Unsupervised Depth and Bokeh Learning from Natural Images Using Aperture Rendering Generative Adversarial Networks

Takuhiro Kaneko

Abstract

Humans can estimate the depth and bokeh effects from a two-dimensional (2D) image on the basis of their experience and knowledge. However, computers have difficulty in doing this because they logically cannot have such experience and expertise. To overcome this limitation, a novel deep generative model called aperture rendering generative adversarial network (AR-GAN) is discussed. AR-GAN makes it possible to control the bokeh effects on the basis of the predicted depth by incorporating an optical constraint of a camera aperture into a GAN. During training, AR-GAN requires only standard 2D images (such as those on the web) and does not require 3D data such as depth and bokeh information. Therefore, it can alleviate the application boundaries that come from the difficulty in collecting 3D data. This technology is expected to enable the exploration of new possibilities in studies on 3D understanding.

Keywords: generative adversarial networks, unsupervised learning, depth and bokeh

1. Introduction

Humans can estimate the depth and bokeh (shallow depth-of-field (DoF)) effects from a two-dimensional (2D) image on the basis of their experience and knowledge. However, computers have difficulty in doing this because they logically cannot have such experience and expertise. However, considering that, in the future, robots will be able to move around us and the real and virtual worlds will be integrated, it will be necessary to create computers that can act or present information on the basis of 3D data such as depth and bokeh information. Considering that a photo is one of the most frequently used forms of data for recording or saving information, understanding 3D information from 2D images will be valuable for various 3D-based applications to reduce installation cost because it enables using easily available 2D images as input.

Three-dimensional understanding from 2D images has been actively studied in computer vision and machine learning. A successful approach is to learn the 3D predictor using direct or photometric-driven supervision after collecting pairs of 2D and 3D data [1] or sets of multi-view images [2]. This approach demonstrates good prediction accuracy due to the ease of training. However, collecting pairs of 2D and 3D data or sets of multi-view images is not always easy or practical because they require special devices such as a depth sensor or stereo camera.

To reduce the data-collection costs, our team is investigating a fully unsupervised approach for learning 3D representations only from images without any additional supervision. In the study published in the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021) [3], I introduced a new deep generative model called aperture rendering generative adversarial network (AR-GAN), which can learn depth and bokeh effects from standard 2D images such as those on the web. Focus cues that are

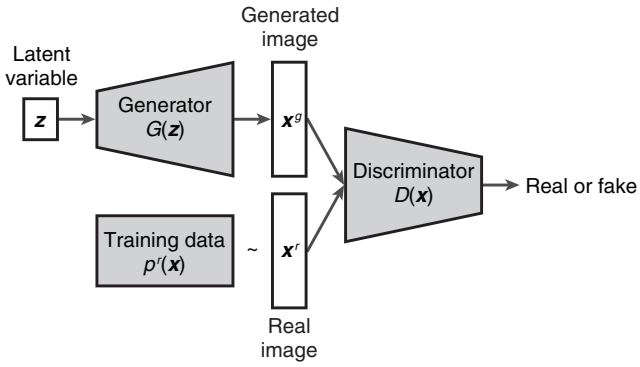


Fig. 1. Architecture of GAN.

inherent in photos but had not been actively studied in previous deep generative models were considered. On the basis of this consideration, our team developed AR-GAN to incorporate aperture rendering (particularly light field aperture rendering [4]) into a GAN [5] (a variant of deep generative models). This configuration allows synthesizing a bokeh image on the basis of the predicted depth and all-in-focus (deep DoF) image using a camera with an optical constraint on the light field.

The rest of this article is organized as follows. In Section 2, I first review two previous studies on which AR-GAN is based: GAN [5] and light field aperture rendering [4]. In Section 3, I explain AR-GAN, which is the main topic of this article. In Section 4, I discuss the experiments on the effectiveness of AR-GAN. In Section 5, I present concluding remarks and areas for future research.

2. Preliminaries

2.1 GAN

GANs [5] can mimic training data without defining their distribution explicitly. This property enables GANs to be applied to various tasks and applications in diverse fields.

