data available in the raw KITTI dataset [25].

1) Loss Function: As described in above, the loss function of our network consists of two loss functions such that

$$\mathcal{L} = \mathcal{L}_{\Phi_F} + \mathcal{L}_{\Phi_R}. \tag{3}$$

The loss has to balance both of $\mathcal{D}_V$ and $\mathcal{D}_S$, and lead synergy without over-fitting to a specific scenario. First of all, we apply point-wise L1-loss directly to the fusion module:

$$\begin{aligned}\mathcal{L}_{\Phi_F} = &\sum_{p\in\Omega(\mathcal{D}_V)} |D_F(p) - \mathcal{D}_V(p)|_1 \\ &+\lambda \sum_{p\in\Omega(\mathcal{D}_S)} |D_F(p) - \mathcal{D}_S(p)|_1,\end{aligned} \tag{4}$$

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

Secondly, since the residual learning strategy is adopted, we use the following loss function for the refinement module such that

$$\begin{aligned}\mathcal{L}_{\Phi_R} = &\sum_{p\in\Omega(\mathcal{D}_V)} |(D_R(p) + D_F(p)) - \mathcal{D}_V(p)|_1 \\ &+\lambda \sum_{p\in\Omega(\mathcal{D}_S)} |(D_R(p) + D_F(p)) - \mathcal{D}_S(p)|_1.\end{aligned} \tag{5}$$

Note that the output of the refinement module is the residual. We thus need to add D_R back to D_F for the final disparity.

2) Training schedule: Our model is trained from scratch with Adam solver [31] using a momentum of 0.9 and a weight decay of 0.0005. The whole training procedure consists of two phases[1]. We sequentially train the fusion and refinement modules for first 50 epochs each with a batch size 32. While training the refinement module, we keep the all parameters in the fusion one. The learning rate is started from 0.001 and then fixed to 0.0001 when the training error stops decreasing. It takes about 10 hour to train the whole modules.

IV. Experimental Results and Discussion

A. Experimental Settings

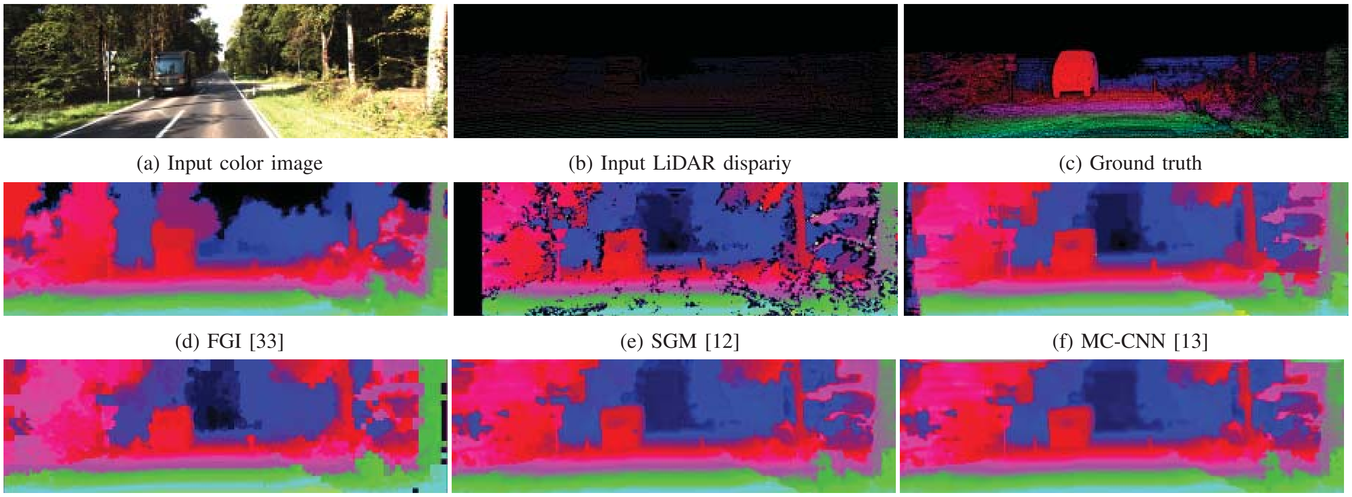
For our experiments, our network was implemented using VLFeat MatConvNet toolbox [38] and trained on a NVIDIA GeForce GTX TITAN X GPU. The inputs of each subnetwork were randomly cropped patches from the training stereo images, the corresponding LiDAR and stereo disparity maps. The input patch sizes were 64 × 64 × 3 for color images and 64 × 64 × 1 for disparity maps. As any other stereo matching algorithms can be applied in our framework, we employed the SGM [12], considering the trade-off between efficiency and accuracy. We also employed MC-CNN [13] to build the ground-truth disparity maps with the confidence estimation technique [30]. However they are not restricted to specific choices of the algorithm.

