# Generating High Resolution Depth Image from Low Resolution LiDAR Data using RGB Image

Kento Yamakawa, Fumihiko Sakaue and Jun Sato
*Nagoya Institute of Technology, Japan*

Keywords: Depth Image, RGB Image, High Resolution, GAN.

Abstract: In this paper, we propose a GAN that generates a high-resolution depth image from a low-resolution depth image obtained from low-resolution LiDAR. Our method uses a high-resolution RGB image as a guide image, and generate high-resolution depth image from low-resolution depth image efficiently by using GAN. The results of the qualitative and quantitative evaluation show the effectiveness of the proposed method.

## 1 INTRODUCTION

In recent years, autonomous driving and driving support for vehicles are advancing, and it is becoming more common to equip vehicles with various sensors. Especially in autonomous driving, it is expected that LiDAR will be installed in addition to the RGB camera (Caesar et al., 2020). RGB cameras can acquire high-resolution images at low cost, but they cannot directly obtain depth images. Although many methods (Eigen et al., 2014; Laina et al., 2016; Godard et al., 2017) have been proposed for estimating depth images from RGB images by using deep neural networks, they are still inaccurate and suffer from the domain shift problem.

LiDARs, on the other hand, have the advantage of being able to acquire depth images directly. However, they have low vertical resolution and are extremely expensive. In order to realize autonomous driving, it is important to obtain accurate high-resolution depth images at low cost.

Thus, we in this paper propose a new method for obtaining accurate high-resolution depth images by combining high-resolution RGB images with low-resolution depth images. In our method, we consider image super resolution as an image inpainting problem for defect images, and use adversarial learning (GAN (Goodfellow et al., 2014)) for obtaining high resolution complemented images from low resolution defect images of depth. We test two different types of generators, and evaluate their performance. The proposed GAN can generate high resolution depth images as shown in Fig. 1 (c) from RGB images and low-resolution depth images as shown in Fig. 1 (a) and (b).

(a) RGB image (input) (b) LR depth image (input) (c) HR depth image (our result)

Figure 1: High resolution depth image generated from low-resolution depth image of LiDAR by using our proposed method. The low-resolution depth image is considered as defect image with holes and image inpainting is conducted for obtaining high resolution depth image with the proposed method.

## 2 RELATED WORK

Many methods have been proposed for estimating depth images from RGB images. While traditional methods use parallax of stereo images (Vogiatzis et al., 2005) (Hirschmuller, 2007), modern methods can estimate a depth image from a single RGB image by using a deep neural network (Eigen et al., 2014; Laina et al., 2016; Godard et al., 2017). However, these methods are not yet accurate enough and also suffer from the domain shift problem. LiDARs, on the other hand, have the advantage of being able to acquire depth images directly. However, they have low vertical resolution and are extremely expensive compered to RGB cameras.



Yamakawa, K., Sakaue, F. and Sato, J.
Generating High Resolution Depth Image from Low Resolution LiDAR Data using RGB Image.
DOI: 10.5220/0010903900003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 659-665
ISBN: 978-989-758-555-5; ISSN: 2184-4321

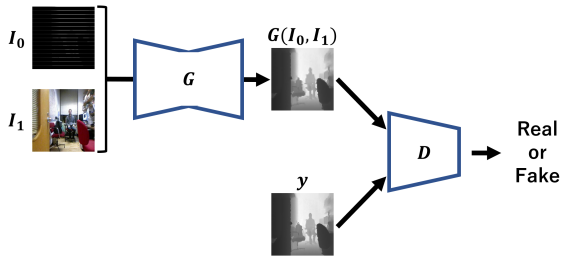



Figure 2: Network structure.

Image super-resolution, which generates a high-resolution image from a low-resolution image, has recently been improved in accuracy by using deep learning (Ledig et al., 2017). The standard technique of super-resolution network is to up-sampling the low-resolution images to obtain high-resolution images. On the other hand, in this research, we consider the image super-resolution as the image inpainting of sparse high-resolution images, and construct a deep neural network for image inpainting of sparse images. In particular, we use a high-resolution RGB image as a guide image, and conduct image inpaiting for sparse high-resolution depth image obtained by enlarging the original low-resolution depth image to the same size as the high-resolution RGB image.

In this paper, the image inpainting is realized by using the Generative Adversarial Network (GAN) (Goodfellow et al., 2014). In GAN, it is known that visually natural images can be generated by training generator and discriminator adversarially. In this paper, the GAN learns image inpainting to generate a high-resolution depth image from a sparse depth image using an RGB image as a guide.