As shown in **Fig. 1**, a GAN is composed of two neural networks: a generator $G(\boldsymbol{z})$ and discriminator $D(\boldsymbol{x})$. These two networks are optimized through a two-player min-max game using an objective function L_{GAN}:

$$L_{\mathrm{GAN}} = \mathbb{E}_{\boldsymbol{x}^r \sim p^r(\boldsymbol{x})}[\log D(\boldsymbol{x}^r)] + \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z})}\left[\log\left(1 - D(G(\boldsymbol{z}))\right)\right],$$

where, given a latent variable $\boldsymbol{z} \sim p(\boldsymbol{z})$, a $G(\boldsymbol{z})$ attempts to generate an image $\boldsymbol{x}^g = G(\boldsymbol{z})$ that can deceive a $D(\boldsymbol{x})$ by minimizing L_{GAN}. By contrast, the $D(\boldsymbol{x})$ attempts to distinguish a generated image $\boldsymbol{x}^g$ from a real image $\boldsymbol{x}^r \sim p^r(\boldsymbol{x})$ by maximizing L_{GAN}. Superscripts r and g denote the real and generated data, respectively. Through this adversarial training, a generative distribution $p^g(\boldsymbol{x})$ reaches close to a real distribution $p^r(\boldsymbol{x})$.

2.2 Light field aperture rendering

Light field aperture rendering [4] is a module that simulates an optical phenomenon (particularly bokeh) on a camera aperture in a differentiable manner. Note that such a differentiable property is necessary for deep neural networks (DNNs), such as a $G(\boldsymbol{z})$ (discussed in Section 2.1), to update the parameters through the backpropagation commonly used for DNNs.

More concretely, as shown in **Fig. 2**, the rendering provides an aperture renderer $R(\boldsymbol{x}_d, \boldsymbol{d})$ that synthesizes a bokeh image $\boldsymbol{x}_s(\boldsymbol{r})$ from an all-in-focus image $\boldsymbol{x}_d(\boldsymbol{r})$ and depth map $\boldsymbol{d}(\boldsymbol{r})$. Here, $\boldsymbol{r}$ indicates the spatial coordinates of the light field on the image plane.

I explain the details in a step-by-step manner. First, a depth map $\boldsymbol{d}(\boldsymbol{r})$ is expanded into a depth map for each view in the light field, i.e., $\boldsymbol{m}(\boldsymbol{r}, \boldsymbol{u})$, using a neural network T:

$$\boldsymbol{m}(\boldsymbol{r}, \boldsymbol{u}) = T(\boldsymbol{d}(\boldsymbol{r})),$$

where $\boldsymbol{u}$ indicates the angular coordinates of the light field on the aperture plane. Subsequently, an all-in-focus image $\boldsymbol{x}_d(\boldsymbol{r})$ is warped into an image for each view of the light field, i.e., $\boldsymbol{l}(\boldsymbol{r}, \boldsymbol{u})$, using the predicted $\boldsymbol{m}(\boldsymbol{r}, \boldsymbol{u})$:

$$\boldsymbol{l}(\boldsymbol{r}, \boldsymbol{u}) = \boldsymbol{x}_d(\boldsymbol{r} + \boldsymbol{u}m(\boldsymbol{r}, \boldsymbol{u})).$$

From this formulation, the left-side images in the light field (5 × 5 images in Fig. 2) represent images

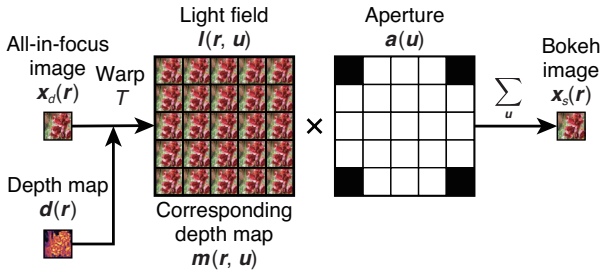


Fig. 2. Processing flow of light field aperture rendering.