In the following, we intensively analyze the performance of our method through comparisons to the state-of-the-art methods of disparity estimation from stereo images [13], sparse disparity interpolation from sparse disparity images [33], and LiDAR and stereo depth fusion [18] on the KITTI benchmark [13] in Section IV-B. We also constructed our own system for outdoor 3D scene reconstruction in Section IV-C and evaluated our proposed method on those dataset.

B. Evaluation on KITTI Dataset

1) Dataset: The KITTI datasets were built by Velodyne HDL-64E LiDAR scanner and 1242 × 375 resolution stereo camera under outdoor environments. For an evaluation, we used the training set of the KITTI stereo evaluation 2015 [14], which provides ground truth disparity maps. However, since no raw LiDAR data was provided in this benchmark, we extracted corresponding LiDAR point cloud data from the raw KITTI dataset. Among 200 training images, 141 images are included in the raw KITTI dataset. These images cover 28 scenes in the raw KITTI dataset, thus, we trained our network on remaining 33 scenes contain 30,159 images by following [37]. We made use of the raw data development

[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 two-phase learning.



(a) Input color image

(b) Input LiDAR dispariy

(c) Ground truth

(d) FGI [33]

(e) SGM [12]

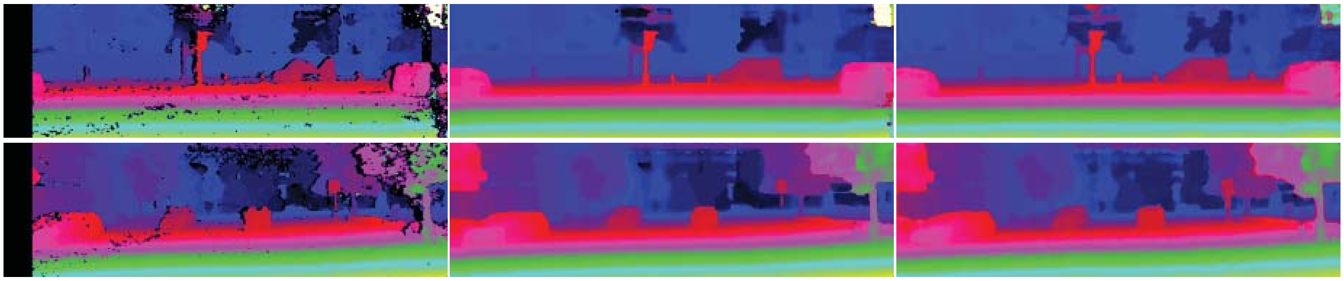
(f) MC-CNN [13]

(g) Probabilistic fusion [18]

(h) Ours wo/Φ_R

(i) Ours

Fig. 7. Qualitative evaluation on KITTI dataset [13].



(a) SGM [12]

(b) MC-CNN [13]

(c) Ours

Fig. 8. Comparison with stereo matching algorithms on KITTI dataset [13]

TABLE I

COMPARISON OF QUANTITATIVE EVALUATION ON THE KITTI BENCHMARK [14].

