논문 22-47-05-11 The Journal of Korean Institute of Communications and Information Sciences '22-05 Vol.47 No.05
https://doi.org/10.7840/kics.2022.47.5.781

자율주행차량을 위한 CycleGAN 기반 Depth Completion 기법

응 웬민찌◆, 유명식°

CycleGAN-Based Depth Completion for Autonomous Vehicles

Minh-Tri Nguyen◆, Myungsik Yoo°

요 약

Depth Completion은 자율 주행 차량에서 장면 이해 및 환경 인식 등을 지원하는 중요한 기술이다. 기존 방식은 두 센서의 상호 보완적 특성을 활용하기 위해 RGB 이미지와 Depth LIDAR 이미지와 같은 다중 모달 입력을 고려했다. 그러나 기존의 자동 인코더 접근 방식은 저차원 공간에서 데이터를 표현하는 데 한계가 있다. 또한 RGB 영상의 광 민감도로 인해 카메라 영상과 LIDAR 영상을 융합할 때 Depth의 불연속성이 발생한다. 본 연구에서는 고밀도의 Depth 재구성을 위해 픽셀 밀도 대신 데이터 분포를 학습하는 데 중점을 둔 CycleGAN을 적용하였다. 또한 Depth 불연속성 문제를 완화하기 위하여 Semantic Segmentation을 추가 입력으로 고려하였다. 제안된 방식은 다양한 도로 환경에서 획득된 동기화된 KITTI 벤치마크 데이터를 활용하여 훈련하고 평가하였다. 실험 결과를 통해 제안된 방식이 Depth Completion의 성능과 효율 측면에서 우수함을 입증하였다.

Key Words : depth completion, cycleGAN, semantic segmentation, autonomous vehicle, sensor fusion

ABSTRACT

Depth completion is a challenging task supporting the purpose of scene understanding and environment perception in an autonomous vehicle. The existing method considered multiple modals input such as RGB images and depth LIDAR images to utilize the complementary characteristics of those two sensors. However, traditional autoencoder approaches have shown limitations in representing the data in low dimensional space. Moreover, depth discontinuity also happened when fusing the camera image and LIDAR image due to the light sensitivity in the RGB image. In our study, we are adapting CycleGAN focusing on learning the distribution of the data rather than the pixel density to reconstruct the depth into dense one. We also consider the semantic segmentation as additional input to mitigate the depth discontinuity problem. Our framework is trained and evaluated on the KITTI benchmark with synchronized data capturing various road scenery. The experimental results prove the proposed framework to be competitive performance and efficient in depth completion task..

※ This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (2021R1A2B5B01002559).

◆ First Author : Department of Information Communication Convergence, Soongsil University, Seoul, South Korea, tri.nguyenminh.2111@gmail.com, 학생회원

° Corresponding Author : School of Electronic Engineering, Soongsil University, Seoul, South Korea, myoo@ssu.ac.kr, 종신회원

논문번호 : 202202-025-C-RN, Received February 25, 2022; Revised March 17, 2022; Accepted March 18, 2022

Ⅰ. Introduction

Nowadays, computer vision and deep learning have proved to be an essential part of the development of the autonomous vehicle. As the crucial key for the ability to operate in the human world, the perception system takes the responsibility of collecting information from the surrounding environment. In particular, the perception system performs various scene understanding tasks such as semantic segmentation, depth completion, object detection, etc. In order to make a precise decision, multiple types of sensors collecting different aspects of data is mounted on the vehicle and by a fusion procedure, the outcome data can be more reliable and well visualized.

