Monocular Depth Estimation Based on Lens Aberrations

Naoki Nishizawa[1], Masako Kashiwagi[1], Nao Mishima[1], Akihito Seki[1]

naoki1.nishizawa@toshiba.co.jp
[1]Toshiba Corporate Research & Development Center, Japan
Keywords: Deep Depth from Aberration Map, Depth Estimation, Lens Aberration, Smartphone Camera

ABSTRACT

We have developed our Deep Depth from Aberration Map (DDfAM) that can obtain a valid depth map from a single-shot image with a monocular camera even if no contextual information exists. In this paper, we explain two of our recent efforts: application for smartphone cameras and depth inference acceleration.

1 Introduction

Depth measurement from a single-shot image is becoming more and more important in various fields, such as autonomous driving of mobile vehicles, augmented reality, and infrastructure inspection. Although a lot of methods for obtaining depth from images have already been proposed, scene-independent depth measurement using small devices is still desired. For example, in infrastructure inspection work, there is a growing need for object measurement that can be easily performed with smartphone cameras to improve work efficiency. Stereo camera systems [1] have been widely used as an example of image-based depth measurement, but these systems require a wide baseline for accurate depth and are difficult to apply to small devices. Recently, various methods for context-based depth estimation using a single image have been proposed [2]. However, these methods are dependent on the learned contextual information, so their robustness to scenes of new domains is an issue.

We proposed a Deep Depth from Aberration Map (DDfAM) [3], a novel approach of physics-based single-shot depth estimation utilizing lens aberration map, which contains various types of aberrations corresponding to positions of image and distance from the image sensor. While it had the advantage of robustly estimating the depth of the scene, the validation of our approach was limited to images from DSLR cameras and had not yet been tested on small devices, especially smartphone cameras. In addition, our conventional network estimated depths by sliding window, resulting in slow processing speed. This means that even if this method could be applied to smartphone cameras, the advantage of smartphones, their ease of use, would be lost.

In this paper, we present a single-shot depth measurement using a smartphone camera with DDfAM. We also propose a network structure that speeds up depth inference for practical use in smartphones. This allows us, for example, to get the depth map immediately as long as we take an image with our smartphone and upload it to the inference server (Fig. 1), which can be applied to efficient inspection systems or real-time size measurement.

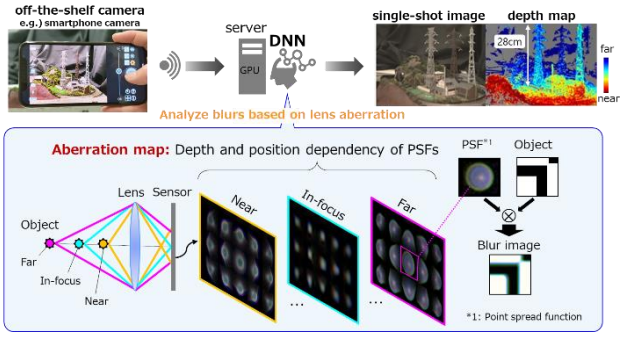


Fig. 1 Overview of DDfAM.

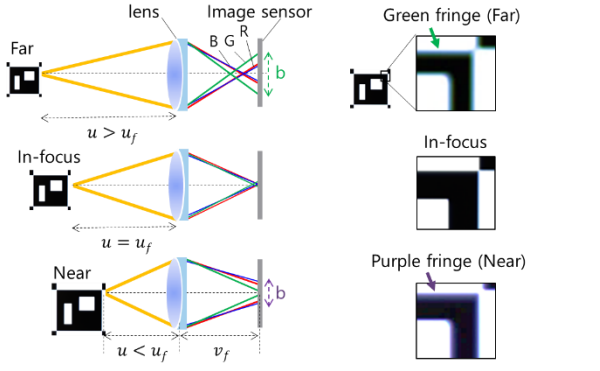
2 Method

2.1 Depth measurement with DDfAM

In the DDfAM framework, as with Depth from Defocus (DfD) [4], we estimate the defocus blur radius from an image. A blur radius b is derived from the following formula,

$$b = \frac{av_f}{2p}\left|\frac{1}{f} - \frac{1}{u} - \frac{1}{v_f}\right|, \quad (1)$$

where u, a, f, F, p, and v_f are object distance, lens aperture, focal length, aperture number, sensor pitch, and distance between the lens and the image sensor, respectively. As can be seen from Eq. (1), the defocus blur can be the same value on the far or near side of the focal plane. Consequently, the DfD method requires two or more images to determine whether the depth is on the far or near side. In contrast, in the DDfAM approach, we estimate "signed" defocus blur, which is unique with measured distance, based on analyzing an aberration



(a) Light path diagram of achromatic lens

(b) Example of fringes produced by actual chromatic aberration

Fig. 2 Optical rays and chromatic aberration.

map.

