Image-quality Improvement of Omnidirectional Free-viewpoint Images by Generative Adversarial Networks

Oto Takeuchi[1], Hidehiko Shishido[1], Yoshinari Kameda[1], Hansung Kim[2] and Itaru Kitahara[1]
[1]*University of Tsukuba, Tsukuba, Ibaraki, Japan*
[2]*University of Surrey, Guildford, Surrey, U.K.*

Keywords: Free-viewpoint Image, Omnidirectional Image, Image-quality Improvement, Generative Adversarial Networks.

Abstract: This paper proposes a method to improve the quality of omnidirectional free-viewpoint images using generative adversarial networks (GAN). By estimating the 3D information of the capturing space while integrating the omnidirectional images taken from multiple viewpoints, it is possible to generate an arbitrary omnidirectional appearance. However, the image quality of free-viewpoint images deteriorates due to artifacts caused by 3D estimation errors and occlusion. We solve this problem by using GAN and, moreover, by focusing on projective geometry during training, we further improve image quality by converting the omnidirectional image into perspective-projection images.

1 INTRODUCTION

Image shooting with an omnidirectional camera (360-camera) is an effective technique for observations around an environment. In recent years, this technique has attracted more attention for its ability to achieve immersive observations in combination with a head-mounted display. In Google Street View (Google, 2007), multi-directional observation with a moving viewpoint is possible by properly choosing omnidirectional images shot from multiple viewpoints.

By applying a 3D estimation process such as Structure from Motion (SfM) to the omnidirectional multi-viewpoint images, it is possible to estimate the position and rotation of the omnidirectional camera and the 3D shape of the target space. We proposed the Bullet-Time video generation method to smoothly switch the viewpoint while gazing at the point to be observed using the estimated information (Takeuchi et al., 2018). In this method, omnidirectional observation is possible only at the captured viewpoint, not at non-captured positions. When the interval between the multi-viewpoint cameras becomes wider, the smoothness of viewpoint movement is degraded. Moreover, another serious problem is the complete inability of the viewer to move the viewpoint from the capturing viewpoint.

Free-viewpoint image generation with the aim of reproducing an appearance from an arbitrary viewpoint is one of the most active research fields in computer vision (Agarwal et al., 2009; Kitahara et al., 2004; Kanade et al., 1997; Shin et al., 2010; Newcombe et al., 2011; Orts-Escolano et al., 2016; Seitz et al., 1996; Levoy et al., 1996; Tanimoto et al., 2012; Matusik et al., 2000; Hedman et al., 2016), but artifacts due to 3D reconstruction errors (caused by an error in correspondence search) and occlusion, which degrade the image quality, are still important research issues. It is possible to improve 3D reconstruction accuracy by using devices that acquire depth information, such as RGB-D cameras (Newcombe et al., 2011; Orts-Escolano et al., 2016; Hedman et al., 2016), but this reduces the simplicity of the capturing system, making it more difficult for use in practical applications. We attempt to solve this issue by using an omnidirectional camera. Among multiple omnidirectional images, there are many overlapping areas due to the wide field of view. As a result, the same region in the 3D space is observed from various viewpoints, and thus the accuracy of the correspondence search can be improved.

Research has been conducted to recover the degraded image quality by using an image reconstruction technique (Barnes et al., 2009). In recent years, methods using deep learning have been proposed

Takeuchi, O., Shishido, H., Kameda, Y., Kim, H. and Kitahara, I.
Image-quality Improvement of Omnidirectional Free-viewpoint Images by Generative Adversarial Networks.
DOI: 10.5220/0008959802990306
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 299-306
ISBN: 978-989-758-402-2; ISSN: 2184-4321

(Pathak et al., 2016; Iizuka et al., 2017), and more natural image-quality improvement has been achieved. However, these methods are based on the assumption that the region to be complemented is known. On the other hand, in free-viewpoint video generation, it is difficult to identify regions of low image quality, since this depends on the capturing condition. This makes it difficult to apply the conventional image reconstruction technique to solving image-quality degradation.