In 2001, LIDAR (Light Detection and Ranging), which is a range measurement sensor, was first introduced to the autonomous vehicle and received a lot of attentions from researchers in the world. Having a similar working principle with RADAR, LIDAR can identify the distance to an object by calculating the time of flight of a signal when traveling from LIDAR to an object and reflecting the sensor. However, instead of using radio waves, LIDAR not only can shoot out multiple rays of laser beams with high frequency but also can rotate 360 degrees to scan the surrounding environment. Moreover, due to the characteristic of laser, LIDAR achieves high precision within a great distance compared to the RADAR. However, data collected from LIDAR or point cloud is often sparse and irregular structure. In particular, when a laser beam hits an object and reflects, it sometimes cannot return to the sensor because of the object's surface, environmental noises such as heavy light illuminance or some extremely adverse weather conditions, etc. As a result, point cloud appears some holes identified as invalid measurement regions. Depth completion is to fill up those holes in sparse depth LIDAR data to generate a dense depth map.

The existing methods such as FusionNet[1], PENet[2] or GuideNet[3] utilize corresponding RGB images as additional input to make the depth prediction and preserve scene structure in depth image. However, since the RGB is highly sensitive to optical changes, the result can be falsely predicted and downgrade the performance. In this paper, we will put semantic segmentation as a third input to compensate for the optical changes in RGB images.

Generative Adversarial Network (GAN), introduced in 2014, had made a significant contribution to image generation. The most outstanding breakthrough of GAN is that it efficiently learns the distribution of data rather than their numerical density value. Therefore, GAN can better represent the data in a low-dimensional manifold than traditional autoencoder architecture. In addition, structure preservation is also an essential key in depth completion. This problem is considered to be similar to the image-to-image translation problem, which is the task of transferring the source domain to the target domain while ensuring the content is consistent. Zhu et al.[4] proposed a cycle-consistent adversarial (CycleGAN) implementing the image-to-image translation concept. In particular, by taking advantage of feature richness in RGB image, the CycleGAN can transfer the color domain of RGB image to the depth domain of the depth image and the sparse depth map can be constructed into a dense and well-structured depth map.

In summary, our contributions are as follow:

- We successfully adapted the CycleGAN architecture for solving the depth completion problem.
- We designed a generator having semantic segmentation as an additional input and introduced a smoothness penalization loss for enhancing the model performance.

Ⅱ. Related Works

IImage-to-Image translation. The purpose of image-to-image translation is to transform the content of an image from the source domain to a target domain. The transformation must be consistent in image content and style between two domains. Generative Adversarial Network (GAN)[5] is achieved multiple outstanding result in the task of generating a realistic image using the adversarial idea. CycleGAN[4] is proposed by Zhu et al. combining original GAN and cycle consistency loss for unpaired image translation task. However, CycleGAN not only has low performance when converting images with complex structures but also poorly generates semantic labels and object details from RGB images. Fu et al.[6] proposed a geometry-consistent generative adversarial network (GcGAN) taking a single RGB as the first input and its corresponding geometry transformation as the second input, and GcGAN can generate the output with the corresponding geometry constraint. Nevertheless, despite achieving some noticeable improvement, GcGAN is still unable to compensate for the problem of generating poor details in CyccleGAN when transforming. In an effort to solve the aforementioned problem, Li et al.[7]. proposed an Asymmetric GAN (AsymGAN) successfully generating the result with fine-grained detail and further improving image realistic. They designed the second autoencoder following with the second discriminator for learning auxiliary of image content, but the complexity of the model is significantly increased.

Depth completion. Depth completion is the task of generating valid depth measurements and replacing unidentified regions in sparse LIDAR images. Two main approaches are single depth image input and multiple modals input. Uhrig et al.[8] proposed a sparse convolution mechanism working on sparse data and used a validity mask to indicate the position of valid depth value. However, the performance of the model gradually decreases in a deeper layer. Lu et al.[9] further improved by focusing on learning and constructing the semantic segmentation information in the depth image. HMSNet[10] proposed by Eldesokey et al. also has some improvement. They designed a multi-functional layer for upsampling and downsampling the sparse depth. In contrast to the single depth input, multiple modals input taking RGB image as additional input outperformed the single depth input method due to the complemented characteristics between RGB image and sparse depth LIDAR image. Mat et al.[11] designed a deep regression model taking a concatenated sparse depth map with the corresponding RGB image to predict the depth value. Eldesokey et al.[12] proposed a normalized convolution layer in CNN for adapting the sparse data and extracted a confidence map indicating the validity of pixels in the sparse data. This approach also came with a small number of parameters, which significantly reduced the size of the model. Besides RGB images, other aspects also is being considered as supportive information to enhance the accuracy of prediction, such as surface normal[13] and semantic segmentation[14].

