International Journal of Control, Automation and Systems 18(1) (2020) 206-216
http://dx.doi.org/10.1007/s12555-019-0350-8
ISSN:1598-6446 eISSN:2005-4092
http://www.springer.com/12555

Real-time Depth Estimation Using Recurrent CNN with Sparse Depth Cues for SLAM System

Sang Jun Lee, Heeyoul Choi, and Sung Soo Hwang*

Abstract: Depth map has been utilized for refinement of geometric information in a variety of fields such as 3D reconstruction and pose estimation in SLAM system where ill-posed problems are occurred. Currently, as learning-based approaches are successfully introduced throughout many problems of vision-based fields, several depth estimation algorithms based on CNN are suggested, which only conduct training of spatial information. Since an image sequence or video used for SLAM system tends to have temporal information, this paper proposes a recurrent CNN architecture for SLAM system to estimate depth map by exploring not only spatial but also temporal information by using convolutional GRU cell, which is constructed to remember weights of past convolutional layers. Furthermore, this paper proposes using additional layers that preserve structure of scenes by utilizing sparse depth cues obtained from SLAM system. The sparse depth cues are produced by projecting reconstructed 3D map into each camera frame, and the sparse cues help to predict accurate depth map avoiding ambiguity of depth map generation of untrained structures in latent space. Despite accuracy of depth cues according to monocular SLAM system degrades than stereo SLAM system, the proposed masking approach, which takes the confidence of depth cues with regard to a relative camera pose between current frame and previous frame, retains the performance of the proposed system with the proposed adaptive regularization in loss function. In the training phase, by preprocessing exponential quantization of ground-truth depth map to eliminate the ill-effects of the captured large distances, the depth map prediction of the proposed system improves more than other baseline methods with accomplishment of real-time system. We expect that this proposed system can be used in SLAM system to refine geometric information for more accurate 3D reconstruction and pose estimation, which are essential parts for robust navigation system of robots.

Keywords: Deep learning, depth estimation, geometry recovery, SLAM, vision based navigation.

1. INTRODUCTION

Depth map, which records depth information of 3D scene as an image format, has been used for more sophisticated development of several technologies in 3D reconstruction, and location recognition [1, 2]. This is because depth map is important information for understanding the geometric structure of an object in three-dimensional space. Such depth map can be acquired easily by using a LiDar or depth sensor, but it is not easy to be used in embedded systems because of high unit price. Another way to estimate depth map is to use a stereo camera. However, in order to acquire depth map from a stereo camera, it is necessary to be accurate calibration both binocular cameras, and disparity matching of each pixel unit which is complicated and computationally intensive that leads no real-time system.

Therefore, research on depth estimation of images captured from monocular camera has been actively studied. In early approaches, statistical models that utilize specially designed textures or filters for a scene is used to estimate depth map [3]. However, those approaches are failed in complex scenes because the designed textures or filters have limits to recover and express depth in such scenes, and are vulnerable to various distance scale variations.

While introduction of deep neural network, which is a learning-based approach, is a successful example in many computer vision fields, several algorithms for depth map generation based on deep neural network have been proposed resolving several limitations in statistical model-based approaches. The deep neural network structure for depth estimation generates robust depth maps to scene

Manuscript received May 10, 2019; revised August 5, 2019; accepted August 5, 2019. Recommended by Associate Editor Hyun Myung under the direction of Editor Jessie (Ju H.) Park. This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education [No. 2016R1D1A3B03934808].

Sang Jun Lee, Heeyoul Choi, and Sung Soo Hwang are with the School of Computer Science and Engineering, Handong Global University, 558, Handong-ro, Heunghae-eup, Buk-gu, Pohang-si 37554, Korea (e-mails: eowjd4@naver.com, hchoi@handong.edu, sshwang@handong.edu).
* Corresponding author.

complexity and scale variation without special feature design through learning.

Since the model based on deep neural network for depth estimation receives an image and generates depth map image, the model is mainly composed of the CNN (Convolutional Nerual Network) model suitable for inferring spatial information [4,5]. However, the CNN-based approach cannot handle temporal information such as video or sequential images.

Currently, camera-based Visual SLAM (Simultaneously Localization and Mapping) systems [6, 7] has considered to use depth information. Visual SLAM is a technique utilized for self-driving car, robotics by generating 3D feature maps and calculating localization (i.e. pose estimation) synchronistically on multi-threading using sequentially acquainted images from camera sensor. Depth map can help accurate localization where there are less feature points, or pure rotation problem is presented. Except for depth sensor or LiDar that is expensive, depth information obtained by stereo camera calculating disparity map cannot be used for Visual SLAM since it cannot guarantee real-time computation. Therefore, the research for depth estimation without any heavy operations such as calculating disparity map is necessary for real-time computation. Even though depth values with regard to the feature points in Visual SLAM calculated by projecting reconstructed 3D map are hard to be full depth map because they are too less and sparse, they can be obtained in real-time. Therefore the sparse depth values can be used for depth cues for depth prediction.

In this paper, we propose a novel depth map estimation algorithm for SLAM system based on deep nerual network in real-time. The proposed deep neural network structure is based on recurrent convolutional neural network using Convolutional GRU(Gated Recurrent Unit), which considers sequential information such as consecutive images utilized for SLAM system. By considering spatio-temporal information while CNN considers only spatial information, the structure of scene is more robustly and constantly recovered than using CNN-based method.

Moreover, this paper presents depth map refinement by using sparse depth cues obtained from SLAM system. Using additional convolutional layers for sparse depth cues and their masks controls reliability of sparse depth cues. And the proposed regularization using sparse depth cues with the confidence of depth cues in loss function can learn accurate structure of scenes even untrained structures by exploring latent space according to depth cues. Because the confidence between monocular SLAM system and stereo SLAM system is different, we propose an analytic confidence value of mask based on the relative pose between camera poses of current frame and previous frame in SLAM system.