Fig. 2 shows the relationship between optical rays passing through a lens and chromatic aberration. Fig. 2(a) shows how the R, G, B rays travel when chromatic aberration is caused by the camera lens when the distance to the object u is closer to (near), the same as (in-focus) and farther from (far) the camera focus distance u_f. The achromatic lens is made of a low-dispersion convex lens and a high-dispersion concave lens to compensate for a blue and red light so that it can suppress the spread of light. However, since it is still difficult to remove lens aberrations completely, the light does not converge to a single point. When $u > u_f$ (far), the green blur is the largest and the edges of texture appear green. When $u < u_f$ (near), the purple blur is the largest and the edges appear purple, which is a mixture of red and blue. These are called green fringing and purple fringing, respectively (Fig. 2(b)).

Actually, due to the off-axial aberrations such as coma aberration, the color pattern of defocus blur depends on the position in the image as well as the distance of an object. We call images which contains this position-dependent and distance-dependent aberrations as aberration map (A-Map). In DDfAM, by analyzing this A-Map with DNN, it is possible to estimate the "signed" defocus blur radius $\hat{b}$ reflecting whether the object is on the near or far plane.

$$\hat{b} = \frac{av_f}{2p}\left(\frac{1}{f} - \frac{1}{u} - \frac{1}{v_f}\right) \qquad (2)$$

This value is positive when the object is on the far plane, and negative when the object is on the near plane. The definition of signed defocus blur makes its value unique with the distance u, regardless of whether the object is in the near or far plane.

In the A-Map analysis network framework, Bayer-array images with the RAW format are used as input. After preprocessing such as demosaicing, resizing, and white balance correction, they are converted to blur maps by DNN. The networks learn the relationship between blur features and ground truth distances, trained as a regression problem with supervision. For collecting training and testing datasets, we develop an indoor experimental system containing a moving stage and a screen as shown in Fig. 3. We put a camera on the moving stage and capture images of MSCOCO [7] tiling patterns displayed on the screen monitor. At the same time, the distance from the camera to the screen is recorded as ground truth. In the case of using a smartphone camera with a small image sensor, the moiré on the monitor may reduce depth accuracy. Therefore, we add the printed pattern images pasted on the screen to training datasets. In this paper, we use images taken for 100 points between 400 and 1200 mm with a focus distance of 600mm as training datasets.

2.2 Application for smartphone cameras

In recent years, high-end smartphones are equipped

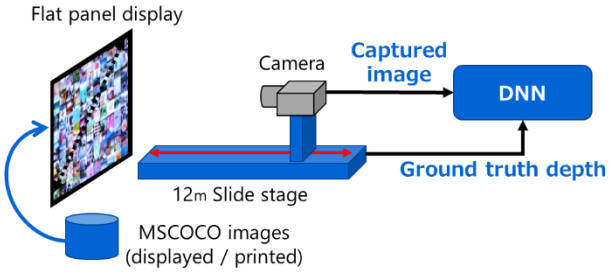
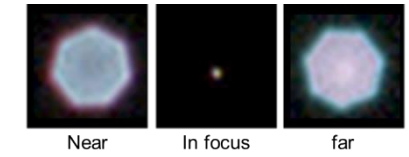
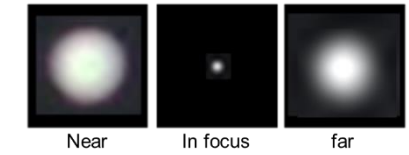


Fig. 3 Experimental system



(a) DSLR camera (Nikon D810 f=50mm f/4)



(b) smartphone camera (iPhone11Pro f=6mm f/2)

Fig. 4 Comparison of PSF images of a DSLR camera and a smartphone camera.

with a dual-camera system consisting of a wide-angle camera (Wide) with a focal length of about 4mm and a telephoto camera (Tele) with a focal length of about 6mm, or even a triple-camera system including an ultra-wide-angle camera. In this paper, we evaluate the performance of DDfAM using the Tele and Wide cameras of iPhone11Pro experimentally.

Fig. 4 shows PSF images taken at three locations: near-plane, in focus, and far-plane. The result for a DSLR camera (Nikon D810, f=50mm, f/4) and a Tele camera (f=6mm, f/2) are shown in Fig. 4(a) and (b) respectively. These PSF images were extracted from the center of the images taken of a monitor with only one pixel lit. While the image sensor of the DSLR camera is full-frame, the image sensor of iPhone11Pro Tele is 1/3.4" and has an area of about 1/72. This reduces the amount of light collected and lowers the signal-to-noise ratio in smartphone cameras. In addition, since the lenses for smartphone cameras are generally smaller and have shorter focal lengths, their resolution and contrast are more likely to be degraded compared to DSLR cameras. Reflecting this fact, while the colored-edge fringe by chromatic aberration is visible in the PSF image of a DSLR camera (Fig. 4(a)), the colored-edge fringe is unclear in the PSF image of a smartphone camera (Fig. 4(b)). Thus, it is more difficult to extract depth cues from smartphone images.