In this paper, we employ deep learning by generative adversarial networks (GAN) to learn the relationship in appearance between generated omnidirectional free-viewpoint (OFV) images and captured images. By using the learning results (Generator of GAN), a method to improve the image quality of OFV images has been developed. It is well known that the variation in training data affects the efficiency of deep learning. The appearance of an omnidirectional image is significantly distorted by a unique optical system. Therefore, when the viewpoint of the omnidirectional camera changes, the appearance of the same region is also drastically changed. In other words, the same region is observed with various appearances. We reduce changes in appearance due to lens distortion to improve the learning efficiency of deep learning. In particular, we divide an omnidirectional image into multiple perspective projection images to reduce the variation in appearance.

2 RELATED WORKS

2.1 Display of Multi-viewpoint Omnidirectional Images

In Google Street View (Google, 2007), it is possible to observe the surrounding view by using omnidirectional images. By switching omnidirectional images shot from multiple viewpoints according to the viewpoint movement specified by the observer, it is possible to grasp the situation in more detail while looking around the scene. By combining image-blending processing and image-shape transform, the observer gets the sensation that he/she is moving around the scene. We also estimate the position and rotation of the omnidirectional camera and the 3D shape of the capturing space by applying 3D reconstruction processing to the multi-viewpoint omnidirectional images. Using the estimated 3D information, we developed the Bullet-Time video generation method to switch the viewpoint while gazing at the point to be observed (Takeuchi et al., 2018). However, the omnidirectional image-switching method has the problem of allowing the viewer to move only at the capturing position.

2.2 Free-viewpoint Images

There has been much research on free-viewpoint images. Model-based rendering (MBR) (Agarwal et al., 2009; Kitahara et al., 2004; Kanade et al., 1997; Shin et al., 2010; Newcombe et al., 2011; Orts-Escolano et al., 2016) reproduces a view from an arbitrary viewpoint using a 3D computer graphics (CG) model reconstructed from multi-viewpoint images of the capturing space. Image-based rendering (IBR) (Seitz et al., 1996; Levoy et al., 1996; Tanimoto et al., 2012; Matusik et al., 2000; Hedman et al., 2016) synthesizes the appearance directly from the captured multiple viewpoint images.

In MBR, the quality of the generated free-viewpoint images depends on the accuracy of the reconstructed 3D CG model. For this reason, when capturing a complicated space where a 3D reconstruction error is likely to occur, an artifact may occur in the generated view. Furthermore, the occlusion inherent in observations with multiple cameras makes it challenging to reconstruct an accurate 3D shape, thus degrading the quality of generated images (Shin et al., 2010).

Since IBR does not explicitly reconstruct the 3D shape but applies a simple shape, it is possible to generate free-viewpoint images without considering the complexity of the capturing space. However, when the applied shape of the capturing space is largely different from the actual shape, the appearance of the generated view is significantly distorted by the image fitting error. To reduce this distortion and generate an acceptable view, it is necessary to increase the number of capturing cameras.

2.3 Image-quality Improvement

Research on image-quality improvement has been conducted actively. There is a method that complements the appearance of the image by finding the corresponding image information using peripheral image continuity (Barnes et al., 2009), and this method has also been applied to complement free-viewpoint video (Shishido et al., 2017). However, this method cannot reconstruct information that is not observed in the image. Various approaches of using convolutional neural networks and GAN to reconstruct information not included in the image have been proposed, but these methods assume that the missing region is known (Pathak et al., 2016; Iizuka et al., 2017). By applying reconstruction utilizing GAN to transform

the entire image (Isola et al., 2017), we propose a method to reproduce an appearance that is equivalent to the captured image by compensating for the image-quality degradation due to movement of the viewpoint.