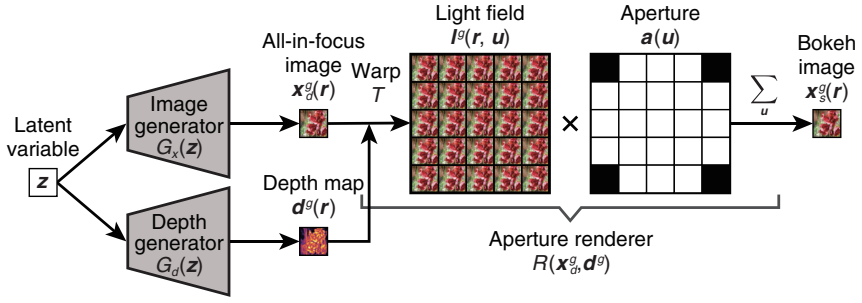


Fig. 3. Processing flow of AR-GAN generator.

when viewing objects from the left side, and the right-side images represent vice versa.

Finally, the $l(r, u)$ is integrated using an aperture $a(u)$ (an indicator that represents the disk-shaped camera aperture and takes ones for views within the aperture (indicated with white in Fig. 2) and zeroes otherwise (indicated with black in Fig. 2)) to render a bokeh image $x_s(r)$:

$$x_s(r) = \sum_u a(u)\, l(r, u).$$

When an object is on the focal plane, the object's position is consistent regardless of the $l(r, u)$. Therefore, no bokeh occurs when the $l(r, u)$ is integrated by the above equation. By contrast, when an object is separate from the focal plane, the object's position varies depending on the $l(r, u)$. Thus, bokeh occurs in this case. Hereafter, r and u are omitted for simplicity except in necessary cases.

3. AR-GAN

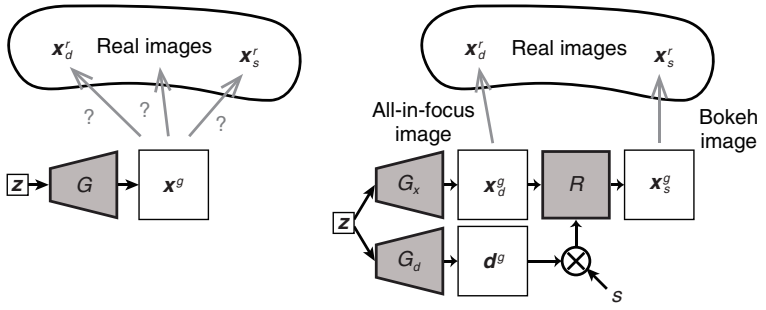
3.1 Problem statement

The problem statement is clarified before explaining the details of AR-GAN. As described in Section 1, AR-GAN is used to learn depth and bokeh effects only from images without additional supervision. In this setting, it is not easy to construct a conditional generator that directly predicts the depth or bokeh effects from an image due to the absence of pairs of 2D and 3D data or sets of multi-view images. Therefore, as an alternative, the aim is to learn an unconditional generator that can generate a tuple of an all-in-focus image x_d^g, depth map d^g, and bokeh image x_s^g from a latent variable z.

AR-GAN uses focus cues as a clue for addressing this challenge. When the training images are highly biased in terms of bokeh effects (e.g., all training images are all-in-focus), it is difficult to gain focus cues from the images. Therefore, it is assumed with AR-GAN that the training dataset includes various bokeh images. Note that this assumption does not mean that the training dataset contains sets of different bokeh images for each instance. Under this assumption, AR-GAN learns the generator in a wisdom of crowds approach.

3.2 Model architecture

The processing flow of the AR-GAN generator is presented in **Fig. 3**. Given a latent variable z, the AR-GAN generator generates an all-in-focus image



(a) Standard GAN learning (b) DoF mixture learning

Fig. 4. Comparison between standard GAN learning and DoF mixture learning.