Methods	Error (%)	Density. (%)
Bicubic upsamp.	15.7	98.7
Guided upsamp. [36]	12.4	86.7
FGI [33]	11.2	99.1
SGM [12]	20.7	92.2
MC-CNN [13]	6.34	99.3
Probabilistic fusion [18]	5.91	99.6
Ours	**4.84**	**99.8**

TABLE II

COMPARISON OF COMPUTATION TIME ON THE KITTI BENCHMARK

Method	SGM [18]	MC-CNN [13]	Prob. [18]	Ours
Time (sec)	0.004	0.822	0.024	0.045

kit [14] to project LiDAR point clouds to the color image coordinate.

2) Disparity Estimation: Fig. 7 and Table I show the qualitative and quantitative comparison results on training set of the KITTI stereo 2015 evaluation [14], respectively. Specifically, our method was evaluated compared to the state-of-the-art methods, such as SGM [12] and probabilistic LiDAR-Stereo fusion method [18]. For quantitative evaluation, a bad-pixel error rate was measured by using the KITTI stereo development kit [14].

Even though the probabilistic LiDAR and stereo fusion method [18] has provided reliable performances thanks to its fusion approaches, it also has shown limited performance due to its limited capacity for reliable fusion. Stereo disparity estimation results only from stereo images, such as SGM [12] and MC-CNN [13], also have shown unreliable performance due to inherent limitations of stereo disparity estimation. Unlike these methods, as in the quantitative results, our proposed network achieved the error rate of 4.84%, which is better than the best available competitor's 5.91%. This is more evident in the qualitative evaluation. As shown in Fig. 8, our method have accurately acquired disparity maps on thin or complex objects, which are typical failure examples in stereo disparity estimation. In addition, reliable depth information is also obtained even in areas where the LiDAR does cannot scan range information, such as outside the viewing angle or high reflectance object. These cases demonstrate that LiDAR and stereo depth information complement each other, and they are jointly fused through the benefits of the context information provided by high receptive field of DC.

3) 3D reconstruction: As shown in Fig. 9, to evaluate our method in a practical manner, we reconstructed the 3D model using estimated depth information. Since the accuracy of disparity estimation is reduced with respect to the square of distance, only significant areas of up to $3m$ were visualized. As shown in the 3D reconstruction results, we can argue

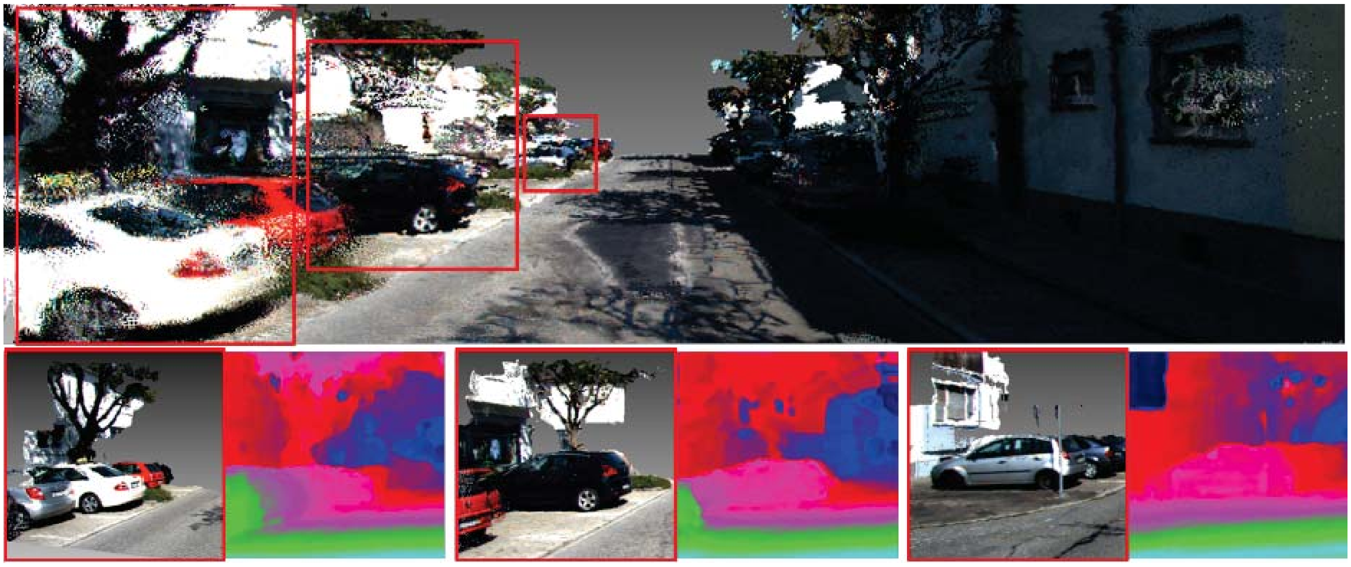
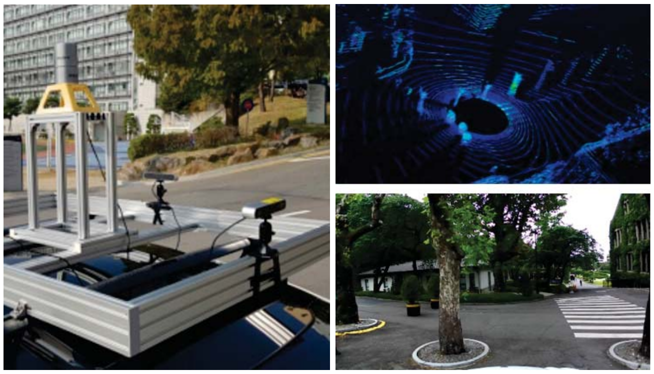