In our study, we will consider the semantic segmentation information as an additional input to compensate for the light sensitivity of the RGB image problem, which makes the depth discontinuity in the predicted result.

Ⅲ. Methodology

3.1 Architecture overview

The overall framework is shown in Figure 1. Our backbone is based on the CycleGAN[4] architecture consisting of two design generators, and two similar design discriminators. As abovementioned, we embed the semantic segmentation as additional input by concatenating with the sparse depth LIDAR. The proposed architecture has two stages:

- In the first stage, set of three input 2D LIDAR depth image, RGB image and semantic segmentation are fed into the generator G_d to generate the dense depth LIDAR image. Then, both of the generated result and the corresponding

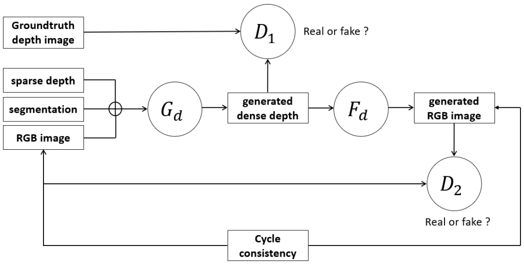

Fig. 1. Overview of overall architecture.

dense depth ground truth will be put into the discriminator D_1 for classification.

- In the second stage, the generated dense depth result from the generator G_d will put into the generator F_d to transfer back to the color domain or the RGB image. Then, the generated RGB image and the original corresponding RGB image will also be put in the second discriminator D_2 for classification purpose.

The detail of the generator G_d/F_d and the discriminator D_1/D_2 will be discussed in the next sector.

3.2 Architecture details

Generator. The architecture of both G_d and F_d is similar. We design the generator having the autoencoder architecture similar with the PENet[2], which has two branches with similar layer layout. The upper aims to predict the dense depth map from RGB image with a confidence map. The lower also aims to upsampling the sparse LIDAR sample with a confidence map. Finally, final depth map is combined from two branches by a fusion metod applied in FusionNet[1]. In particular, we denoted the depth map and confidence map generated from RGB image as d_{rgb} and c_{rgb}; the depth map and the confidence map upsampled from sparse LIDAR as d_d and c_d. The final depth D is:

$$D = \frac{e^{c_{rgb}(x,y)} * d_{rgb}(x,y) + e^{c_d(x,y)} * d_d(x,y)}{e^{c_{rgb}(x,y)} + e^{c_d(x,y)}}$$

The detail of the generator is shown as Figure 3. However, we also make some modifications to the decoder. We replace the traditional deconvolutional layer causing heavy checkerboard artifact as shown in Figure 2 with the resize convolution layer[15]. It consists of a nearest-neighbor upsampling layer followed by a convolution layer.

Discriminator. The architecture of both D_1 and D_2 is similar. We design discriminator based on the DCGAN[16] architecture. The backbone of the discriminator is constructed by applied series of convolutional block. Each block consists of three consecutive layers: a convolution layer, a BatchNorm layer and a ReLU activation layer. Finally, a sigmoid layer is placed at the end to output a probability indicating which one is real of fake sample. The detail of the discriminator is

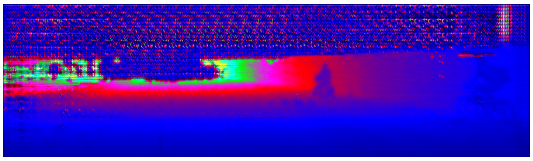

Fig. 2. Heavy checkerboard artifact.