3 IMAGE-QUALITY IMPROVEMENT OF OFV IMAGE

Figure 1 shows an overview of our proposed method. By applying SfM to multi-viewpoint omnidirectional images capturing the target space, the position and rotation of each omnidirectional camera and the 3D point cloud of the target space are estimated. Based on the estimated camera parameters, the 3D point cloud is projected onto each omnidirectional image plane to generate sparse depth images. By interpolating the gap among the projected points, dense omnidirectional depth images are generated at each viewpoint. It is possible to synthesize an omnidirectional image at any viewpoint by using the omnidirectional depth image and the captured omnidirectional image as the texture. As a result, we obtain a dataset of actually captured omnidirectional images and synthesized omnidirectional images at the same viewpoints.

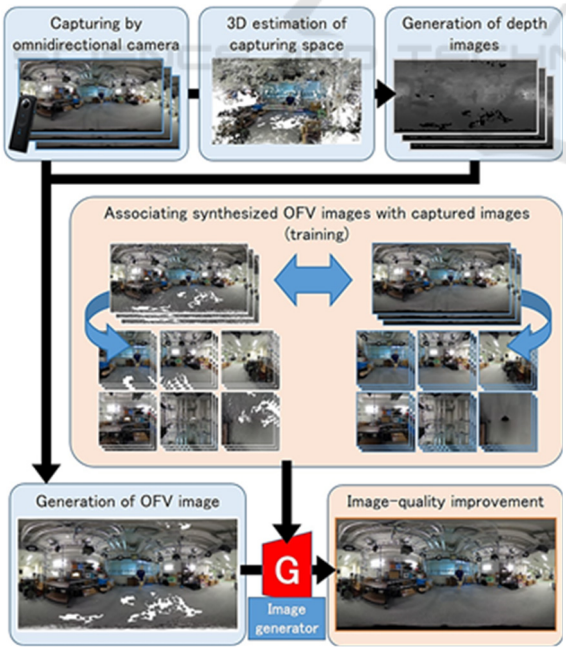


Figure 1: Image-quality improvement of OFV images.

We apply the dataset to GAN, which learns a way to generate the appearance of an image from the synthesized image. By using the results of deep learning (image generator) provided by GAN, the image-quality of the synthesized OFV image can be improved.

4 GENERATION METHOD FOR OFV IMAGES

4.1 Capturing Multiple Omnidirectional Images and 3D Estimation

Multiple omnidirectional images are captured at various viewpoints surrounding a target object. Due to the active research and development on 3D information estimation from multi-viewpoint images, some excellent SfM libraries (Wu, 2011; Schönberger et al., 2016; Sweeney et al., 2015) have become available. However, these libraries are usually based on perspective projection, which is different from the projective geometry of an omnidirectional image. Therefore, in our method, we divide an omnidirectional image into perspective images (i.e., virtually setting cameras using perspective geometry) and apply an SfM library to each perspective projection image captured by a virtual camera. As a result, the camera parameters of the images and sparse 3D point clouds are estimated. The position and orientation of the omnidirectional camera can be calculated from the estimated camera parameters of the virtual cameras (Takeuchi et al., 2018). Based on the estimated camera parameters and sparse 3D point cloud, multi-view stereo processing (Seitz et al., 2006) is carried out to obtain a dense 3D point cloud.

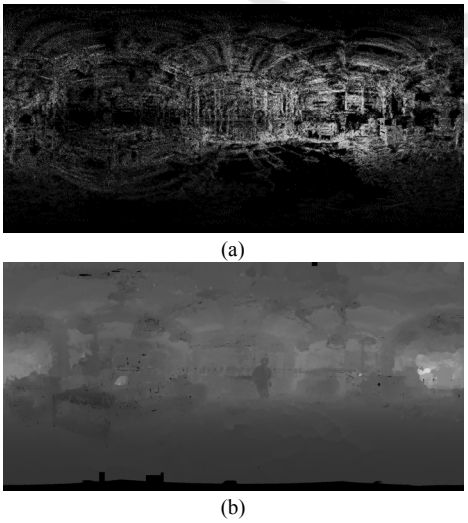
4.2 Generation of Omnidirectional Depth Images

By calculating the distance from each viewpoint of multiple omnidirectional cameras to the 3D point cloud estimated in Section 4.1, we generate the sparse omnidirectional depth image shown in Figure 2(a). We calculate the color difference between the projected 3D point cloud and the pixel of the captured image at the viewpoint where the depth information is generated. The color difference is calculated as the Euclidean distance between the two colors described in the CIELAB color space. This color difference increases when the 3D information of the point cloud is estimated incorrectly. In order to reduce the error of 3D information, we apply threshold processing to the color difference. When the color difference is 20.0 or more, the depth value is not calculated.