By taking space increasing discretization that quantizes depth information in exponential map to avoid the problem of large distances, more accurate learning is operated by the proposed system.

In summary, the contributions of this paper are as follows:

1) We propose a recurrent CNN model for depth estimation tailored to SLAM system. The proposed structure learns spatio-temporal information by convolutional GRU cells that can remember past convolutional layers effectively.
2) We propose the way to apply sparse depth cues obtained from SLAM system to the proposed network structure with the confidence of depth cues to control reliability of depth cues.
3) We suggest regularization term in loss function that consists of utilizing depth cues. Using depth cues in regularization improves generation of reliable structures.
4) We utilize an exponential quantization method to discretize depth map for avoiding any ill-effects of large distances in depth information.

This paper is organized as follows: In Section 2, we introduce several deep neural network-based depth estimation methods with prior knowledge that need to understand this paper. And then, we demonstrate the proposed methods in Section 3. We explain the experiments in Section 4 and conclude in Section 5.

2. RELATED WORK & PRIOR KNOWLEDGE

2.1. Related work

Recently, while CNN-based depth estimation has been actively studied, the first approach is that Eigen *et al*. [4] proposed a model based on coarse-fine approach to solve inherent ambiguity about scale of depth. The coarse network captures a coarse global structure based on entire image, and the fine network refines captured structure more locally. Moreover, they suggest a scale-invarient error to help measure depth relations beyond the scale in log space. Even though this method is fast and capture global and local structures, this model has limitation to take low resolution and accuracy.

Fu *et al*. [5] proposed a model seen depth map estimation as an ordinal regression problem. The proposed ordinal regression model and space increasing quantization of depth map are proposed to address ill-effects of large distances. By applying scene understanding modular that utilizes skip connections with cross-channel interactions, this model can preserve structure of scenes well. Even though this method is very accurate by the proposed strategies, it is hard to be real-time process because the model utilizes pre-trained deep feature extractors that have an amount of parameters.

Those CNN-based approaches are conducted to estimate depth map on each of individual images. In contrast

with the CNN-based approaches, a recurrent CNN-based approach is proposed for SLAM system by Kumar *et al*. [8]. This method utilizes Convolutional LSTM(Long-Short Term Memory) [9]. The results of depth estimation based on recurrent CNN are good for SLAM system by exploring spatio-temporal information of depth. However, the method does not utilize any depth cues that can be produced by SLAM system that can improve the estimation of depth map.

2.2. Depth preprocessing

Depth information obtained from a sensor has the depth values which are recorded according to each distance. Each distance is rectified by the range of the used sensor to be captured within 16 bit range. By the way, the larger depth values have the larger errors as discussed in [5] because of the limitation of depth sensor.

Even though depth information is less rich if the depth values become larger, the depth values is quantized by uniform discretization in a depth map as:

$$\tau_i = \alpha + (\beta - \alpha) * \frac{1}{K}, \tag{1}$$

where the interval is $[\alpha, \beta]$, K is the number of sub-intervals with discretization thresholds $\tau_i \in \{\tau_0, ..., \tau_K\}$.

To address this problem, Fu *et al*. [5] proposed SID (Spacing-increasing discretization) which discretizes depth values in log space to down-weight the training losses in regions with regard to the large depth values. The quantization using SID strategy is following equation:

$$\tau_i = e^{\log(\alpha) + \frac{\log(\beta/\alpha * i)}{K}}. \tag{2}$$

Note that we use α and β as the range of intensity of the depth image, and K is 80 as suggested in [5]. The SID strategy helps to be more accurate prediction as shown the experiments in Section 4.

2.3. Convolutional GRU

GRU(Gated Recurrent Unit) cell [10] is proposed for RNN(Recuurent Neural network) model to learn time series data and to overcome gradient vanishing problem that could lose information of past layers. GRU cell is an alternative cell to LSTM(Long-Short Term Memory) [11], which is more efficient with high accuracy of prediction.

While a GRU cell only covers temporal information, Convolutional GRU cell proposed by [12] is designed to preserve spatio-temporal information by exploring spatial domain through convolutional layers on GRU architecture. A convolutional GRU cell is constructed as:

$$z_t = \sigma\left(W_z \star x_t + U_z \star h_{t-1}\right), \tag{3}$$

$$r_t = \sigma\left(W_r \star x_t + U_r \star h_{t-1}\right), \tag{4}$$

$$\widetilde{h_t} = \tanh\left(W \star x_t + U \star (r_t \circ h_{t-1})\right), \tag{5}$$

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \widetilde{h_t}, \tag{6}$$

where z_t is an update gate at time t, r_t is a reset gate at time t, and h_t is the updated hidden state at time t, σ is the sigmoid function. W_z, W_r, W and U_z, U_r, U are 2D-convolutional kernels, x_t is the input at time t, and operation $\star$ indicates convolution and operator $\circ$ is for element-wise multiplication between two tensors. The hidden recurrent representation h_t preserves the spatial topology as $h_t = (h_t(i,j))$ where $h_t(i,j)$ is a feature vector at location (i,j). Using this simple architecture, convolutional GRU can learn the structure of spatial information with temporal information.