2.3 Fast inference network

Conventional A-Map analysis network is based on

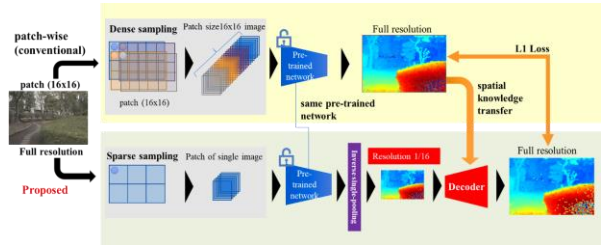


Fig. 5 Fast inference network with sparse sampling.

dense patch-wise processing, where the input images are converted into many patches with densely overlapping, and this causes slow inference. For practical use, a real-time analysis will be required for some applications. Then, we propose a network that speeds up the depth inference process as shown in Fig. 5. Instead of dense sampling, we introduce sparse sampling that converts the input images to patches without overlapping. Then an upsampling decoder converts the blurs estimated by the pre-trained patch-wise network to the depth maps. The structure of the decoder is based on MSGNet [5], a method of upsampling depth resolution, and there is a shortcut connection of the patch-wise features to the decoder similar to U-Net [6] suitable for the patch-processing network. For training the decoder, to restore the resolution lost in sparse sampling, we propose a transfer learning method that takes depth map by the dense patch-wise processing of the pre-trained network as supervision.

We use the following L1 loss function to train the decoder, fixing the encoder parameter $\hat{\theta}$,

$$L(\phi) = \sum_{I \in N} \left| \hat{o} - g_{\phi}\left(p^{-1}\left(f_{\hat{\theta}}(p(I))\right)\right)\right|, \qquad (3)$$

where $\hat{o}$ is the output by the pre-trained network f and p is the sparse sampling operator.

3 Results

3.1 Evaluation for smartphone cameras

We used the Tele (f=6mm, f/2) and Wide (f=4.25mm, f/1.8) cameras of iPhone11Pro and downsampled the captured images to 2016x1512 (1/2 of original size) for the depth analysis. For comparison, we also show stereo depth maps calculated by Semi Global Matching [1][8] to images of Tele and Wide cameras with about 15mm baseline. We note that this is not an equal comparison because DDfAM and stereo methods are based on different principles and the baseline of the stereo method is fixed.

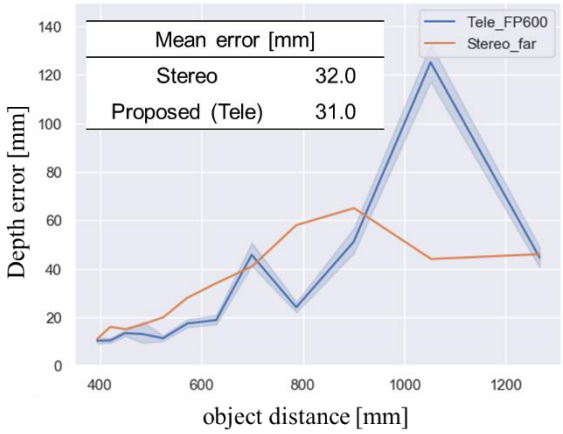


Fig. 6 Result of mean depth error.

For quantitative evaluation, we used our experimental system stated above. Moving the stage from the object distance of 400mm to 1200mm, we captured images of patterns displayed on the screen monitor and calculated the mean error between the estimated depth map for pattern images, which should ideally be uniform, and ground truth distances. The results of the depth error at each distance are shown in Fig. 6, along with the depth error averaged over the measured distance range. As the result shows, the mean errors of the proposed and stereo method are almost the same in this evaluation.

The qualitative results for indoor and outdoor scenes are shown in Fig. 7. As for the stereo results, depth can be estimated in the region where feature points can be matched, however, errors are observed in textures in the same direction as the disparity (e.g., horizontal lines or repeating patterns). Moreover, there are some areas where stereo matching fails despite the presence of texture in Fig. 7(b), since iPhone11Pro's optical image stabilization may cause a calibration shift when the camera is tilted. In contrast, the proposed method for the Tele camera can estimate valid depth from near to far, independent of the pattern and the tilt of the camera.

3.2 Evaluation for inference speed

For evaluation of the fast inference network, we used a DSLR camera (Nikon D810, f=50mm, f/4), whose depth performance had already been verified. The qualitative results of depth analysis for an outdoor scene are shown in Fig. 8. The input image is shown in Fig. 8(a). While a context-based monocular depth estimation method [2] cannot estimate the correct depth map (Fig.