Since we cannot estimate the depth value of the pixels where the 3D point cloud is not projected, as shown in Figure 2(a), there are vast missing regions in a depth image. We interpolate these regions using a cross bilateral filter (Chen et al., 2012). The cross bilateral filter uses two different modal images (e.g., a color image and the depth image). It filters one of the images based on the other one that has smaller observation noise. In our case, depth images having much observation noise are filtered using captured color images having smaller observation noise. The following filter equations are applied:

$$D_p = \frac{\sum_{r \in N} d(\boldsymbol{p},\boldsymbol{r})\, c(I_p,I_r)\, D_r}{\sum_{r \in N} d(\boldsymbol{p},\boldsymbol{r})\, c(I_p,I_r)},$$
$$d(\boldsymbol{p},\boldsymbol{r}) = exp\left(-\frac{\|\boldsymbol{p}-\boldsymbol{r}\|_2}{2\sigma_1^2}\right), \quad (1)$$
$$c(I_p,I_r) = exp\left(-\frac{\|I_p-I_r\|_2}{2\sigma_2^2}\right),$$

where $\boldsymbol{p}$ is the pixel coordinate of interest, $\boldsymbol{r}$ is the reference pixel coordinate, D is the depth value, I is the luminance value, N is the set of reference pixel coordinates, and σ is a constant. $d(\boldsymbol{p},\boldsymbol{r})$ and $c(I_p,I_r)$ are weights for distance and color difference, respectively. We calculate the depth value by weighting the distance between the pixel position of interest and the reference pixel position as well as the color difference on the captured image. As a result, as shown in Figure 2(b), it is possible to interpolate the depth image while maintaining the contour of the captured image.



(a)

(b)

Figure 2: Generated omnidirectional depth image. (a): Before interpolation processing. (b): After interpolation processing.

4.3 Generation of OFV Image

As shown in Figure 3, an OFV image at an arbitrary viewpoint is generated from the omnidirectional image captured in Section 4.1 and the omnidirectional depth image created in Section 4.2.