# 3 GENERATING HIGH RESOLUTION DEPTH IMAGE FROM GAN

In this research, we propose a network that generates high-resolution depth images by image inpainting technique that reconstructs the missing part of the depth image. A low-resolution depth image is considered as a high-resolution image lacking information, and the task of inpainting the missing part is learned using GAN to generate a high-resolution depth image. By inputting the high-resolution RGB image as a guide image to GAN, the high-frequency components lacking in the low-resolution depth image are complemented by the high-resolution RGB image, and more accurate high-resolution depth image is generated.

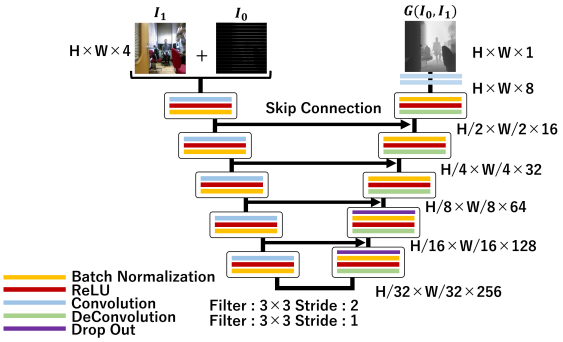



Figure 3: Generator of Proposed method 1.

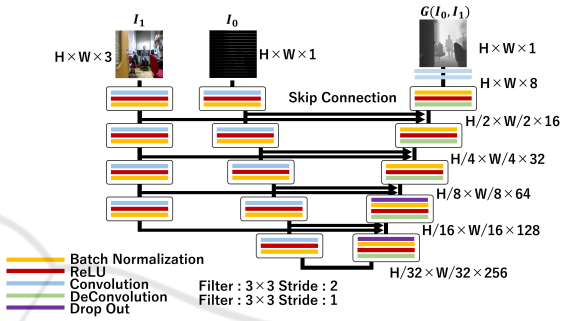



Figure 4: Generator of Proposed method 2.

## 3.1 Network Structure

The network used in this paper consists of a Generator that outputs a high-resolution depth image by inputting a low-resolution depth image and a high-resolution RGB image, and a Discriminator that discriminates between a ground truth depth image and a generated depth image. The low-resolution depth images are converted to sparse high-resolution images before being input to the Generator. However, in the following part of this paper, we use the term low-resolution depth image for sparse high-resolution images obtained by converting low resolution depth images.

The Generator is based on U-net (Ronneberger et al., 2015), which has been used in many image generation tasks. U-net has skip connections, which propagate the feature map of each layer in the encoder to each layer in the decoder. By using the skip con-

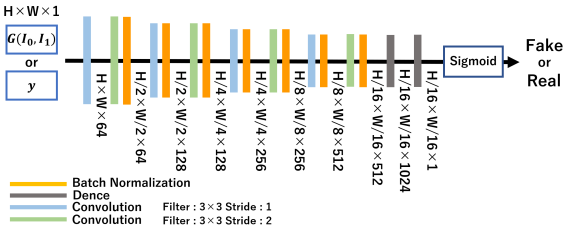
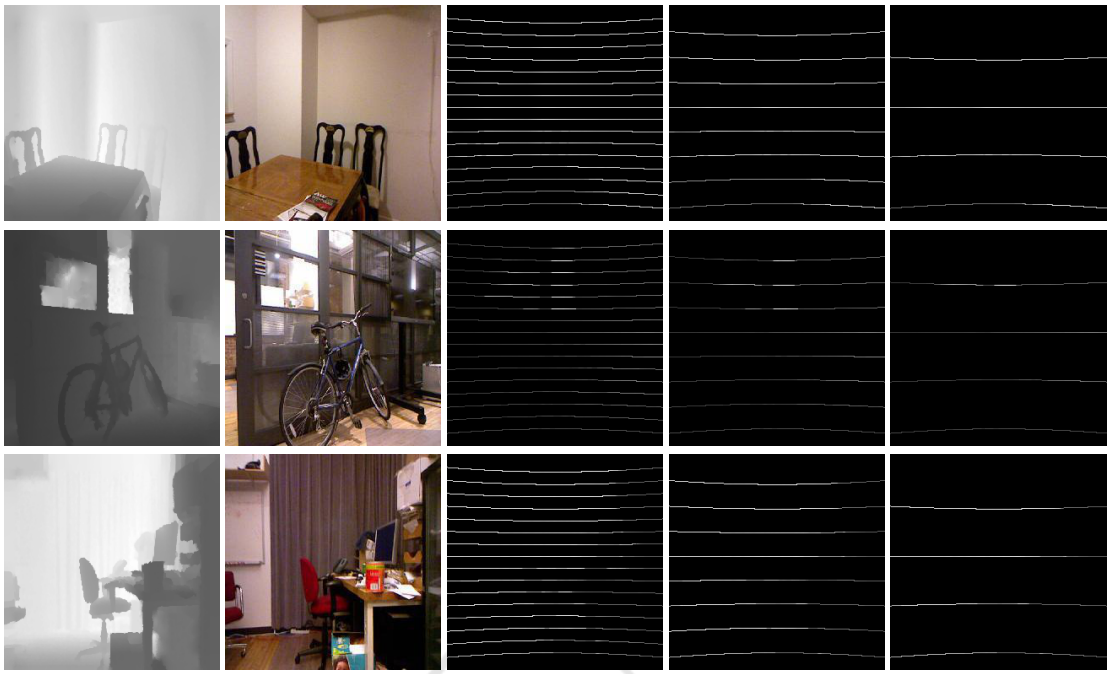