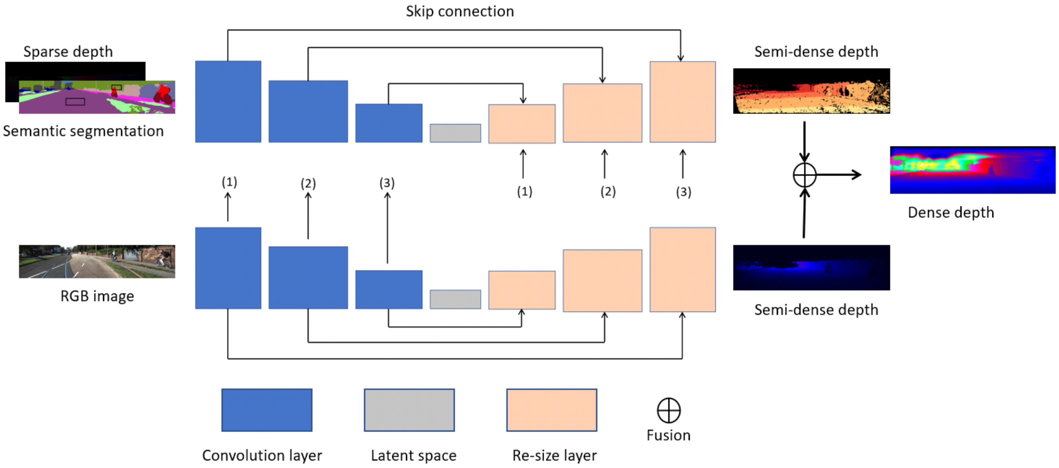

Fig. 3. Overview of the generator architecture

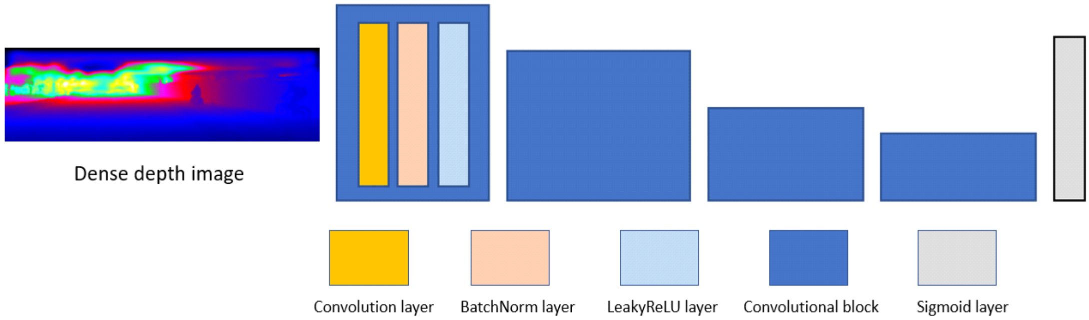

Fig. 4. Overview of the discriminator architecture.

shown as Figure 4.

3.3 Loss function

The cycle consistency loss. The cycle consistency[4] is applied to constrain the consistency when transforming between from RGB domain into depth domain and reversing. The loss is formulated as:

$$l_{cycle} = E_{x \sim p_{data(d)}} \| F_d(G_d(d) - d) \|_1 + E_{x \sim p_{data(rgb)}} \| F_d(G_d(rgb) - rgb) \|_1,$$

where d denoted as predicted depth map, rgb denoted as corresponding rgb image sample.

Structural similarity reconstruction loss. We further constrain the transformation by using the SSIM[17] metric to ensure the scene structure and the original content between input RGB image and the generated RGB image in stage two of the model. The loss is formulated as:

$$l_{ssim} = 1 - SSIM\ (rgb, \widehat{rgb}),$$

where rgb is the input RGB image, and the $\widehat{rgb}$ is the RGB image generated by the generator F_d.

Depth loss. Following the study by Carvhalho et al.[18], we decide to use L_1 to minimize the loss between the generated depth value and the groundtruth. The loss is formulated as:

$$l_{depth} = | 1_{gt>0} \odot (d - gt) |_1,$$

where d denoted as predicted depth map, gt denoted as corresponding groundtruth sample. $1_{gt>0}$ denoted as the indicator showing valid depth value in depth LIDAR image.