3. METHODS

3.1. Proposed network

The proposed network architecture in this paper is shown in Fig. 1. The architecture of the proposed network accepts consecutive N images as inputs and generates corresponding depth images through encoder-decoder

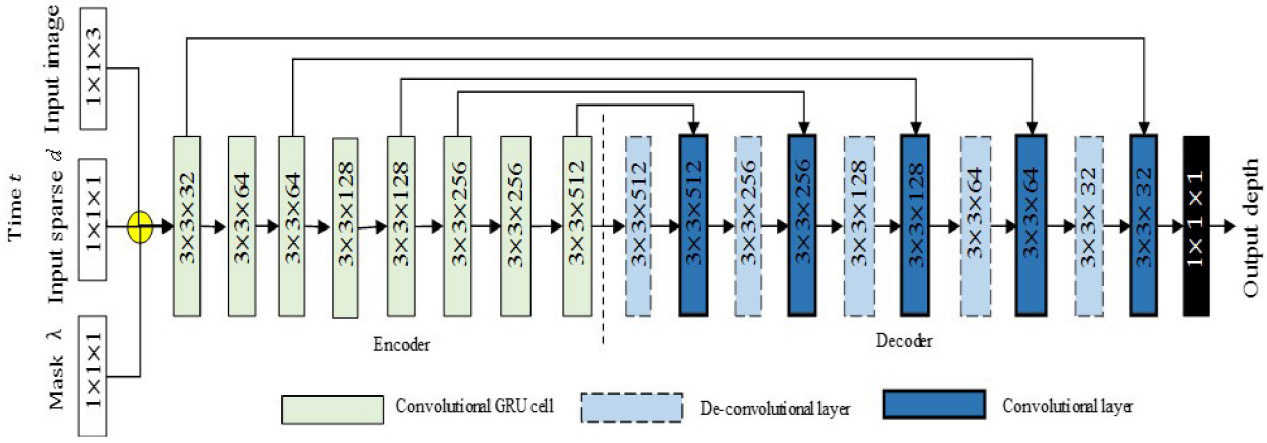


Fig. 1. The network architecture of the proposed system.

architectures. The encoder framework consists of Convolutional GRU cells in each layer, and the decoder framework is composed of de-convolutional layers and convolutional layers. To decode encoded learning parameters to depth image, a de-convolutional layer functions as upsampling process, and a successive convolutional layer is operated to learn latent space with regard to the upsampled image and reduce dimension of tensor size.

The proposed model is in the form of a U-net [13] structure to improve speed and performance similar to [8]. U-net architecture is proposed for image segmentation that accomplishes real-time system, and the proposed skip connections between encoder and decoder frameworks help to understand structure of scenes. The proposed system also applies the skip connections between layers of encoder-decoder structure concatenating correspo-nding convolutional layers. At the end of the model, there is a 1×1 convolutional layer to generate the depth parameters of the decoded parameters. Moreover, all depth images that are used for ground-truth in training phase and sparse depth cues are quantized by SID strategy discussed in Section 2.2. This preprocessing prevents ill-effects of large distances.

Fig. 2 shows unfolded structure of the proposed system since this method is proposed as an recurrent network model. For successive N times, a RGB image and sparse depth cues obtained SLAM system with the confidence λ at time t are inputted to the proposed network. The parameters of the convolutional GRU cells for the previous time are passed to the cells of next time, and the model conducts regression between all predicted depth images and ground-truth depth maps with respect to all times. To this end, the sparse depth cues and the confidence value are used for regularization in regression function. The learning is operated by minimizing the proposed loss function in regression based on back-propagation.

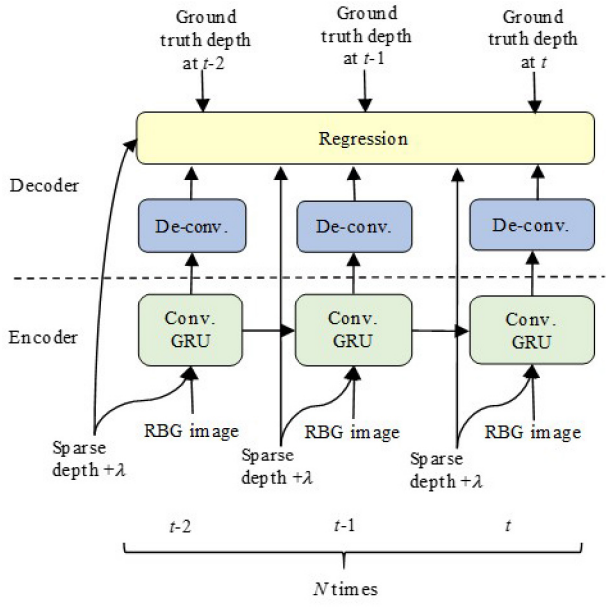


Fig. 2. The unfolded structure of the proposed network model.

In Section 3.2, we show how to obtain sparse depth cues in SLAM system with respect to monocular system and stereo system with their confidence. In Section 3.3, we propose regularization using sparse depth cues with the confidence in the used loss function. In Section 3.4, we explain details for implementation of this system.

3.2. Acquisition of depth cues and the confidence

The Visual SLAM system generates 3D features in 3D map by triangulating matched 2D features between the current frame and previous frame, and estimates the camera pose of current frame using matched 3D-2D feature points by minimizing geometric measurement. Therefore, depths of pixels according to 2D features in current frame can be calculated by matched 3D features as following (7):

$$\text{depth}\,(\mathbf{X},\mathrm{P}) = \frac{\text{sign}\,(\det\,\mathrm{M})\,\omega}{\|\mathrm{m}^3\|}, \tag{7}$$

where $\mathbf{X} = (\mathrm{X},\mathrm{Y},\mathrm{Z},1)^{\mathrm{T}}$ is a reconstructed 3D point, and $\mathrm{P} = [\mathrm{M} \mid \mathrm{p}]_4]$ is a projection matrix that transforms 3D point $\mathbf{X}$ in absolute 3D coordinates to a 2D point $\mathbf{x} = \omega(x,y,1)^{\mathrm{T}}$ in a camera coordinates, i.e., $\mathrm{P}\mathbf{X} = \mathbf{x}$ [14].

At this time, the stereo SLAM system is no need for an initialization step because the stereo system can directly generate 3D feature points using matched 2D features between two binocular cameras. In addition, estimated depth values using (7) in the stereo SLAM system that two binocular cameras are pre-calibrated are same with the depth values obtained from depth sensors. Thus, depth values calculated in the stereo SLAM system have high reliability that can be used for depth cues. Fig. 3 shows acquired depth cues on stereo SLAM system. As shown in the figure, sparse but many depth values of features are illustrated. Those features are reliable so long as they are matched.