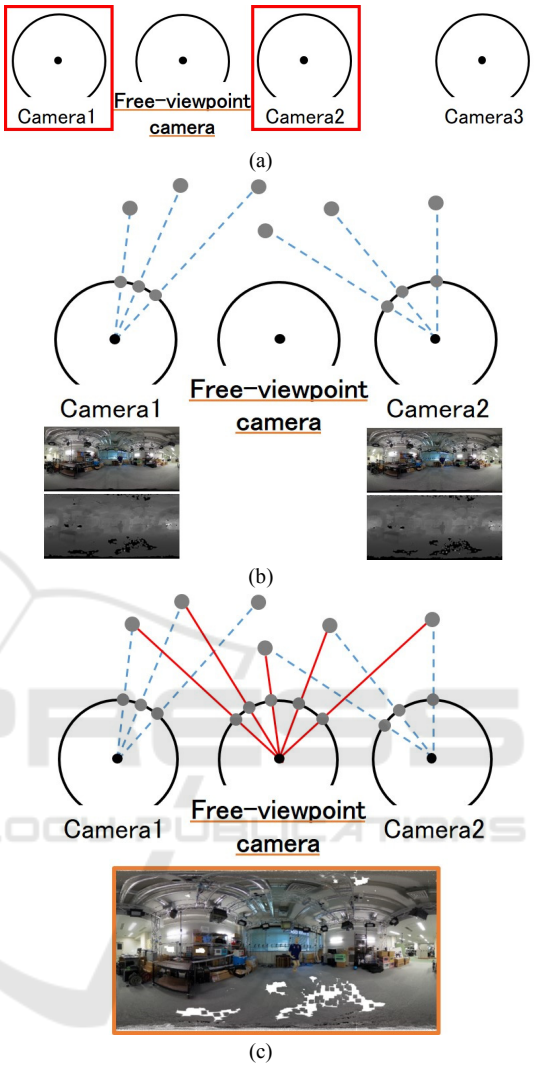


Figure 3: Generation method for an OFV image. (a): Select multi-view cameras to be used for free-viewpoint image. (b): By referring to the depth information, every pixel value (color information) of the captured multiple omnidirectional images is projected onto a 3D space. (c): The OFV image is generated by back-projecting these 3D point clouds onto the omnidirectional image plane.

When the free viewpoint for generating a new omnidirectional image is determined, the distance from the free viewpoint to each multi-view camera is calculated. Then, a certain number of multi-view cameras are selected in order from the closest one (Figure 3(a)). By referring to the depth information, every pixel value (color information) of the captured multiple omnidirectional images is projected onto a 3D

space to generate a dense 3D point cloud model (Figure 3(b)). The OFV image is generated by back-projecting these 3D point clouds onto the omnidirectional image plane at the free-viewpoint. When different point clouds are projected on the same pixel of a free-viewpoint image, the closer point cloud from the free viewpoint is adopted to remove the hidden surface (Figure 3(c)).

With the same processing, it is possible to generate a free-viewpoint image at the viewpoint where multiple omnidirectional images are actually captured. We prepare a learning dataset (a pair of synthesized free-viewpoint images and captured images) for GAN used in image-quality improvement, which is described in the next section.

5 IMAGE-QUALITY IMPROVEMENT

Some artifacts are observed in the OFV images generated in Section 4.3. Typical causes of these artifacts include 3D shape estimation errors and missing 3D information due to occlusion. This section describes how to reduce these problems using GAN. In this research, we employee Pix2Pix (Isola et al., 2017) as a way to implement GAN. Pix2Pix is a type of conditional GAN that learns the correspondence between two images of different styles, such as line-drawn images and photos or aerial photos and maps, and then converts one to the other. In this research, Pix2Pix is applied to image conversion between a free-viewpoint image and a captured image to improve the quality of free-viewpoint images.

Pix2Pix consists of two networks: an image generator and a discriminator. A pair of pre-conversion and post-conversion images are prepared as training data, the pre-conversion image is input to the image generator, and either the image generated by the image generator or the prepared post-conversion image is input to the discriminator. The discriminator determines which image is input. Learning is done while the images compete with each other, so the image generator can deceive the discriminator, while the discriminator can make an accurate decision. After learning, image conversion is achieved by using an image generator.

As the training data, the OFV image synthesized at the capturing viewpoint in Section 4.3 is prepared as the pre-conversion image, and the omnidirectional image captured in Section 4.1 is prepared as the post-conversion image. After the image generator is trained using the training data, an OFV image at the virtual viewpoint is given as an input to the learned image generator to generate a highly realistic image with reduced image-quality degradation.

We focus on the projective geometry of learning images to achieve learning efficiency. The diversity of appearance among learning samples increases, making learning difficult because omnidirectional images based on equirectangular projections cause a significant change in appearance due to the movement of the viewpoint based on their projection characteristics. Therefore, we reduce the diversity of appearance by dividing the omnidirectional images into multiple perspective projection images and then perform efficient GAN learning. In this paper, as shown in Figure 4, we adopt cube mapping to divide an omnidirectional image into six image planes and construct an image generator using perspective projection images on each plane.

Figure 4: Division of an omnidirectional image into six image planes by cube mapping.