Smooth penalization loss. As mentioned before, due to the light sensitivity of RGB, we use the semantic segmentation image to compensate the problem. Inspired by [19], we adapt the second-order differential loss to reduce the impact of the optical changes. The loss is formulated as:

$$l_{smooth} = \frac{1}{N} \sum_{i=1}^{N} (|\nabla_x^2 d_i| * e^{|\nabla_x^2 s_i|} + |\nabla_y^2 d_i| * e^{|\nabla_y^2 s_i|}),$$

where $\nabla_x^2 d_i$, $\nabla_y^2 d_i$ denoted as the second-order derivative in the x-direction and y-direction of the generated depth image, $\nabla_x^2 s_i$, $\nabla_y^2 s_i$ denoted as the second-order derivate in the x-direction and y-direction of the semantic segmentation image. e is the exponential function. N is the total number of pixels.

Adversarial loss. In the DCGAN, adversarial loss using BCE loss is used to help the discriminator D_1/ D_2 classifying between the generated sample and the groundtruth sample. The loss is formulated as:

$$l_{adv} = E_{x \sim p_{data(gt)}} [\log D_1(x)] + E_{x \sim p_{data(d)}} [1 - \log D_1(G_d(x))]$$

The above loss function is used for the discriminator D_1 in the first stage with the sample generated by G_d. The adversarial loss for D_2 is similar but with the sample generated by F_d.

The overall loss used in training is formulated as:

$$loss = \lambda_1 l_{cycle} + \lambda_2 l_{ssim} + \lambda_3 l_{depth} \\ + \lambda_4 l_{smooth} + \lambda_5 l_{adv},$$

where λ_1, λ_2, λ_3, λ_4, λ_5 are the hyper parameters controlling the weight of each loss.

Ⅳ. Experiment Results

4.1 Dataset

In our study, KITTI dataset is used for training and evaluating. The dataset contains total of 85898 training samples, 1000 verification samples, 1000 test samples. Each RGB frame is synchronized with project LIDAR point to 2D image plane. During training, we reduce the image dimension from 1216x352 to 512x256 for saving time and memory capacity.

4.2 Evaluation metric

We use 4 standard metric using in depth completion benchmark: root mean square error (RMSE) [mm], mean absolute error (MAE) [mm], root mean square error of the inverse depth (iRMSE) [1/km], and mean absolute error of the inverse depth (iMAE) [1/km].

4.3 Performance evaluation

Table 1. shows the performance comparison between different architecture. Compared to IR_L2[9], NConv-CNN[12] and PwP[21], we have better result and suffer less noise in the upper region in the depth image. CSPN[20] and Sparse-to-Dense[11] show that they fail to fill out the valid depth in the upper area. Moreover, DeepLIDAR[13] shows that they are losing some content and scene structure compared to the RGB image. Overall, our result prove to have better in visual and smoother in depth transition as shown in Figure 5.