Figure 5: Network structure of discriminator.



(a) HR depth image (ground truth) (b) RGB image (c) LR depth image ($n = 16$) (d) LR depth image ($n = 8$) (e) LR depth image ($n = 4$)

Figure 6: Example of dataset images.

nections, the input information is propagated to the decoder part, and the image conversion can be realized without losing the detailed information of the input image. In each layer of the proposed network, processing was performed in the order of Convolution $\rightarrow$ ReLU $\rightarrow$ Batch Normalization. In order to suppress overfitting, Batch Normalization (Ioffe and Szegedy, 2015) and Drop out layer were incorporated in the bottom two layers of image generation in each layer. When outputting the image, tanh was used as the activation function.

In this research, we propose two methods, method 1 and 2, for using high-resolution RGB image and low-resolution depth image in the Generator. The network structure of method 1 is shown in Fig. 3. This structure has traditionally been used to combine multiple pieces of information, where low resolution depth images ($H \times W \times 1$) and high resolution RGB image ($H \times W \times 3$)) are combined by concat before inputting into the Generator, so the input image is $H \times W \times 4$. On the other hand, the network structure of method 2 is shown in Fig. 4. In this method, the RGB image and the depth image are convolved separately, and the image features in each layer are combined by skip connection. By using this method, it is possible to retain the high resolution information of the RGB image and convolve it into the image feature of the next layer. Therefore, when an image is generated in the decoder of the U-net, a higher resolution depth image can be generated by convolving the image feature that holds the high resolution information.

We next explain the Discriminator used in the proposed method. The structure of the Discriminator used in the proposed method is shown in Fig. 5. We used Patch GAN for the Discriminator. Patch GAN (Pathak et al., 2016) cuts out the image into fine patches and determines the validity of the image for each patch. By using this structure, it becomes possible to judge the validity of image with respect to a local region of the image, and the image validity is measured in various sizes. In each layer, processing is performed in the order of Convolution $\rightarrow$ Batch Normalization. Sigmoid was used as the output activation function.

## 3.2 Network Training

Let $G^*$ be the Generator obtained by training the GAN. Then, the training of our GAN can be described as follows:

$$G^* = \arg\min_G \max_D \mathcal{L}_{GAN}(G,D) + \lambda \mathcal{L}_{L1}(G) \quad (1)$$

where, $\mathcal{L}_{GAN}$ represents the adversarial loss shown in the following equation:

$$\begin{aligned}\mathcal{L}_{GAN}(G,D) &= E_{y\sim p_{data(y)}}[\log D(y)] \\ &+ E_{I_0,I_1\sim p_{data(I_0,I_1)}}[\log(1-D(G(I_0,I_1)))]\end{aligned}$$

On the other hand, $\mathcal{L}_{L1}$ represents L1 loss shown in the following equation:

$$\mathcal{L}_{L1}(G) = E_{y,I_0,I_1\sim p_{data(y,I_0,I_1)}}[||y-G(I_0,I_1)||_1]$$

where, $y$ is the ground truth of the high-resolution depth image, $I_0$ is a low-resolution depth image, and $I_1$ is a high-resolution RGB image.