In contrast with the stereo SLAM system, the monocular SLAM system, which uses a single camera, generates sparse few and unstable depth values. This is because generation of 3D features is operated between current frame

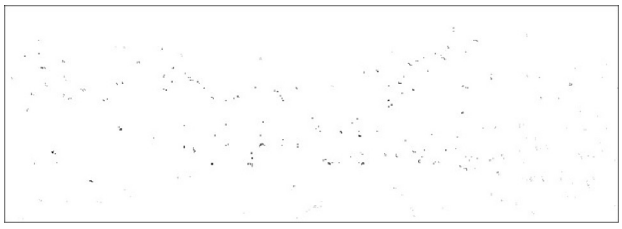


Fig. 3. Sparse depth cues obtained from stereo SLAM system. This figure is inverted for clear visualization.

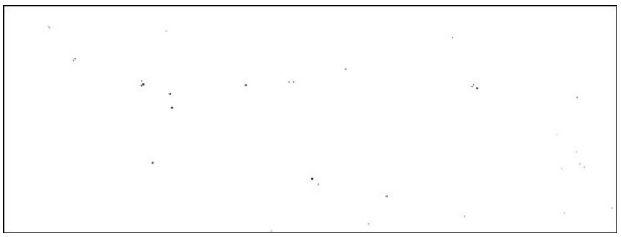


Fig. 4. Sparse depth cues obtained from monocular SLAM system.

and consecutive previous frames, those 3D features are filtered through consistency measurement between 3D-2D data association [15]. In addition, accurate generation of 3D features is affected by relative camera movement known as scale drift [14]. The larger rotation than translation, the more errors are incremented owing to that matched features may be generated as ideal points. Moreover, the 3D map reconstructed by monocular SLAM system has no absolute scale since the initial map is generated up to scale. Therefore, monocular SLAM system generates different scale of map as often as the system is run. As such reasons, sparse few and unstable depth values of features are estimated as drawn in Fig. 4.

To deal with those problems, scale drift and scale ambiguity, the proposed system concatenates a RGB image with sparse depth cues as a 1×1 convolutional layer and another 1×1 convolutional layer filled with the confidence λ where the depth cues are non-zero as a mask. The 1×1 convolutional layer for sparse depth cues prevent scale ambiguity by learning the relationship suitable scale for ground-truth depth map. And another 1×1 convolutional layer used as a mask filters non-zero values in the depth cues, controls whether the sparse depth cues are reliable by the analytically calculated confidence, and updates the confidence to be more suitable by learning.

The confidence λ can be calculated as follows. To compensate the confidence if translation is relatively larger than rotation in monocular SLAM system, λ can be obtained as:

$$\lambda = \min\left(1, \frac{\exp(\triangle \mathrm{t})}{\triangle \theta}\right), \tag{8}$$

where $\exp(\cdot)$ is the function that takes an exponential value, $\triangle \mathrm{t} \in \mathrm{R}^3$ and $\triangle \theta \in \mathrm{R}^3$ are respectively differential translation and rotation between two frames at time t and $t-1$. The reason why taking $\min(\cdot)$ is to avoid exceeding the maximum weight of 1. Given camera poses $[\mathrm{R} \mid \mathrm{t}]_t$, $[\mathrm{R} \mid \mathrm{t}]_{t-1} in \mathrm{R}^{3\times4}$ (R is a 3×3 rotation matrix, t is a 3×1 translation vector), $\triangle$ t and $\triangle \theta$ are calculated as:

$$\triangle \mathrm{t} = \|\mathrm{t}_t - \mathrm{t}_{t-1}\|_2^2, \tag{9}$$

$$\triangle \theta = \frac{\mathrm{v}_t \cdot \mathrm{v}_{t-1}}{\|\mathrm{v}_t\| \, \|\mathrm{v}_{t-1}\|}, \tag{10}$$

where v is the rotation axis of a rotation matrix represented by exponential map [16], and can be obtained by

$$\mathrm{v} = \frac{1}{2\sin\varphi}\begin{bmatrix} r_{32} - r_{23} \\ r_{13} - r_{31} \\ r_{21} - r_{12} \end{bmatrix}, \tag{11}$$

where

$$\varphi = \arccos\left(\frac{\mathrm{trace}(\mathrm{R}) - 1}{2}\right), \tag{12}$$

$$\mathrm{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \in SO(3). \tag{13}$$

In the stereo SLAM system, since obtained depth values are accurate, the confidence α is set to one for the stereo system.

3.3. Loss function with proposed regularization

To learn the proposed deep neural network model, proposed loss function is needed for regression minimizing the error. Scale-Invarient Error proposed by Eigen *et al.* [4] solves the regression problem in log space to reduce the effects of projected distance as:

$$\mathrm{Loss}(\mathrm{y}, \mathrm{y}^*) = \frac{1}{n}\sum_i l_i^2 - \frac{0.5}{n^2}\left(\sum_i l_i\right)^2, \tag{14}$$

where $l_i = \log y_i - \log y_i^*$ is the difference between predicted depth values and ground-truth depth values in log space, and n is the number of valid depth values.

The proposed system utilizes additional regularization to enforce the restoration of the structure according to the sparse depth cues in the depth prediction. Thus, the proposed loss function including the regularization is:

$$\mathrm{L}(y, y^*, d) = \mathrm{Loss}(y, y^*) + \frac{1}{m}\sum_i \|\lambda_{d_i>0} \cdot (y_i - d_i)\|_2^2, \tag{15}$$

where d_i is the sparse depth cues, and m is the number of depth cues if $d_i > 0$. The learning is performed by back-propagation minimizing the proposed loss function $\mathrm{L}(y, y^*, d)$.