$\boldsymbol{x}_d^g = G_{\boldsymbol{x}}(\boldsymbol{z})$ and depth map $\boldsymbol{d}^g = G_{\boldsymbol{d}}(\boldsymbol{z})$ using an all-in-focus image generator $G_{\boldsymbol{x}}(\boldsymbol{z})$ and a depth generator $G_{\boldsymbol{d}}(\boldsymbol{z})$, respectively. Subsequently, the aperture renderer $R(\boldsymbol{x}_d, \boldsymbol{d})$ (explained in Section 2.2) synthesizes a bokeh image $\boldsymbol{x}_s^g = R(\boldsymbol{x}_d^g, \boldsymbol{d}^g)$. Using this configuration, AR-GAN makes it possible to generate a tuple of an all-in-focus image $\boldsymbol{x}_d^g$, depth map $\boldsymbol{d}^{\mathrm{g}}$, and bokeh image $\boldsymbol{x}_s^g$ using a camera with an optical constraint on the light field.

3.3 Training method

As shown in Fig. 1, a typical GAN applies a $D(\boldsymbol{x})$ to the final output of the $G(\boldsymbol{z})$ (i.e., $\boldsymbol{x}_s^g$ in the case of the AR-GAN generator). However, in the AR-GAN generator, three modules, i.e., $G_{\boldsymbol{x}}(\boldsymbol{z})$, $G_{\boldsymbol{d}}(\boldsymbol{z})$, and $R(\boldsymbol{x}_d^g, \boldsymbol{d}^g)$ are trainable. Therefore, they compete for roles if there is no constraint. For example, they can fall into an extreme solution (e.g., $R(\boldsymbol{x}_d^g, \boldsymbol{d}^g)$ learns strong bokeh effects and $G_{\boldsymbol{x}}(\boldsymbol{z})$ learns over-blurred images).

To alleviate this problem, AR-GAN is trained using DoF mixture learning. **Figure 4** illustrates the comparison between the standard GAN learning and DoF mixture learning. In the standard GAN learning shown in Fig. 4(a), the $G(\boldsymbol{z})$ attempts to cover the overall real image distribution using generated images without any constraint. Consequently, it cannot determine to make a generated image $\boldsymbol{x}^{\mathrm{g}}$ close to a real all-in-focus image $\boldsymbol{x}_d^r$ or a real bokeh image $\boldsymbol{x}_s^r$ (indicated with question marks "?" in Fig. 4(a)).

By contrast, as shown in Fig. 4(b), in DoF mixture learning, the AR-GAN generator attempts to represent the real image distribution using generated images, the bokeh degrees of which are adjusted by a scale factor s. More concretely, the GAN objective function presented in Section 2.1 is rewritten as follows:

$$L_{\mathrm{AR-GAN}} = \mathbb{E}_{\boldsymbol{x}^r \sim p^r(\boldsymbol{x})}[\log D(\boldsymbol{x}^r)] + \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z}), s \sim p(s)} \left[\log\left(1 - D\left(R(G_{\boldsymbol{x}}(\boldsymbol{z}), sG_{\boldsymbol{d}}(\boldsymbol{z}))\right)\right)\right],$$

where $s \in [0, 1]$; when $s = 0$, an all-in-focus image $\boldsymbol{x}_d^g$ is generated, whereas when $s = 1$, a bokeh image $\boldsymbol{x}_s^g$ is rendered. Intuitively, the aperture renderer $R(\boldsymbol{x}_d^g, \boldsymbol{d}^g)$, which has an optical constraint on the light field, functions as a bokeh image prior. This prior encourages a generated all-in-focus image $\boldsymbol{x}_d^g$ to approximate a real all-in-focus image $\boldsymbol{x}_d^r$ (indicated by the "All-in-focus image" in Fig. 4(b)) and promotes a generated bokeh image $\boldsymbol{x}_s^g$ to mimic a real bokeh image $\boldsymbol{x}_s^r$ (indicated by the "Bokeh image" in Fig. 4(b)). Consequently, $\boldsymbol{d}^g$, which connects $\boldsymbol{x}_d^g$ and $\boldsymbol{x}_s^g$, is also optimized. In this manner, the DoF mixture learning allows optimizing $G_{\boldsymbol{x}}(\boldsymbol{z})$, $G_{\boldsymbol{d}}(\boldsymbol{z})$, and $R(\boldsymbol{x}_d^g, \boldsymbol{d}^g)$ together under an optical constraint.