Fig. 9. Examples of outdoor 3D scene reconstruction using proposed method in KITTI dataset [25].



(a) Our system

(b) LiDAR and stereo data

Fig. 10. Illustration of our system. (a) Our system equipped with low-channel LiDAR and stereo camera. (b) Example of LiDAR 3D point cloud and stereo color image acquisition.

that our method can successfully reconstruct 3D map even in challenging outdoor environment.

C. Evaluation on YONSEI Dataset

1) Dataset: We further evaluated our network on our multi-sensor data acquisition system, built for outdoor 3D scene reconstruction. We took various scene data under challenging outdoor environments. As shown in Fig. 10, our recording platform is equipped with a ZED stereo camera of 1280 × 720 resolution, and Velodyne HDL-32E of 32 channels. In particular, the YONSEI dataset contains 32,549 LiDAR–stereo sequential frame sets. This dataset was recorded at 10 Hz. In comparison to the experiments of the KITTI benchmark Section IV-B, our own YONSEI dataset enables us to prove the stability and robustness of the proposed network even with lower channel LiDAR sensor, which is a recent trend in low-cost 3D LiDAR scanners. We just evaluated our network, which was trained on the KITTI benchamrk, on our dataset for an additional evaluation.

2) 3D reconstruction: Despite of the lower number of LiDAR channels in the YONSEI benchmark, our proposed method has acquired accurate depth information and reliable 3D scene reconstruction results even under challenging outdoor conditions, as shown in Fig. 11. This proves that the proposed method is reliable to acquire and reconstruct 3D geometry information in the challenging outdoor environment.

D. Computation Complexity

Table II evaluated the computational complexity of our methods compared to the state-of-the-art algorithms in handling stereo images of size 1242 × 375 and 64 channel LiDAR data. As shown in the results, our algorithm is highly efficient than other algorithms, such as MC-CNN [13], which proves the main contribution of this paper in that accurate depth information can be obtained through an efficient computation of our network.

V. Conclusion

We presented the CNN architecture for high-precision depth estimation. Our network started from the fusion process to encode the complementary characteristics of sparse 3D LiDAR and dense stereo depth in a boosting manner. Our network has also employed the compact convolution module that can be applied to various real time systems. To the best of our knowledge, this network is the first CNN model specifically designed for LiDAR and stereo depth fusion. We also built large-scale datasets using the KITTI raw LiDAR data, and augmented the raw LiDAR scans by adapting off-the-shelf stereo algorithm and confidence measure. We also collected data by our own multi-sensor acquisition system, and demonstrated that our network outperforms the state-of-the-art algorithms. In future work, our network can benefit from the incorporation of additional stereo matching network by improving the input depth information.