3.4. Implementation details

This model is implemented by Pytorch with GTX 1080 GPU. We build this system using [64, 128, 128,256,256 512,512] layers for encoder and [512, 256, 256, 128, 128, 64, 64, 32] layers for decoder. The image size is set by proposed in [4]. For the learning phase, we utilize early stopping to prevent over-fitting and gradient clipping set to 0.5 to avoid gradient explosion in training. We set epochs to 20 and learning rate to 0.0001, the training is done within 20 epochs by early stopping. All layers in decoder

utilize batch normalization with Relu operation. The last layer applies sigmoid operation to predict depth map for regression. All strides in this model are one and filter sizes are 3×3.

We utilize ORB SLAM 2 [6] as the SLAM system to obtain depth cues for both of monocular system and stereo system. All confidences are calculated through relative camera poses between consecutive frames. If a frame has no consecutive frames i.e., first frame or last frame, it uses own frame as consecutive frames.

4. EXPERIMENT

All experiments are conducted on GTX 1080 GPU. For the experiment, we evaluate the proposed model on the KITTI dataset [17]. The KITTI benchmark dataset provides high quality image sequences that are captured by car-mounted cameras and Velodyne LiDar sensors of outdoor scenes as shown in Fig. 5. It provides 56 scenes and we split 28 scenes for training and validation and 23 scenes for test. We set time steps N to 3 for training the convolutional GRU cells.

To evaluate the proposed system, we use standard metrics proposed by [4]. For the accuracy metric(higher is better), the accuracy is calculated by $\max(y_i/(y_i^*), y_i^*/y_i) = \delta < threshold$. For the error metric (lower is better), variety metrics are used such as Abs Rel (Absolute relative difference), Squa Rel(Squared relative difference), and RMSE (Root mean squared error) suggested in [4]. We only use one predicted depth map by last decoder for all tests.

We first compare among the proposed methods. Note that we call the proposed basic model that utilizes Conv. GRU (Convolutional GRU) cell as the Baseline, and we apply several approaches such as SID and mono (monocular) cues or stereo cues with or without the reg. (regularization) term. Note that we do not apply regularization term for the monocular cues because in monocular system, monocular cues has scale ambiguity and it cause gradient explosion in the regularization term.

In the experiment as shown in Table 1, It can be seen that the performance of methods that the proposed strategies are applied is improved. For example, a method that the SID approach is applied is better than one that does not apply it in terms of accuracy and error metrics due to its reduction of errors, and a method utilizing depth cues outperforms those that do not apply depth cues, by preserving structure of the depth cues.

The performance without the regularization term of methods utilizing stereo cues or mono cues is similar to each other. This is because the first two convolutional layers for depth cues and a mask learn suitable weights for monocular or stereo cues even though monocular cues have scale ambiguity. Learning by the regularization term outperforms methods that learn without the regularization term, by directly exploring the depth cues to predict structures according to the depth cues.

We also compare the proposed methods with other methods proposed for depth prediction. We report the results of other methods from [5] as testing similar environments. The results are shown in the Table 2. In the results, except the DORN (ResNet) [5] model, the pro-

Fig. 5. An RGB image and ground truth of depth in KITTI dataset. The ground truth depth is given as semi-dense format obtained by a Velodyne sensor.

Table 1. Comparison of depth prediction with the proposed methods on KITTI dataset.

Method	Accuracy metric (higher is better)			Error metric (lower is better)		
	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Abs Rel	Squa Rel	RMSE
Baseline(Conv.GRU)	0.833	0.954	0.978	0.114	0.639	4.297
Baseline+SID	0.842	0.954	0.978	0.117	**0.602**	4.280
Baseline+mono cues	0.837	0.954	0.978	0.116	0.633	4.207
Baseline+SID+mono cues	0.843	0.956	0.979	**0.112**	0.692	4.119
Baseline+ stereo cues	0.842	0.954	0.978	0.127	0.820	4.109
Baseline+ stereo cues+reg.	0.844	0.957	0.978	0.120	0.721	4.059
Baseline+SID+stereo cues	0.851	0.959	0.979	0.118	0.754	4.046
Baseline+SID+stereo cues+reg.	**0.860**	**0.964**	**0.981**	**0.112**	0.647	**3.974**

Table 2. Comparison of depth prediction with the other methods on KITTI dataset.

Method	Accuracy metric (higher is better)			Error metric (lower is better)		
	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Abs Rel	Squa Rel	RMSE
Make 3D [18]	0.601	0.820	0.926	0.280	3.012	8.734
Eigen *et al.* [4]	0.692	0.899	0.967	0.190	1.515	7.156
Liu *et al.* [19]	0.647	0.882	0.961	0.217	1.841	6.986
Depthnet [8]	0.828	0.945	0.972	0.127	0.838	4.505
Baseline(Conv.GRU)	0.833	0.954	0.978	0.114	0.639	4.297
Baseline+SID	0.842	0.954	0.978	0.117	**0.602**	4.280
Baseline+SID+mono cues	0.843	0.956	0.979	**0.112**	0.692	4.119
Baseline+SID+stereo cues+reg.	**0.860**	**0.964**	**0.981**	**0.112**	0.647	**3.974**
DORN (ResNet) [5]	**0.932**	**0.984**	**0.994**	**0.072**	**0.307**	**2.727**

posed methods outperforms other methods. DORN model as an CNN-based model utilizes a pre-trained deep feature model called ResNet 101, which is designed by stacking 101 deep convolutional layers to extract features of scenes, so it is very accurate but cannot be accomplished as real-time depth prediction system. It is reported in [20] that ResNet 101 spends almost 0.2 sec on same hardware (GTX 1080 GPU) and extra layers proposed in DORN spend more times.