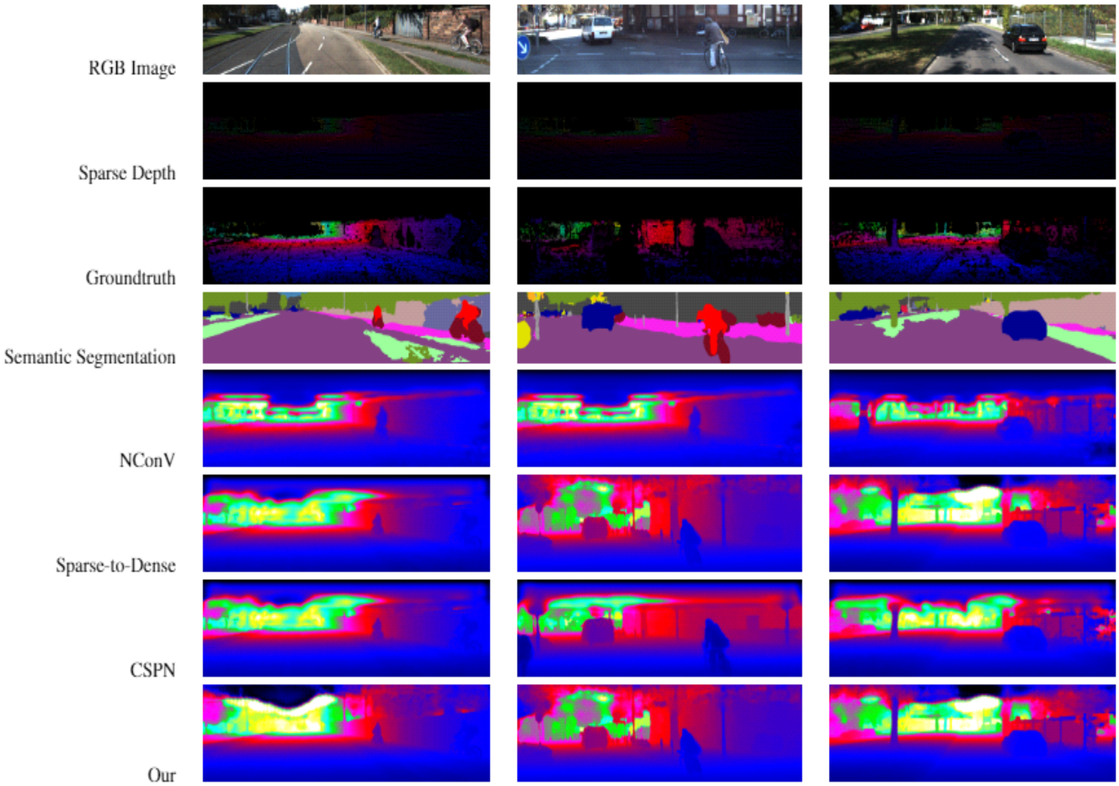

Fig. 5. Visualization of the comparison result

Table 1. Performance comparison of others architecture

Method	Input	RMSE [mm]	MAE [mm]	iRMSE [1/km]	iMAE [1/km]
IR_L2 [9]	depth-only	901.43	292.36	4.92	1.35
CSPN [20]	multiple-input	1019.64	279.46	2.93	1.15
NConv-CNN [12]	multiple-input	829.98	233.26	2.60	1.03
Sparse-to-Dense [11]	multiple-input	814.73	249.95	2.80	1.21
PwP [21]	multiple-input	777.05	235.17	2.42	1.13
DeepLIDAR [13]	multiple-input	758.38	226.50	2.56	1.15
Our	multiple-input	746.96	267.71	2.24	1.10

Ⅴ. Conclusion

We proposed an architecture using a CycleGAN that takes multiple inputs to generate a dense depth map. We further concatenated the semantic segmentation with the sparse depth LIDAR image to compensate for the sensitivity of the RGB image to lighting conditions in the surrounding environment. In our future study, we will combine the task of semantic segmentation with the CycleGAN and build a multi-task architecture.

References

[1] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool, “Sparse and noisy lidar completion with rgb guidance and uncertainty,” in *2019 16th Int. Conf. MVA*, pp. 1-6, Tokyo, Japan, May 2019.

[2] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, “Towards precise and efficient image guided depth completion,” *arXiv e-prints, arXiv-2103*, 2021.

[3] J. Tang, F. P. Tian, W. Feng, J. Li, and P. Tan, “Learning guided convolutional network for depth completion,” *IEEE Trans. Image Process.*, vol. 30, pp. 1116-1129, Dec. 2020.

[4] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2223-2232, Venice, Italy, Oct. 2017.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and Y. Bengio, “Generative adversarial nets,” *Advances in NIPS*, vol. 27, Monreal, Canada, Dec. 2014.

[6] H. Fu, M. Gong, C. Wang, K. Batmanghelich, K. Zhang, and D. Tao, “Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping,” in *Proc. IEEE/CVF Conf. CVPR*, pp. 2427-2436, Califonia, USA, Jun. 2019.