6 EXPERIMENTS

6.1 Experimental Environment

We conducted demonstration experiments on the effect of improving the image quality of OFV images by deep learning and on the impact of image division on learning efficiency. As shown in Figure 5, we installed a tripod with an omnidirectional camera (RICOH THETA S) at 42 viewpoints in the indoor environment (University of Tsukuba) and captured multi-view omnidirectional images. For the processing, we used a notebook PC with the following specifications: CPU: Intel Core i7-7700HQ 2.8 GHz,

GPU: NVIDIA GeForce GTX 1060, Memory: 16 GB RAM. SfM was executed using VisualSFM (Wu, 2011). We generated 42 OFV images at the capturing viewpoints using the method described in Section 4.3. Of these, we used 22 OFV images, as well as images shot from the same viewpoint as these images, as the Pix2Pix training data.

To verify the learning effect of GAN by the image division described in the previous section, we trained the image generator for the case of using an equirectangular image as is and for the case of using a perspective projection image divided by cube mapping. The OFV image based on equirectangular projection was 2,048 × 1,024 pixels, each perspective projection image was 512×512 pixels, and the number of learning steps was 1,000 epochs. For evaluation, we input the 20 OFV images that were not used for training to the image generator and observed the generated images. Moreover, the image quality was quantitatively evaluated using the peak signal-to-noise ratio (PSNR), which is one of the image-quality evaluation indexes.

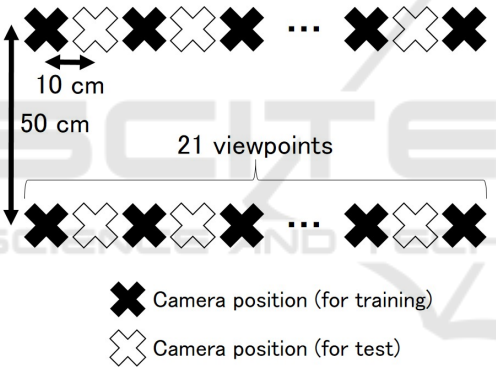


Figure 5: Arrangement of omnidirectional cameras in capturing experiments (viewed from above).

6.2 Results

Figure 6 compares examples of the images generated. Figure 6(a) is an OFV image (before image-quality improvement) made by the method described in Section 4. Figures 6(b, c) are OFV images with improved image quality: Figure 6(b) is the case where the divided image is input, and Figure 6(c) is the case where the omnidirectional image of the equirectangular projection is input. Figure 6(d) shows the captured image (correct image). Comparing Figure 6(a) with (b, c), we can confirm that the image generator constructed by deep learning improves the missing regions in the image. Comparing Figure 6(b) with (c), the former, which uses the divided images as input, produces a more precise image with fewer artifacts and less blur.

Using the average value of PSNR calculated from OFV images at 20 viewpoints, we perform a quantitative evaluation on the effect of image-quality improvement and the presence or absence of image division. Table 1 shows the evaluation results.

Table 1 shows that PSNR is improved and the image generator constructed by deep learning improves the image quality. In addition, the image-dividing method produces a higher PSNR value than the non-dividing method, thus confirming the effectiveness of image division.

Table 1: Average PSNR with standard deviation in 20 viewpoints images.

Before image-quality improvement	After image-quality improvement	
	With image division	Without image division
12.68 (±1.37) dB	27.39 (±0.45) dB	23.01 (±0.18) dB

7 CONCLUSIONS

In this paper, we proposed an image-quality improvement method for OFV images using GAN. We reconstructed the 3D information of the capturing space from an omnidirectional multi-viewpoint image and generated the OFV image after interpolation of the depth information by image processing. By using deep learning (GAN), we improved the image quality of artifacts and the missing regions observed in conventional free-viewpoint images. By focusing on the projective geometry during training, we raised the performance of image-quality improvement by converting an omnidirectional image into perspective projection images.

This work was partially supported by JSPS KAKENHI Grant Number 17H01772 and by JST CREST Grant Number JPMJCR14E2, Japan.