By training the network as shown in Eq. (1), we obtain Generator which generates high-resolution depth images from low-resolution depth images.

# 4 DATASET

We next explain the data set used in this research. In order to learn the proposed network, pairs of depth image and RGB image is required. Therefore, we constructed a training dataset using NYU Depth Dataset (Silberman and Fergus, 2011). NYU Depth is an indoor image dataset, which consists of 2284 pairs of depth image and RGB image. The depth and RGB images obtained from this dataset were resized to $256 \times 256$, and 2184 pairs were used for training and 100 pairs were used for testing. In this research, we conducted two experiments, a synthetic image experiment in which a low-resolution depth image obtained from LiDAR was created from a high-resolution depth image synthetically, and a real image experiment in which real low-resolution depth images were obtained from LiDAR (Velodyne VLP-16). In both cases, in order to investigate the change in accuracy due to the difference in the amount of information in the low-resolution depth image, we created datasets with different vertical resolutions, $n = 16, 8, and 4$, for low-resolution depth images. That is the number of vertical scan lines of the LiDAR was 16, 8 and 4. The example dataset used in our experiments is shown in the Fig. 6.

# 5 EXPERIMENTS

## 5.1 Synthetic Image Experiments

We next show the results of synthetic image experiments, in which a high-resolution depth image is generated from a low-resolution depth image and an RGB image by using the proposed method. For comparison, we also generated the high-resolution image from just a low-resolution depth image.

Table 1: Accuracy of recovered high-resolution depth image.

| | | LiDAR only | method 1 | method 2 |
|---|---|---|---|---|
| $n = 16$ | RMSE ↓ | 6.6462 | 5.7329 | **5.6673** |
| | PSNR ↑ | 32.187 | 33.4756 | **33.5886** |
| | SSIM ↑ | 0.9453 | 0.9525 | **0.9529** |
| $n = 8$ | RMSE | 11.6271 | **9.2953** | 9.3441 |
| | PSNR | 27.2126 | **29.3198** | 29.1475 |
| | SSIM | 0.9117 | **0.9289** | 0.9239 |
| $n = 4$ | RMSE | 19.2588 | **15.4661** | 16.3567 |
| | PSNR | 22.7405 | **24.7400** | 24.2165 |
| | SSIM | 0.8828 | **0.9020** | 0.8914 |

Generator and Discriminator were trained for 5000 epochs. The batch size was 32, and Adam (Kingma and Ba, 2014) was used with a learning rate of 0.001 for learning optimization.