[7] Y. Li, S. Tang, R. Zhang, Y. Zhang, J. Li, and S. Yan, “Asymmetric GAN for unpaired image-to-image translation,” *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5881-5896, Jun. 2019.

[8] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, “Sparsity invariant cnns,” in *2017 Int. Conf. 3DV*, pp. 11-20, Qingdao, China, Jun. 2017.

[9] K. Lu, N. Barnes, S. Anwar, and L. Zheng, “From depth what can you see? Depth completion via auxiliary image reconstruction,” in *Proc. IEEE/CVF Conf. CVPR*, pp. 11306-11315, China, Jun. 2020.

[10] Z. Huang, J. Fan, S. Cheng, S. Yi, X. Wang, and H. Li, “Hms-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion,” *IEEE Trans. Image Process.*, vol. 29, pp. 3429-3441, Dec. 2019.

[11] F. Ma and S. Karaman, “Sparse-to-dense: Depth prediction from sparse depth samples and a single image,” in *2018 IEEE ICRA*, pp. 4796-4803, Brisbane, QLD, Autralia, May 2018.

[12] A. Eldesokey, M. Felsberg, and F. S. Khan,

"Confidence propagation through cnns for guided sparse depth regression," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 42 no. 10, pp. 2423-2436, Jul. 2019.

[13] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys, "Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image," in *Proc. IEEE/CVF Conf. CVPR*, pp. 3313-3322, Califonia, USA, Jun. 2019.

[14] M. Jaritz, R. De Charette, E. Wirbel, X. Perrotton, and F. Nashashibi, "Sparse and dense data with cnns: Depth completion and semantic segmentation," in *2018 Int. Conf. 3DV*, pp. 52-60, Verona, Italy, Sep. 2018.

[15] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, Oct. 2016.

[16] Y. Yu, Z. Gong, P. Zhong, and J. Shan, "Unsupervised representation learning with deep convolutional neural network for remote sensing images," in *Int. Conf. Image and Graphics*, pp. 97-108, vol. 10667, Shanghai, China, Dec. 2017.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600-612. Apr. 2004.

[18] M. Carvalho, B. Le Saux, P. Trouvé-Peloux, A. Almansa, and F. Champagnat, "On regression losses for deep depth estimation," in *2018 25th IEEE ICIP*, pp. 2915-2919, Athens, Greece, Oct. 2018.

[19] C. Zhao, G. G. Yen, Q. Sun, C. Zhang, and Y. Tang, "Masked GAN for unsupervised depth and pose prediction with scale consistency," *IEEE Trans. Neural Netw. and Learn. Syst.*, vol. 32, no. 12, pp. 5392-5403, Dec. 2020.

[20] X. Cheng, P. Wang, and R. Yang, "Learning depth with convolutional spatial propagation network," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 42, no. 10, pp. 2361-2379, Oct. 2019.

[21] Y. Xu, X. Zhu, J. Shi, G. Zhang, H. Bao, and H. Li, "Depth completion from sparse lidar data with depth-normal constraints," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, pp. 2811-2820, Califonia, USA, Jun. 2019.

응 웬민찌 (Minh-Tri Nguyen)

Tri Minh Nguyen received the B.Eng. degree in computer engineering from the University of Information Technology, Vietnam National University−Ho Chi Minh City, Ho Chi Minh City, Vietnam, in 2019. He is currently pursuing the master's degree with Soongsil University. His research interests include Visible Light Communication(VLC).

유 명 식 (Myungsik Yoo)

Myungsik Yoo received his B.S. and M.S. degrees in electrical engineering from Korea University, Seoul, Republic of Korea, in 1989 and 1991, and his Ph.D. in electrical engineering from State University of New York at Buffalo, New York, USA in 2000. He was a senior research engineer at Nokia Research Center, Burlington, Massachusetts. He is currently a professor in the school of electronic engineering, Soongsil University, Seoul, Republic of Korea. His research interests include visible light communications, sensor networks, Internet protocols, control, and management issues.