REFERENCES

Google, 2007. Google Street View. See: https://www.google.com/streetview/

Takeuchi, O., Shishido, H., Kameda, Y., Kim, H., and Kitahara, I., 2018. Generation Method for Immersive Bullet-Time Video Using an Omnidirectional Camera in VR Platform. Proc. of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia, pages 19-26.

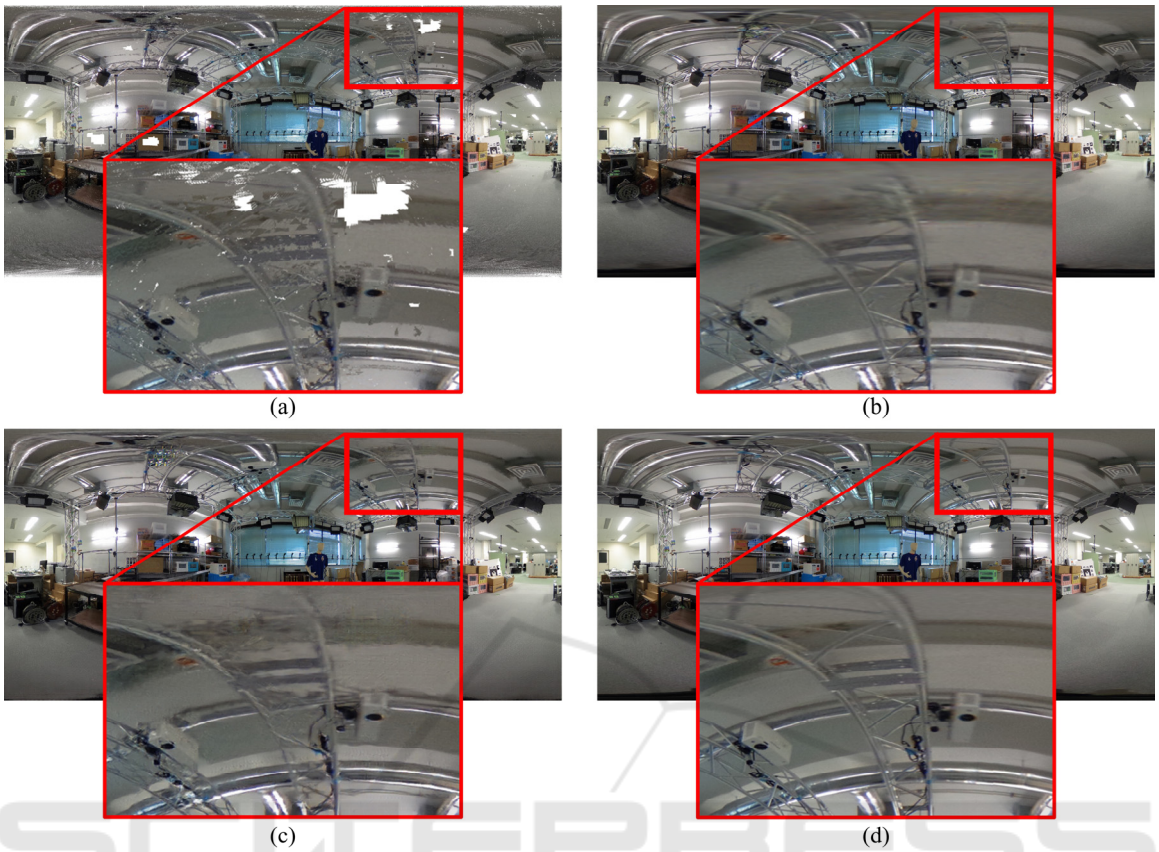


Figure 6: Comparison results (top) and enlarged views (bottom). (a): OFV image (no image-quality improvement). (b): Proposed method using learning by image division (with image-quality improvement). (c): Proposed method using learning with omnidirectional images (with image-quality improvement). (d): Correct image (captured image).

Agarwal, S., Snavely, N., Simon, I., Seitz, S. M., and Szeliski, R., 2009. Building Rome in a Day. International Conference on Computer Vision, 8 pages.