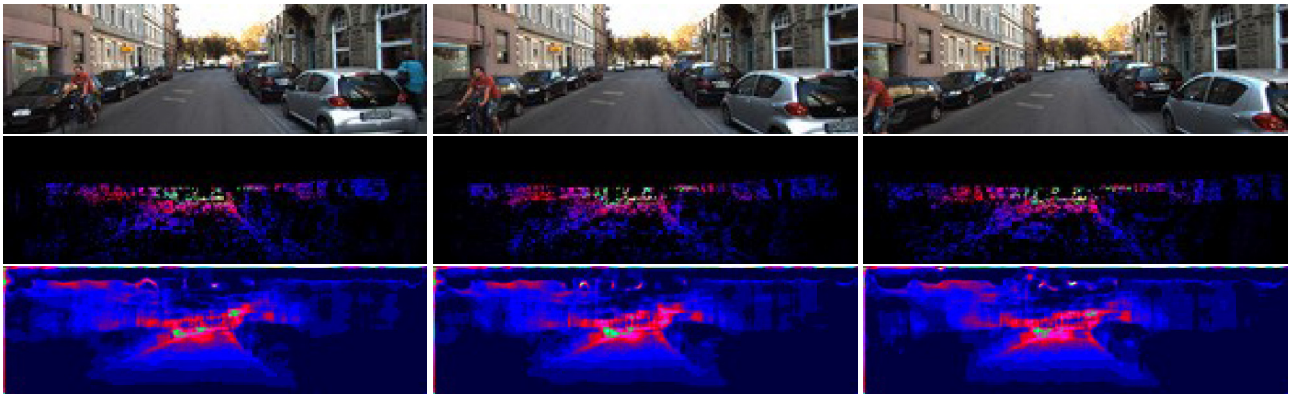
In contrast, the proposed methods do not require the pre-trained deep CNN model for feature extraction because the proposed methods train features of structures through temporal information by a recurrent model. Therefore, the proposed methods performs in real-time system by spending mean 0.08 sec (i.e., 12.5 Hz) with high accuracy of depth map prediction.

The time required to calculate monocular cues, stereo cues, or the confidence λ is negligible. This is because such information is given by SLAM system. The depth cues are directly obtained by projecting observed 3D map points into the current camera frame by using (7), and the camera poses for the confidence are simultaneously calculated in SLAM.

For other methods, Make 3D [18], Eigen *et al.* [4] and Liu *et al.* [19] can accomplish the real-time system as reported in their papers, but accuracy and resolution are lower than the proposed methods. Speed of Depthnet [8] is similar to the proposed system.

As the proposed system is a recurrent model, the system can predict N depth images. It means that it can be seen $12.5 \times N$ Hz by generating N consecutive depth images simultaneously as shown in Fig. 6. Therefore, the proposed model is suitable for SLAM systems that needs to accomplish 30 Hz processing for real-time system. Note that the predicted depth is only accurate in where the ground truth depth is given because the ground truth depth is semi-dense.

Moreover, this system generate very robust depth images in temporal domain. Fig. 7. shows depth prediction on subsequent N images. The proposed system is hard to find differences between generated subsequent depth images.



$t-2$ $t-1$ t

Fig. 6. Simultaneously generated N consecutive depth images by the proposed model. The proposed model can generate N consecutive images through the individual N decoder. The first row indicates RGB images in each time, the second row shows ground truth depth information, and the third row denotes the simultaneously generated depth images according to each time. More bright values indicate more larger distances.

