Learning 3D Information from 2D Images Using Aperture Rendering Generative Adversarial Networks toward Developing a Computer that "Understands the 3D World"

Takuhiro Kaneko
Distinguished Researcher, NTT Communication Science Laboratories

Abstract

When people look at photos, they can estimate three-dimensional (3D) information, such as depth, from their experience and knowledge, but computers have difficulty in doing so because they cannot have such experience and knowledge. We spoke to Takuhiro Kaneko, a distinguished researcher who developed a novel deep learning model that can learn 3D information from standard 2D images such as those on the web.

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

Learning of depth and bokeh effects from natural images with aperture rendering generative adversarial networks

—What is research on "learning of depth and bokeh effects from natural images" about?

It is difficult to record the three-dimensional (3D) world in which we live in as it is. For this reason, it is common to record and store two-dimensional (2D) images such as photographs instead of 3D information.

When people look at photos, they can estimate 3D information, such as depth, from their previous experience and knowledge. However, computers have difficulty in doing so because they do not have such experience or knowledge. In the future, when we think about scenarios where robots support our lives, it will become essential for computers to understand the 3D world. The easiest way for computers to learn is to provide a large number of pairs of 2D images and 3D information as training data. Learning would be easy because they know the correct answer. However, this method requires special devices such as depth sensors and stereo cameras and is costly.

For this reason, in this research, we created a deep learning model that can learn 3D information from standard 2D images such as those taken with ordinary

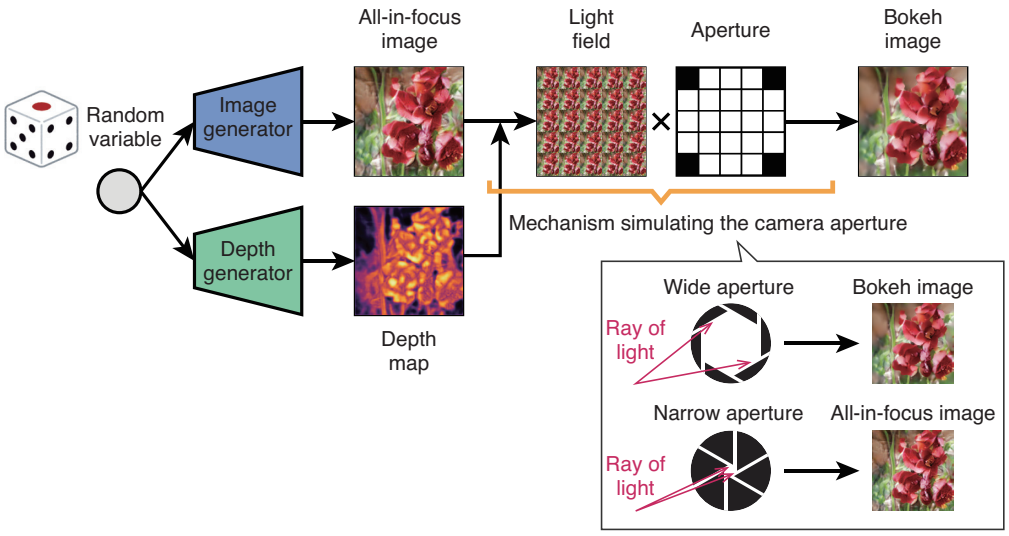

Fig. 1. Process flow in the AR-GAN generator.

cameras and those on the web. When we look at the photos, we see that the focus is usually on the object, and the background is often blurry. Using these defocus cues as clues, we carry out learning of 3D information, especially depth and bokeh effects. In other words, if the problem of projecting 3D information onto 2D images through a camera is the forward problem, for this research, the goal is to solve the inverse problem. This problem is challenging because it is the so-called "ill-posed problem" of estimating 3D information from only 2D images that lack various information.

—Specifically, through what mechanisms are you carrying out learning?

This research is based on a technology called a generative adversarial network (GAN). GAN is a so-called "unsupervised learning model," which does not require pre-determined correct answers, and consists of two neural networks; namely, a generator and discriminator. The generator generates a "fake image" from a given random variable. On the other hand, the discriminator distinguishes two types of images, "real images" and the "fake images" generated by the generator. Because they work adversarially, wherein the generator tries to trick the discriminator while the discriminator tries to identify the fake images precisely, learning can be carried out while competing, and as a result, the generator can produce realistic images.

Since GAN is a specialized technology for producing 2D images, it has no connection with the 3D world. We, therefore, proposed the "aperture rendering GAN (AR-GAN)," which incorporates the optical properties of the camera into the GAN. "Aperture" refers to the aperture of the camera. By incorporating optical constraints due to the aperture of the camera when projecting the 3D world into 2D images, the generator can learn by associating 2D images, depth maps, and bokeh effects.