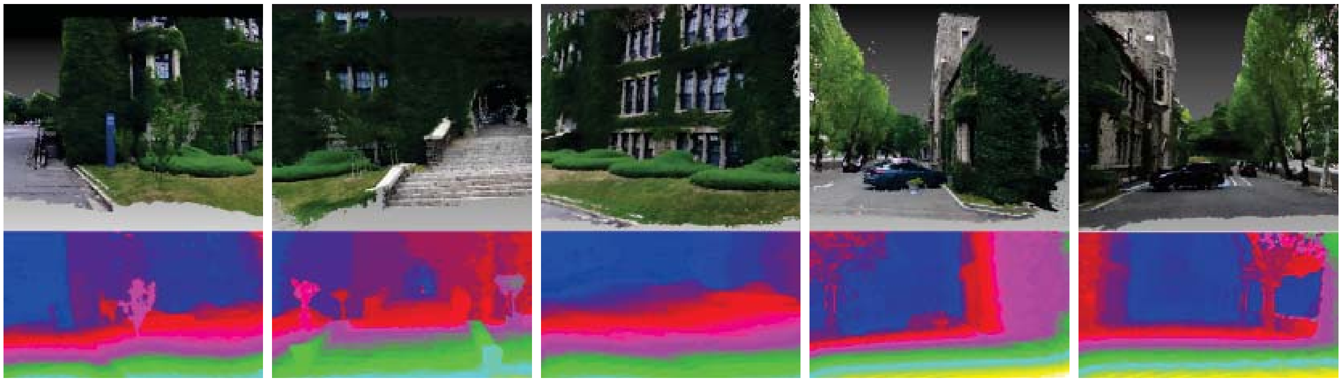


Fig. 11. Examples of outdoor 3D scene reconstruction results using the proposed method in our database.

References

[1] D. Glasner, S. Bagon, and M. Irani, Super-Resolution from a Single Image, In Proc. of the IEEE International Conf. on Computer Vision (ICCV), 2009.

[2] M. Hornacek, C. Rhemann, M. Gelautz, and C. Rother, Depth super resolution by rigid body self-similarity in 3d, In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2013.

[3] G. Riegler, M. Ruther, and H. Bischof, ATGV-net: Accurate depth super-resolution, In Proc. of the European Conf. on Computer Vision (ECCV), 2016.

[4] C. Dong, C. C. Loy, K. He, and X. Tang, Learning a deep convolutional network for image super-resolution, In Proc. of the European Conf. on Computer Vision (ECCV), 2014.

[5] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, A Noise-Aware Filter for Real-Time Depth Upsampling. In ECCV Workshops, 2008.

[6] J. Dolson, J. Baek, C. Plagemann, and S. Thrun, Upsampling Range Data in Dynamic Environments, In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2010.

[7] N. Schneider, L. Schneider, P. Pinggera, U. Franke, M. Pollefeys, and C. Stiller, Semantically guided depth upsampling, In Proc. of the German Conference on Pattern Recognition (GCPR), 2016.

[8] T. Hui, C. C. Loy, and X. Tang, Depth map super-resolution by deep multi-scale guidance, In Proc. of the European Conf. on Computer Vision (ECCV), 2016.

[9] A. Geiger, M. Roser, and R. Urtasun, Efficient large-scale stereo matching, In Proc. of the Asian Conf. on Computer Vision (ACCV), 2010.

[10] T. Kanade and M. Okutomi, A stereo matching algorithm with an adaptive window: Theory and experiment, In Proc. IEEE Int. Conf. on Robotics and Automation (ICRA), 1994.

[11] P. Heise, S. Klose, B. Jensen, and A. Knoll, PM-Huber: patchmatch with huber regularization for stereo matching, In Proc. IEEE International Conf. on Computer Vision (ICCV), December 2013.

[12] H. Hirschmueller, Stereo processing by semiglobal matching and mutual information, IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 30(2):328341, 2008.

[13] J. Zbontar and Y. LeCun, Stereo matching by training a convolutional neural network to compare image patches, Journal of Machine Learning Research, 17(132):2, 2016.