A remaining challenge specific to unsupervised depth and bokeh learning is the difficulty in distinguishing whether blur occurs ahead of or behind the focal plane. For this challenge, on the basis of the observation that the focused image tends to be placed at the center of a photo, AR-GAN uses the center focus prior, which encourages the center to be focused while promoting the surroundings to be behind the focal plane. In practice, this prior is only used at the beginning of training to determine the learning direction.

4. Experiments

4.1 Image and depth synthesis

The previous AR-GAN study [3] demonstrated the utility of AR-GAN using various natural image datasets, including flower (Oxford Flowers [6]), bird (CUB-200-2011 [7]), and face (FFHQ [8]) datasets.

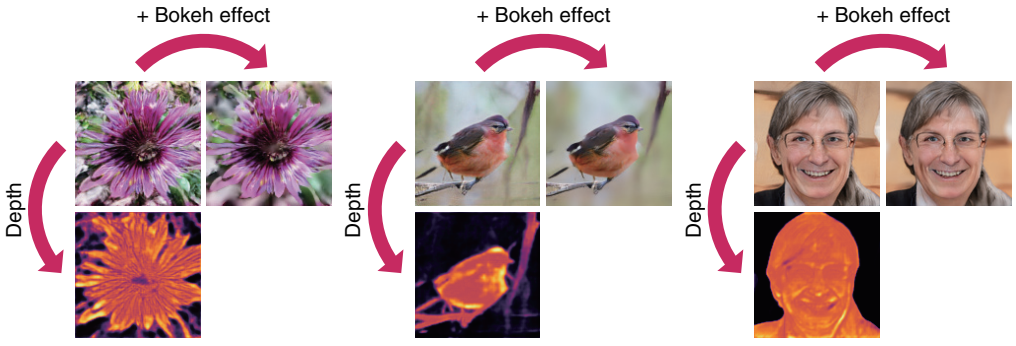


Fig. 5. Examples of generated images and depth maps.

Fig. 6. Examples of bokeh rendering and depth prediction.

The implementation details are omitted because of space limitations. See that AR-GAN study [3] if interested in the implementation details.

Figure 5 shows examples of generated images and depth maps. AR-GAN succeeds in generating a tuple of an all-in-focus image (upper left), bokeh image (upper right), and depth map (lower left) in every setting. For example, in Fig. 5(a), the background is blurred while the foreground is unchanged in bokeh conversion (the conversion from the upper left to upper right). In depth prediction (the transformation from the upper left to lower left), the depth map (lower left) corresponding to the image (upper left) is successfully predicted. A light color indicates the foreground while a dark color indicates the background. Recall that the training data are only 2D images, and depth and bokeh effects are not provided as supervision. In this manner, learning depth and bokeh effects only from 2D images is the main strength of AR-GAN.

4.2 Application to bokeh rendering and depth prediction

As discussed in Section 3.1, AR-GAN learns an unconditional generator that generates a tuple of an all-in-focus image $\boldsymbol{x}_d^g$, depth map $\boldsymbol{d}^g$, and bokeh image $\boldsymbol{x}_s^g$ from a latent variable $\boldsymbol{z}$. Therefore, it cannot be directly used to convert a given image to the bokeh image or depth. However, AR-GAN can generate sets of all-in-focus and bokeh images or sets of all-in-focus images and depth maps artificially and abundantly by randomly changing the latent variable. By using these data, we can learn a bokeh renderer (i.e., a converter that converts an all-in-focus image to a bokeh image) and depth predictor (i.e., a predictor that predicts a depth map from an image) in a supervised manner.

Figure 6 shows example results obtained with the bokeh renderer and depth predictor mentioned above. A photo I took was used as an input (Fig. 6(a)). The

bokeh renderer synthesizes a bokeh image (Fig. 6(b)), and the depth predictor predicts a depth map from the input image (Fig. 6(c)). Similar to the results in Fig. 5, the background is blurred while the foreground remains unchanged in the bokeh conversion (the conversion from (a) to (b)), and the depth map corresponding to the input image is predicted in the depth prediction (the transformation from (a) to (c)).