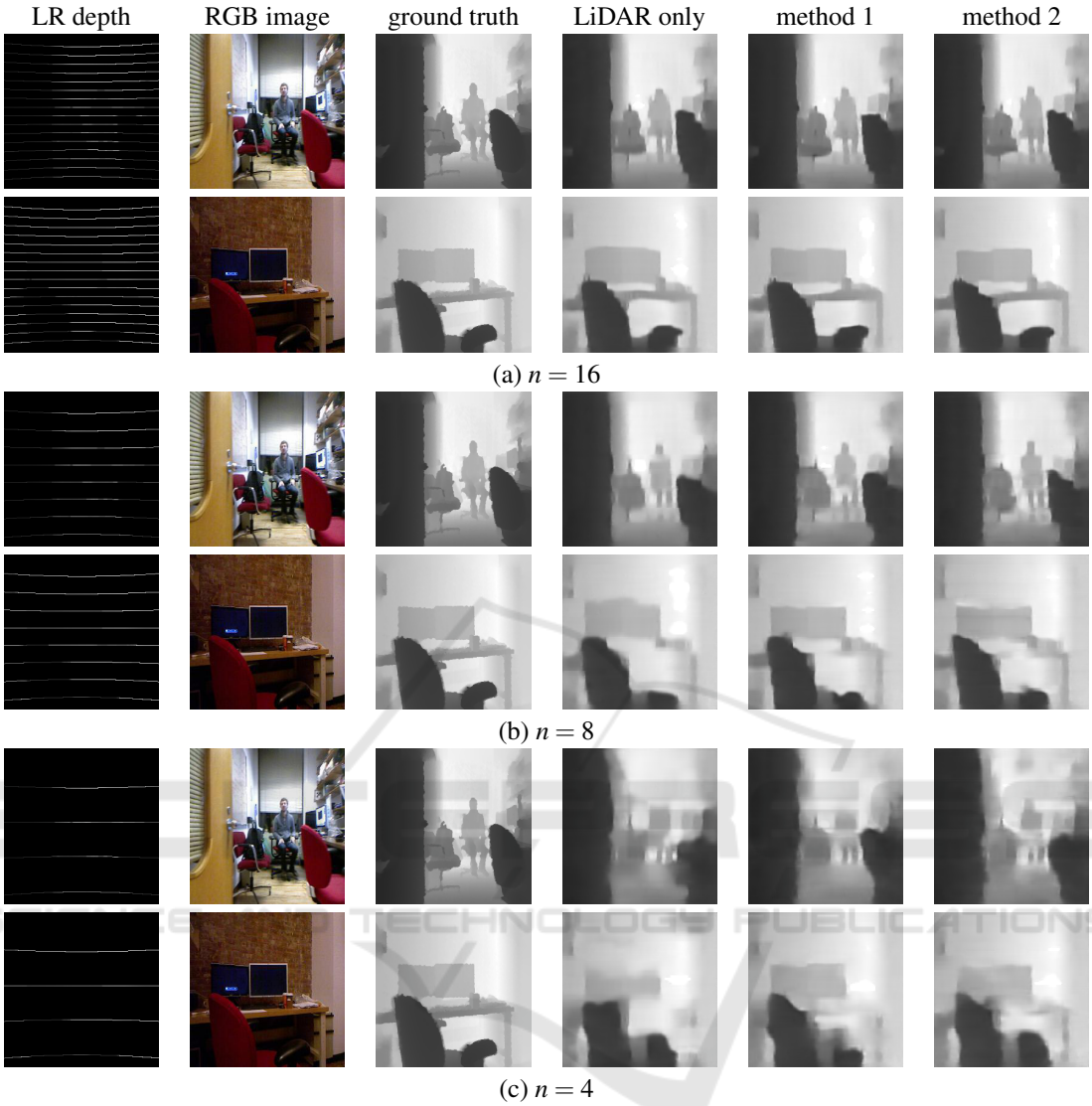
For each low-resolution depth image of $n = 16$, 8, and 4, the network was trained by using 2184 training data, and high-resolution depth images were generated from 100 test low-resolution images by using the trained network.

The experimental results are shown in Fig. 7. From the result of $n = 16$ in Fig. 7 (a), we find that the difference between the proposed method and the existing method with only depth images is small. However, as the vertical resolution of the input depth image decreases to $n = 8$ and $n = 4$, the degradation of the result in the existing method becomes very large, and we find that the proposed method combining RGB images can recover the high-resolution depth image more accurately. For example, we find that the shapes of the desk and chair are distorted in the existing method, whereas the proposed method can recover them more accurately.

Table 1 shows the accuracy of the recovered 100 high-resolution depth images in RMSE, PSNR, and SSIM. From this table, we find that in any case of vertical resolution, the proposed method using the RGB image and the depth image can generate more accurate high-resolution depth images than the existing method.

## 5.2 Real Image Experiments

We next show the results obtained from real image experiments. Similar to the synthetic image experiments, training was performed with NYU Depth dataset, and the low-resolution depth image obtained from a LiDAR (Velodyne VLP-16) were input to the trained network to evaluate the performance of the proposed method. We tested the proposed method and the existing method while changing the vertical resolution of LiDAR to $n = 16$, 8 and 4. Calibration of the data between the RGB camera and LiDAR was conducted in advance by using projective transformation.




(a) $n = 16$

(b) $n = 8$

(c) $n = 4$

Figure 7: Synthetic image experiments.

The recovered high resolution depth images obtained from the proposed method and the existing method are shown in Fig. 8. As shown in this figure no difference was observed when $n = 8$ and $n = 4$, but in the result of $n = 16$, we fined that the monitor and desk in the upper scene were recovered more accurately by the proposed method.

Although we need more systematic evaluations, the results of the synthetic image experiments and real image experiments show the effectiveness of the proposed method.

## 6 CONCLUSION

In this paper, we proposed a method for obtaining high-resolution depth images from low-resolution depth data obtained from LiDAR. In particular we proposed a GAN based network that combines high-resolution RGB image with low-resolution depth image.