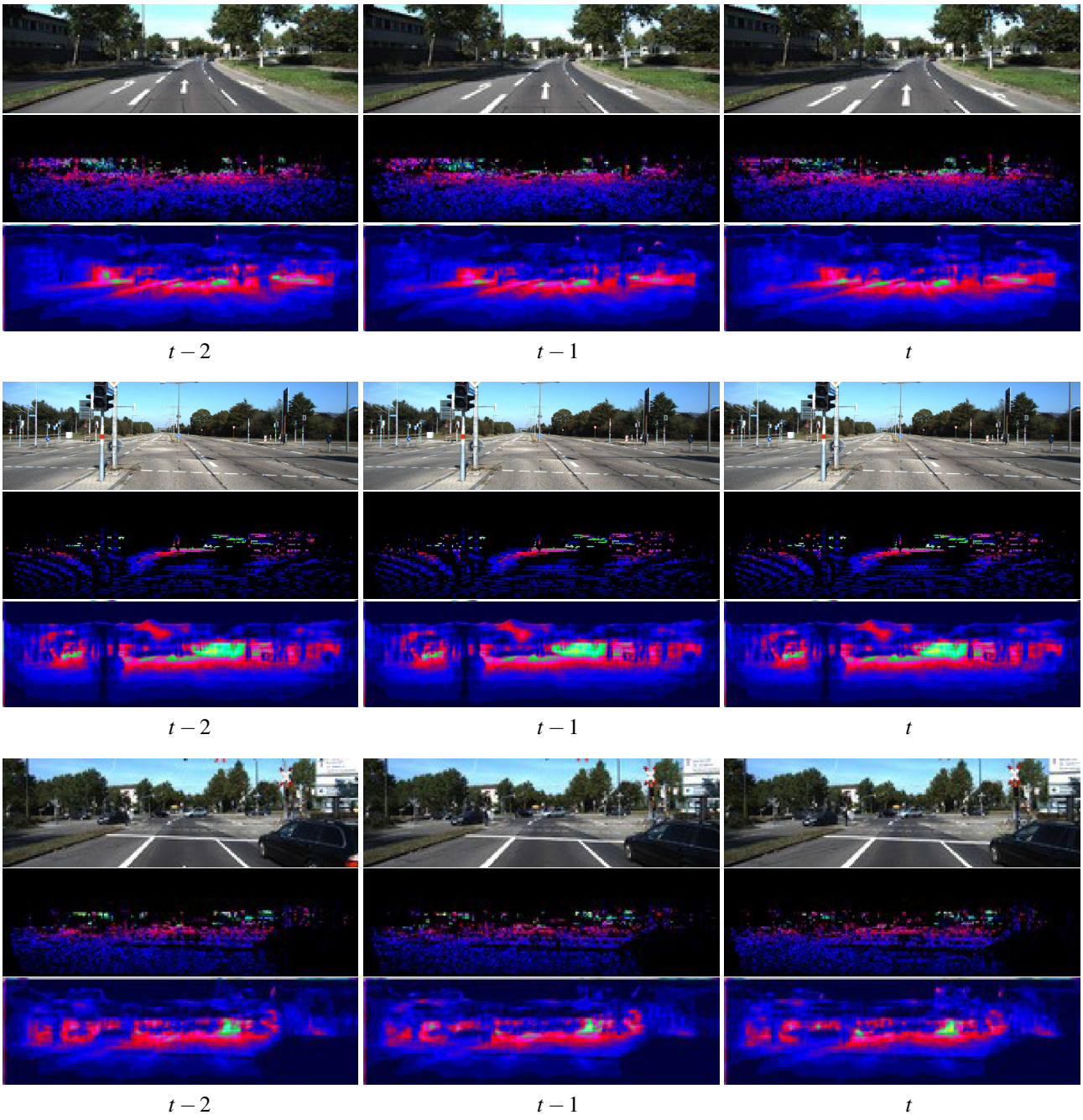


Fig. 7. Generated N subsequent depth images. Each depth map is generated by last decoder. The first row indicates RGB images in each time, the second row shows ground truth depth information in each time, and the third row denotes the generated depth images in each time. More bright values indicate more larger distances.

In addition, Fig. 8 shows examples of depth map prediction among the proposed methods and other methods. We only illustrate DepthNet with the proposed methods Baseline, Baseline+SID+mono cues, and Baseline+SID+stereo cues+reg. This is because Make 3D, Eigen *et al.* or Liu *et al.* is fast but has low resolution with less accuracy excepting DORN as a non real-time system. DepthNet is an approach with similar performance with the proposed systems. We show the improvement of the system than DepthNet and differences among the different proposed strategies in Fig. 8.

In the figure, Baseline is similar to DepthNet but more accurate in the form of scenes. Using either mono cues or stereo cues preserves edges of structures such as cars, buildings, or trees better than Baseline and DepthNet. Moreover, using stereo cues with regularization gener-

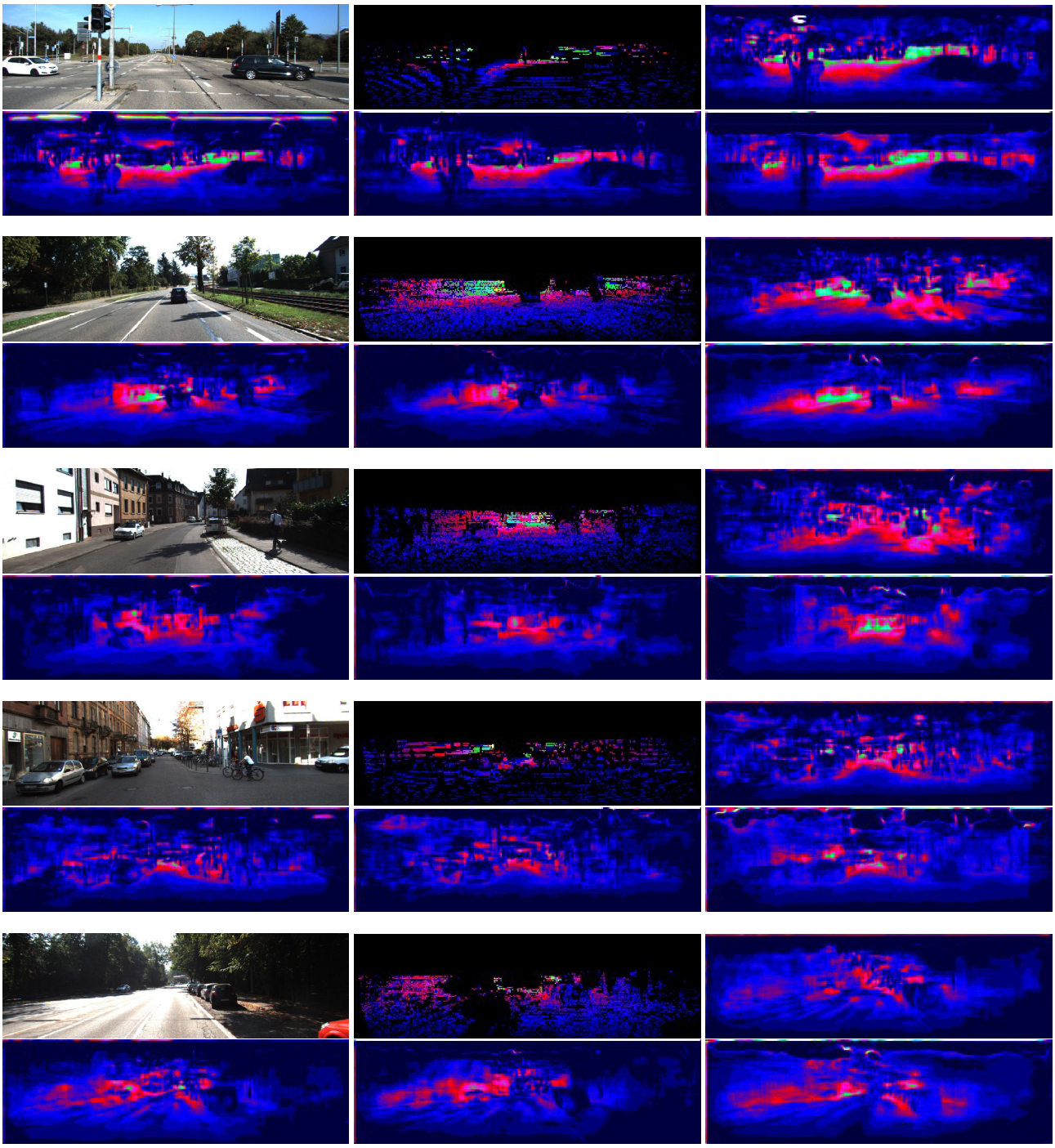


Fig. 8. Examples of depth map prediction among the proposed method with other methods. In each row (6 images), the first image is a RGB image, the second image is given the ground-truth depth, the third image indicates generated depth by DepthNet [8], the fourth image shows generated depth by Baseline, the fifth image is the depth map generated by Baseline+SID+mono cues, and the sixth image shows the depth generated by Baseline+SID+stereo cues+regularization. More bright values indicates more larger distances.

ates more accurate images than using mono cues. This is because stereo cues preserve more accurate edges of structures than mono cues. Therefore, it shows that using accurate depth cues improves the performance of depth prediction.