Figure 1 shows the process flow in the AR-GAN generator. The "image generator" is also included in the GAN and generates an all-in-focus image when given a random variable. The depth generator generates the depth map that is paired with the image, and is unique to AR-GAN. The system then uses these pairs of generated data to enable a mechanism simulating the camera aperture. The light field consists of 25 images, which represent what the image looks like in the aperture of the camera. In the center, the image created by light coming in through the center of the aperture is displayed. When the object is offset from the center to the right, an image of the object viewed from the right part of the aperture is displayed; and when the object is offset to the top of the aperture, an image of the object viewed from the top part of the aperture is displayed.

It may be easier to imagine the mechanism by moving your face up, down, left, and right while looking at the object in front of your eyes. Objects with focus are clearly visible because they are in the same

position and do not move when you move your face, but objects that are farther away move more when you move your face. Therefore, summation of these images produces images with bokeh effects. This is how the mechanism enables synthesizing a bokeh image on the basis of the predicted depth map and all-in-focus image using a camera with an optical constraint on the light field.

—How is the progress of the research and what are the challenges going forward?

Currently, we are able to generate flower images, bird images, and human face images. Learning can take days, but once it is done, the system can generate in a few seconds bokeh images that cannot be distinguished from the real images. At present, I think that if we narrow down the types of objects, we would be able to generate reasonable images.

As for the challenges, one is that depending on the type of object, some are easier to learn while others are not as amenable to learning. For example, for human face images, the shape and position of the parts are fixed to some degree, making it easy to learn them. On the other hand, images with many variations, such as those of animals taken from different viewpoints and distances, are difficult to learn.

Also, the larger the image size, the more details need to be synthesized, and the more processing time it takes, making learning more difficult. I think there will be various issues that will emerge as we expand the scope of applications in the future. As expected, the difficulties inherent with ill-posed problems remain.

Developing a computer that understands the 3D world

—What will this technology enable in the future?

Our mission as researchers is to "develop computers that are highly compatible with humans." For this reason, it is essential to understand the 3D world; but the cost of data collection is likely to be a barrier to applying the technology to a wider range of fields. In that regard, I think this research is beneficial because it is excellent in terms of data collection. In the future, we will be able to create robots that can move freely around the 3D world, build 3D worlds without dissonance in virtual space, and develop tools to create 3D objects in virtual space.

Another advantage is that optimized models can be built as long as we can collect 2D images. For example, if you gather photos taken by a well-known photographer, you can build a model that learns the bokeh effects unique to that photographer. Currently, communication using images such as through social media has become very common. If we can easily add the bokeh effects, it may become easier to create more attractive photos. I think this is particularly useful in the three areas of robotics, content generation, and entertainment.

—What are the future prospects and initiatives on collaborations with other fields?

Since it is basic research, it is difficult to set specific targets for practical use at this time; but we plan to continue to improve performance by increasing accuracy and resolution. We have focused on the camera aperture, but it is interesting to note that adding physical constraints enables creating a more reliable computer.

One of the key technical areas of the Innovative Optical and Wireless Network (IOWN) initiative is "Digital Twin Computing," where computations are performed using digital twins of various industries, things, and humans. In order to merge the real world with the virtual world, it is necessary for computers to properly understand the real world, so I believe that this technology can also contribute to these areas.

I feel that the importance of integrating media generation technology with various fields is increasing as the technology matures. Currently, we are conducting research that combines computer science, such as computer vision and machine learning, with physics, such as optics. Going forward, we will continue to

focus our efforts on collaborations with people from other fields, such as computer graphics for image creation and photonics for photography, as well as on cross-discipline implementation.

—Could you give a message to young researchers and future business partners?

NTT laboratories are conducting extensive research ranging from basic research to applications, and in particular, our laboratory, NTT Communication Science Laboratories, is conducting research on how to improve communication between humans and between humans and machines. In the past, we have often been limited to basic research, but recently, the distance between basic research and applied research has been narrowing, and I feel that the opportunities for engaging in research while considering real-world problems are increasing. We are also seeing more and more cases of the technology presented at international conferences being embedded in applications and deployed as a service on the web. I think it is interesting to see the output, for example, by creating a solution that actually converts voice, rather than stopping at tinkering with formulas.

There are limitations to doing research individually. Our laboratory is also actively collaborating with universities and accepting interns, so I hope that we can continue to actively collaborate particularly with students and young researchers who want to create something or change something.

As for business partners, since we have been able to create interesting ideas and technologies from the research side, we need their help in linking these ideas and technologies to services. On the other hand, we can get ideas for research by receiving feedback from people in the service field who have profound knowledge of real problems. So, going forward, we hope to continue to actively collaborate also with business partners.

■ Interviewee profile

Takuhiro Kaneko received his M.E. degree from the University of Tokyo in 2014. He joined NTT the same year as a member of NTT Communication Science Laboratories. In 2020, he completed his Ph.D. at the University of Tokyo. He has been a distinguished researcher at NTT Communication Science Laboratories since 2020. He is engaged in research on computer vision, signal processing, machine learning, and deep learning, particularly dealing with image synthesis, speech synthesis, and voice conversion. He received the Hatakeyama Award from the Japan Society of Mechanical Engineers, the ICPR 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. For details: https://www.kecl.ntt.co.jp/people/kaneko.takuhiro/