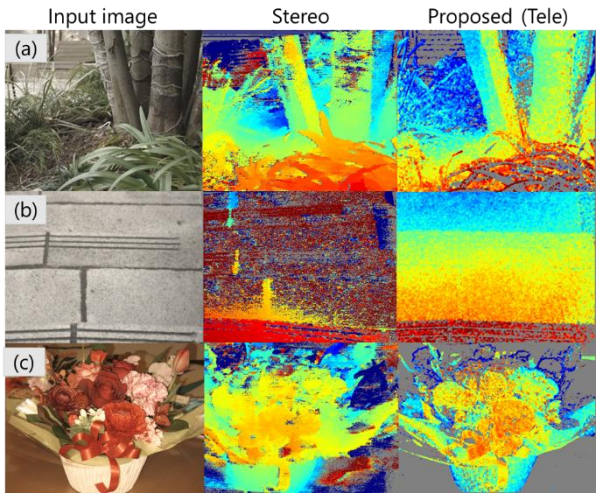


Fig. 7 Qualitativte results for indoor and outdoor scenes. Near objects are displayed in red and far objects in blue. Gray indicates there is no depth cue.

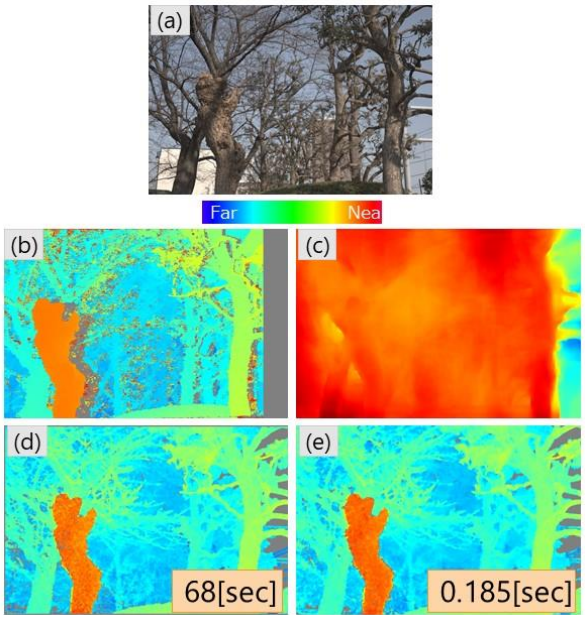


Fig. 8 Comparisons of qualitative results for various methods. (a) An input image. (b) Depth of stereo method[1][8] as a reference. (c) Depth of context-based method [2]. (d) Depth of patch-wise network. (e) Depth of proposed network.

8(c)) in this scene without contextual information, the proposed method (sparse sampling) estimates the valid depth map (Fig. 8(e)), almost equivalent to the conventional method (Fig. 8(d)) and stereo method (Fig. 8(b)) with two DSLR cameras at 250mm baseline. On the other hand, the inference time for an image with the resolution 1845x1232 is reduced from 68 seconds to 0.185 seconds using the proposed method by only 4GB GPU memory, which is 360 times faster than the conventional method.

4 Conclusions

In this paper, we propose a depth measurement method using a single-shot image of a smartphone camera. The evaluation results show that our method can analyze depth robustly to the scene in both indoor and outdoor images, even when using images of smartphone cameras, which have fewer depth cues than DSLR cameras. This allows us to perform depth measurement robustly to the environment at a low cost, without additional devices. In addition, we propose a network structure that speeds up the depth inference process, and we confirm this network can infer 300 times faster than the conventional network with patch-based architecture, demonstrating the potential of real-time processing. In the future, we will continue to improve the accuracy and study the implementations of our products and services.

References

[1] H. Hirschmuller, “Accurate and efficient stereo processing by semi-global matching and mutual information,” 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), Vol. 2, pp. 807–814 (2005).

[2] Z. Li and N. Snavely, “Megadepth: Learning single-view depth prediction from internet photos,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2041–2050 (2018).

[3] M. Kashiwagi, N. Mishima, T. Kozakaya, and S. Hiura, “Deep Depth from Aberration Map,” In Proceedings of the IEEE International Conference on Computer Vision, pp. 4070–4079 (2019).

[4] H. Tang, S. Cohen, B. Price, S. Schiller, and K. N. Kutulakos, “Depth From Defocus in the Wild,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2740-2748 (2017).

[5] T. Hui, C. C. Loy, and X. Tang, “Depth Map Super-Resolution by Deep Multi-scale Guidance,” In European conference on computer vision, pp. 353–369 (2016).

[6] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” In International Conference on Medical image computing and computer-assisted intervention, pp. 234–241 (2015).

[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 4, pp. 652–663 (2017).

[8] OpenCV. https://opencv.org/.