5. CONCLUSION

In this paper, we present a recurrent CNN model for depth prediction. This model is proposed for SLAM systems that should be real-time system. By utilizing

valuable information such as camera poses and sparse depth cues, the proposed strategies improve the proposed model. The proposed strategies in this paper are as follows. We propose a model that uses convolutional GRU to explore spatio-temporal information, and we utilizes sparse depth cues. To prevent scale problems of the sparse depth cues, additional convolutional layers with a confidence value calculated by given camera poses help to train the model using sparse depth cues. Moreover, using the proposed regularization estimates more accurate and robust structures of scenes according to the depth cues. By using space increasing descritization approach, training has more less errors by eliminating ill-effects of large distances. In the experiments, the proposed methods is more accurate than other methods with real-time system. the proposed methods can learn the features of structures without any deep feature extractor by recurrent CNN architecture, and accomplish high accuracy with real-time speed. This model can be used for SLAM system to operate accurate localization and mapping by using the predicted depth maps which are robust in spatio-temporal domain. For the future work, we will modify this model to unsupervised learning using depth cues with camera poses and apply ordinal regression model to improve the accuracy of prediction.

REFERENCES

[1] G. J. Moon and Q. Zhihua, "An autonomous underwater vehicle as an underwater glider and its depth control," *International Journal of Control, Automation, and Systems*, vol. 13, no. 5, pp.1212-1220, 2015.

[2] N. Metni and T. Hamel, "Visual tracking control of aerial robotic systems with adaptive depth estimation," *International Journal of Control, Automation, and Systems*, vol. 5, no. 1, pp.51-60, 2007.

[3] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1226-1238, 2002.

[4] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems*, pp. 2366-2374, 2014.

[5] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2002-2011, 2018.

[6] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source slam system for monocular, stereo, and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255-1262, 2017.

[7] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: large-scale direct monocular SLAM," *Proc. of European Conference on Computer Vision*, pp. 834-849, September 2014.

[8] A. CS. Kumar, S. M. Bhandarkar, and M. Prasad, "Depthnet: a recurrent neural network architecture for monocular depth prediction," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 283-291, 2018.

[9] S. H. I. Xingjian, Z. Chen, H. Wang, H., D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: a machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, pp. 802-810, 2015.

[10] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *Proc. of Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014.

[11] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," *Advances in Neural Information Processing Systems*, pp. 473-479, 1997.

[12] M. Siam, S. Valipour, M. Jagersand, and N. Ray, "Convolutional gated recurrent networks for video segmentation," *Proc. of IEEE International Conference on Image Processing (ICIP)*, pp. 3090-3094, September, 2017.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Proc. International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234-241, October, 2015.

[14] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2nd ed, New York, NY, USA, 2003.

[15] G. Younes, D. Asmar, E. Shammas, and J. Zelek, "Keyframe-based monocular SLAM: design, survey, and future directions," *Robotics and Autonomous Systems*, vol. 98, pp. 67-88, 2017.

[16] N. K. Ibragimov, *Elementary Lie Group Analysis and Ordinary Differential Equations*, vol. 197, New York, Wiley, 1999.

[17] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354-3361, June 2012.

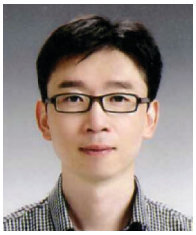
[18] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: learning 3d scene structure from a single still image," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824-840, 2009.

[19] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," *Proc. of Conference on Computer Vision and Pattern Recognition*, 2010.

[20] J. Johnson, Report of CNN Resnet Speed, https://github.com/jcjohnson/cnn-benchmarks, September 2017.

[21] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: self-supervised depth completion from lidar and monocular camera," *Proc. of IEEE International Conference on Robotics and Automation*, 2019.

Sang Jun Lee received his B.S. degree in Computer science and Engineering from Handong Global University, Pohang-si, Korea, in 2017. He is currently pursuing an M.S. degree in the Dept. of Information Technology at the Handong Global University. His research interests include the SLAM system for the localization of self-driving cars, robotics, or users who use augmented reality, virtual reality, as well as 3D reconstruction, and optimization of these technologies using machine learning.



Heeyoul Choi is an assistant professor at Handong Global University. He was a visiting researcher in MILA at University of Montreal from 2015 to 2016. He worked for Samsung Advanced Institute of Technology for 5 years, and was a post-doctoral researcher in Psychological and Brain Science at Indiana University, Indiana, from 2010 to 2011. He received his B.S. and M.S. degrees from Pohang University of Science and Technology, Korea, in 2002 and 2005, respectively, and his Ph.D. degree from Texas A&M University, Texas, in 2010. His research interests cover deep learning and cognitive science.

Sung Soo Hwang received his B.S. degree in Electrical Engineering and Computer Science from Handong Global Unveristy, Pohang, Korea in 2008, and his M.S and Ph.D. degrees in Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 2010 and 2015, respectively. His research interests include image-based 3D modeling, 3D data compression, augmented reality, and Simultaneous Localization and Mapping system.