Note that the data required for training the bokeh renderer and depth predictor are only the data generated by AR-GAN, and no additional data are needed. That is to say, in this setting, we can learn a bokeh renderer and depth predictor in a fully unsupervised manner, similar to AR-GAN. This is a strength of an AR-GAN-based approach.

5. Conclusion and future work

This article explained AR-GAN, which is a new deep generative model enabling the unsupervised learning of depth and bokeh effects only from natural images. Since we live in the 3D world, human-oriented computers are expected to understand the 3D world. For this challenge, AR-GAN is effective because it can eliminate the requirement of 3D data during training. AR-GAN is expected to enable the exploration of new possibilities in studies on 3D understanding.

AR-GAN will also be useful for many applications in various fields such as environmental understanding in robotics, content creation in advertisements, and photo editing in entertainment. For example, AR-GAN can learn a data-driven model from collected images. Using this strength, a data-driven bokeh renderer reflecting a famous photographer can be constructed if we can collect his/her photos. Thus, AR-GAN can be used to obtain more natural and impactful bokeh images and enrich the functionality of photo-editing applications (e.g., smartphone applications for social media).

Future work includes further improvement of depth and bokeh accuracy since unsupervised learning of depth and bokeh effects is an ill-posed problem, and there is room for improvement. Our team is tackling this challenge, and my latest paper [9] has been accepted to CVPR 2022. Due to space limitations, details of this are omitted. Please check my latest paper [9] if interested in the details.

References

[1] D. Eigen, C. Puhrsch, and R. Fergus, "Depth Map Prediction from a Single Image Using a Multi-scale Deep Network," Proc. of the 27th International Conference on Neural Information Processing Systems (NIPS 2014), pp. 2366–2374, Montreal, Canada, Dec. 2014.

[2] R. Garg, V. Kumar B. G., G. Carneiro, and I. Reid, "Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue," Proc. of the 14th European Conference on Computer Vision (ECCV 2016), pp. 740–756, Amsterdam, The Netherlands, Oct. 2016.

[3] T. Kaneko, "Unsupervised Learning of Depth and Depth-of-field Effect from Natural Images with Aperture Rendering Generative Adversarial Networks," Proc. of the 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), pp. 15679–15688, Virtual, June 2021.

[4] P. P. Srinivasan, R. Garg, N. Wadhwa, R. Ng, and J. T. Barron, "Aperture Supervision for Monocular Depth Estimation," Proc. of CVPR 2018, pp. 6393–6401, Salt Lake City, USA, June 2018.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," Proc. of NIPS 2014, pp. 2672–2680, Montreal, Canada, Dec. 2014.

[6] M. Nilsback and A. Zisserman, "Automated Flower Classification over a Large Number of Classes," Proc. of the Sixth Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP 2008), pp. 722–729, Bhubaneswar, India, Dec. 2008.

[7] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[8] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," Proc. of CVPR 2019, pp. 4401–4410, Long Beach, USA, June 2019.

[9] T. Kaneko, "AR-NeRF: Unsupervised Learning of Depth and Defocus Effects from Natural Images with Aperture Rendering Neural Radiance Fields," CVPR 2022, pp. 18387–18397, New Orleans, USA, June 2022.

Takuhiro Kaneko
Distinguished Researcher, Media Information Laboratory, NTT Communication Science Laboratories.

He received a B.E., M.S., and Ph.D. from the University of Tokyo in 2012, 2014, and 2020. He joined NTT Communication Science Laboratories in 2014 and has been a distinguished researcher since 2020. His research interests include computer vision, signal processing, and machine learning. He is currently working on image synthesis, speech synthesis, and voice conversion using deep generative models such as GANs. He received the Hatakeyama Award from the Japan Society of Mechanical Engineers, the ICPR 2012 Best Student Paper Award in the 21st International Conference on Pattern Recognition (ICPR 2012), and the Dean's Award for Best Doctoral Thesis from the University of Tokyo. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Information Processing Society of Japan (IPSJ), and the Acoustic Society of Japan (ASJ).