We conducted synthetic and real image experiments to generate a high-resolution depth images using the proposed network. In the synthetic image experiments, we used NYU Depth dataset for training and testing, and showed that the proposed method can generate high-resolution depth images

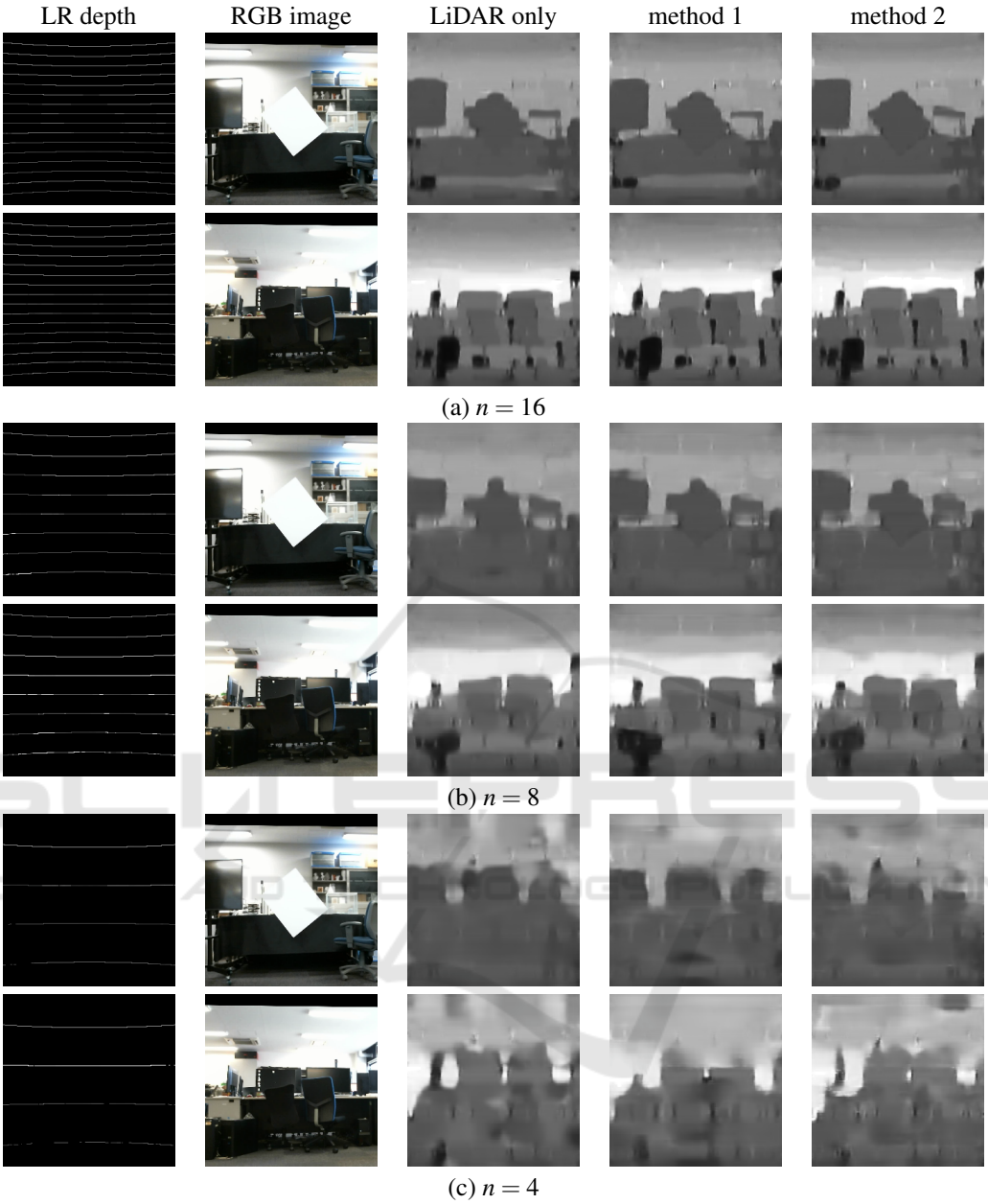



Figure 8: Real image experiments.

more accurately than the existing method that uses only low-resolution depth images as input. We also conducted experiments using real LiDAR data and showed that the proposed method can generate more accurate high-resolution depth images.

Although the study is still in its early stage, we will evaluate various network structures in the future to show the effectiveness of image inpainting using guide images.

# REFERENCES

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631.

Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*.

Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Hirschmuller, H. (2007). Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., and Navab, N. (2016). Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Silberman, N. and Fergus, R. (2011). Indoor scene segmentation using a structured light sensor. In *Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition*.

Vogiatzis, G., Torr, P. H., and Cipolla, R. (2005). Multi-view stereo via volumetric graph-cuts. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 391–398. IEEE.