[14] M. Menze and A. Geiger, Object scene flow for autonomous vehicles, In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015.

[15] K. Nickels, A. Castano, and C. Cianci, Fusion of lidar and stereo range for mobile robots, In Proc. Int. Conf. on Advanced Robotics, 2003.

[16] H. Badino, D. Huber, and T. Kanade, Integrating LIDAR into stereo for fast and improved disparity computation, In IEEE Proc. 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011.

[17] V. J. C ech, and R. Horaud, High-resolution depth maps based on TOF-stereo fusion, In Proc. IEEE Conf. on Robotics and Automation (ICRA), 2012.

[18] W. Maddern and P. Newman, Real-time probabilistic fusion of sparse 3D LIDAR and dense stereo, In Proc. IEEE Conf. on Intelligent Robots and Systems (IROS), 2016.

[19] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and Ra. Urtasun, Monocular 3D Object Detection for Autonomous Driving, In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016.

[20] J. Engel, T. Schps, and D. Cremers, LSD-SLAM: Large-Scale Direct Monocular SLAM, In Proc. of the European Conf. on Computer Vision (ECCV), 2014.

[21] H. Yoo, J. Son, B. Ham, and K. Sohn, Real-time rear obstacle detection using reliable disparity for driver assistance, Expert Systems With Applications, 56(1):186-196, 2016.

[22] A. Geiger, J. Ziegler, and C. Stiller, StereoScan: Dense 3d Reconstruction in Real-time, In Proc. IEEE Intelligent Vehicles Symposium (IV), 2011.

[23] C. Premebida, J. Carreira, J. Batista, and U. Nunes, Pedestrian Detection Combining RGB and Dense LIDAR Data, In Proc. IEEE Conf. on Intelligent Robots and Systems (IROS), 2014.

[24] M. Sharma, S. Chaudhury, and B. Lall, Kinect-Variety Fusion: A Novel Hybrid Approach for Artifacts-free 3DTV Content Generation, In Proc. IEEE Int. Conf. on Pattern Recognition (ICPR), 2014.

[25] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, Vision meets Robotics: The KITTI Dataset, The International Journal of Robotics Research (IJRR), 32(11):1231-1237, 2012.

[26] J. Long, E. Shelhamer, and T. Darrell, Fully Convolutional Networks for Semantic Segmentation, In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015.

[27] H. Noh, S. Hong, and B. Han, Learning Deconvolution Network for Semantic Segmentation, In Proc. IEEE International Conf. on Computer Vision (ICCV), 2015.

[28] N. Mayer, E. Ilg, P. Husser, P. Fischer, D. Cremers, A. Dosovitskiy, T. Brox, A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation, In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016.

[29] J. Kim, J. Kwon Lee and K. M. Lee,,Accurate Image Super-Resolution Using Very Deep Convolutional Networks, In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2015.

[30] M. Poggi and S. Mattoccia, Learning from scratch a confidence measure, In Proc. of the British Machine Vision Conference (BMVC), 2016.

[31] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, View Synthesis by Appearance Flow, In Proc. of the European Conf. on Computer Vision (ECCV), 2016.

[32] N. Silberman, P. Kohli, D. Hoiem, R. Fergus, Indoor Segmentation and Support Inference from RGBD Images, In Proc. of the European Conf. on Computer Vision (ECCV), 2012.

[33] Y. Li, D. Min, M. N. Do, and J. Lu, Fast Guided Global Interpolation for Depth and Motion, In Proc. of the European Conf. on Computer Vision (ECCV), 2016.

[34] F. Yu, V. Koltun, Multi-Scale Context Aggregation by Dilated Convolutions, arXiv, 2016.

[35] M. Liu, O. Tuzel, and Y. Taguchi, Joint geodesic upsampling of depth images, In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2013.

[36] J. Dolson, J. Baek, C. Plagemann, and S. Thrun, Upsampling Range Data in Dynamic Environments, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2010.

[37] C. Godard, O. M. Aodha, and G. J. Brostow, Unsupervised Monocular Depth Estimation with Left-Right Consistency, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017.

[38] Online: http://www.vlfeat.org/matconvnet/