Kitahara, I. and Ohta, Y., 2004. Scalable 3D Representation for 3D Video in a Large-Scale Space. Presence: Teleoperators and Virtual Environments, 13(2):164-177.

Kanade, T., Rander, P., and Narayanan, P. J., 1997. Virtualized reality: Constructing virtual worlds from real scenes. IEEE MultiMedia, 4(1):34–47.

Shin, T., Kasuya, N., Kitahara, I., Kameda, Y., and Ohta, Y., 2010. A Comparison Between Two 3D Free-Viewpoint Generation Methods: Player-Billboard and 3D Reconstruction. 3DTV Conference, 4 pages.

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohli, P., Shotton, J., Hodges, S., and Fitzgibbon, A., 2011. KinectFusion: Real-time dense surface mapping and tracking. IEEE International Symposium on Mixed and Augmented Reality, 10 pages.

Orts-Escolano, S., Rhemann, C., Fanello, S., Chang, W., Kowdle, A., Degtyarev, Y., Kim, D., Davidson, P. L., Khamis, S., Dou, M., Tankovich, V., Loop, C., Cai, Q., Chou, P., Mennicken, S., Valentin, J., Pradeep, V., Wang, S., Kang, S. B., Kohli, P., Lutchyn, Y., Keskin, C., and Izadi, S., 2016. Holoportation: Virtual 3D Teleportation in Real-time. Proc. of the 29th Annual Symposium on User Interface Software and Technology, pages 741-754.

Seitz, S. M., and Dyer, C. R., 1996. View Morphing. Proc. of SIGGRAPH, pages 21-30.

Levoy, M. and Hanrahan, F., 1996. Light Field Rendering. Proc. of SIGGRAPH, pages 31-42.

Tanimoto, M., 2012. FTV: Free-viewpoint television. Signal Processing: Image Communication, 27(6):555–570.

Matusik, W., Buehler, C., Raskar, R., Gortler, S. J., and McMillan, L., 2000. Image-Based Visual Hulls. Proc. of SIGGRAPH, pages 369-374.

Hedman, P., Ritschel, T., Drettakis, G., and Brostow, G., 2016. Scalable Inside-out Image-based Rendering. ACM Transactions on Graphics, 35(6):231:1–231:11.

Barnes, C., Shechtman, E., Finkelstein, A., and Goldman, D. B., 2009. PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing. ACM Transactions on Graphics, 28(3):24:1–24:11.

Shishido, H., Yamanaka, K., Kameda, Y., and Kitahara, I., 2017. Pseudo-Dolly-In Video Generation Combining 3D Modeling and Image Reconstruction. ISMAR 2017 Workshop on Highly Diverse Cameras and Displays for Mixed and Augmented Reality, pages 327-333.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A., 2016. Context Encoders: Feature Learning by Inpainting. IEEE Conference on Computer Vision and Pattern Recognition, 9 pages.

Iizuka, S., Simo-Serra, E., and Ishikawa, H., 2017. Globally and Locally Consistent Image Completion. Proc. SIGGRAPH, 36(4)21-30.

Isola, P., Zhu, J., Zhou, T., and Efros, A. A., 2017. Image-to-Image Translation with Conditional Adversarial Networks. IEEE Conference on Computer Vision and Pattern Recognition, 10 pages.

Wu, C., 2011. VisualSFM: A Visual Structure from Motion System. See: http://ccwu.me/vsfm

Schönberger, J. L. and Frahm, J., 2016. Structure-from-Motion revisited. IEEE Conference on Computer Vision and Pattern Recognition, 10 pages.

Sweeney, C., Höllerer, T. H., and Turk, M., 2015. Theia: A Fast and Scalable Structure-from-Motion Library. ACM International Conference on Multimedia, 4 pages.

Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R., 2006. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms, IEEE Conference on Computer Vision and Pattern Recognition, 8 pages.

Chen, L., Lin, H., and Li, S., 2012. Depth image enhancement for Kinect using region growing and bilateral filter. Proc. of the 21st International Conference on Pattern Recognition, 4 pages.