Technique enables real-time rendering of scenes in 3D

December 7 2021, by Adam Zewe

$$\Phi^{3D} : \mathbb{R}^3 \to \mathbb{R}^n$$

$$c(r) = \blacksquare$$
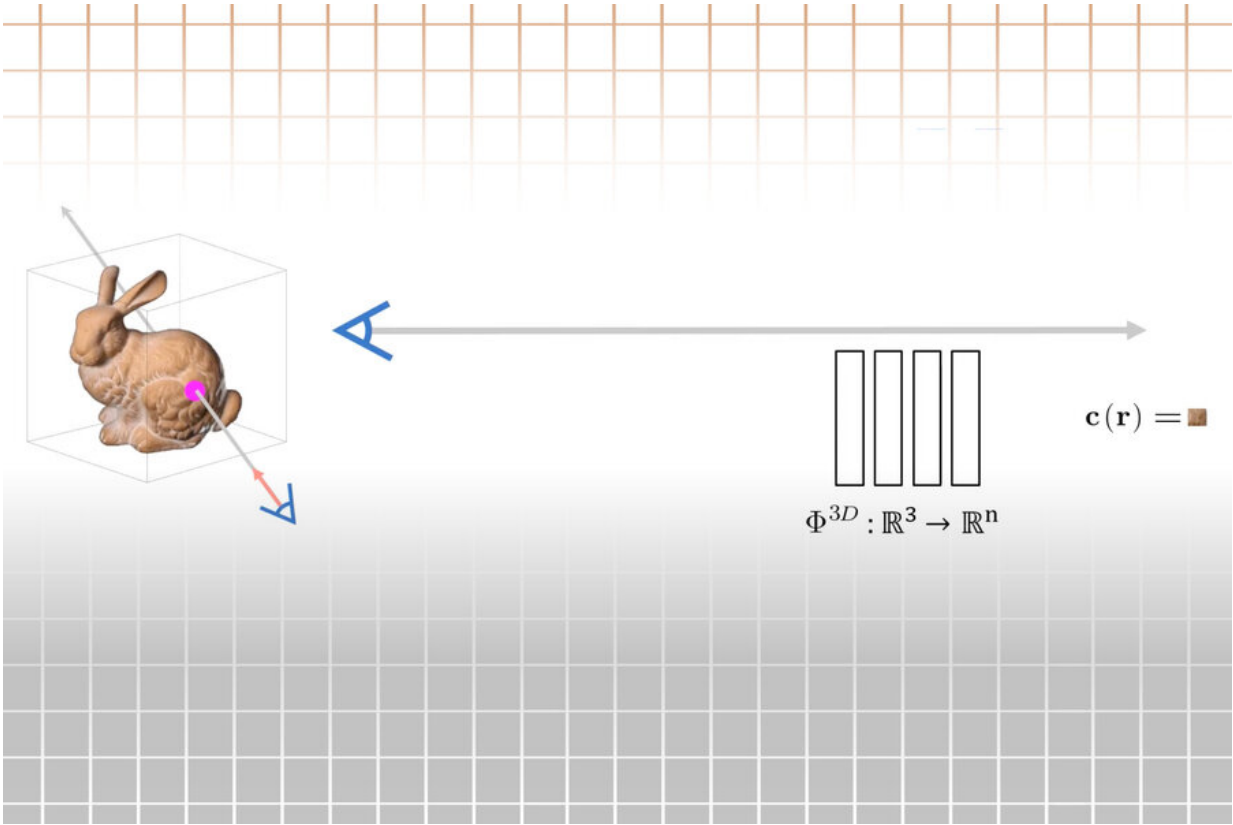
To represent a 3D scene from a 2D image, a light field network encodes the 360-degree light field of the 3D scene into a neural network that directly maps each camera ray to the color observed by that ray. Credit: Massachusetts Institute of Technology

Humans are pretty good at looking at a single two-dimensional image

and understanding the full three-dimensional scene that it captures. Artificial intelligence agents are not.

Yet a machine that needs to interact with objects in the world—like a robot designed to harvest crops or assist with surgery—must be able to infer properties about a 3D scene from observations of the 2D images it's trained on.

While scientists have had success using neural networks to infer representations of 3D scenes from images, these machine learning methods aren't fast enough to make them feasible for many real-world applications.

A new technique demonstrated by researchers at MIT and elsewhere is able to represent 3D scenes from images about 15,000 times faster than some existing models.

The method represents a scene as a 360-degree light field, which is a function that describes all the light rays in a 3D space, flowing through every point and in every direction. The light field is encoded into a neural network, which enables faster rendering of the underlying 3D scene from an image.

The light-field networks (LFNs) the researchers developed can reconstruct a light field after only a single observation of an image, and they are able to render 3D scenes at real-time frame rates.

"The big promise of these neural scene representations, at the end of the day, is to use them in vision tasks. I give you an image and from that image you create a representation of the scene, and then everything you want to reason about you do in the space of that 3D scene," says Vincent Sitzmann, a postdoc in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and co-lead author of the paper.

Sitzmann wrote the paper with co-lead author Semon Rezchikov, a postdoc at Harvard University; William T. Freeman, the Thomas and Gerd Perkins Professor of Electrical Engineering and Computer Science and a member of CSAIL; Joshua B. Tenenbaum, a professor of computational cognitive science in the Department of Brain and Cognitive Sciences and a member of CSAIL; and senior author Frédo Durand, a professor of electrical engineering and computer science and a member of CSAIL. The research will be presented at the Conference on Neural Information Processing Systems this month.

Mapping rays

In computer vision and computer graphics, rendering a 3D scene from an image involves mapping thousands or possibly millions of camera rays. Think of camera rays like laser beams shooting out from a camera lens and striking each pixel in an image, one ray per pixel. These computer models must determine the color of the pixel struck by each camera ray.

Many current methods accomplish this by taking hundreds of samples along the length of each camera ray as it moves through space, which is a computationally expensive process that can lead to slow rendering.

Instead, an LFN learns to represent the light field of a 3D scene and then directly maps each camera ray in the light field to the color that is observed by that ray. An LFN leverages the unique properties of light fields, which enable the rendering of a ray after only a single evaluation, so the LFN doesn't need to stop along the length of a ray to run calculations.
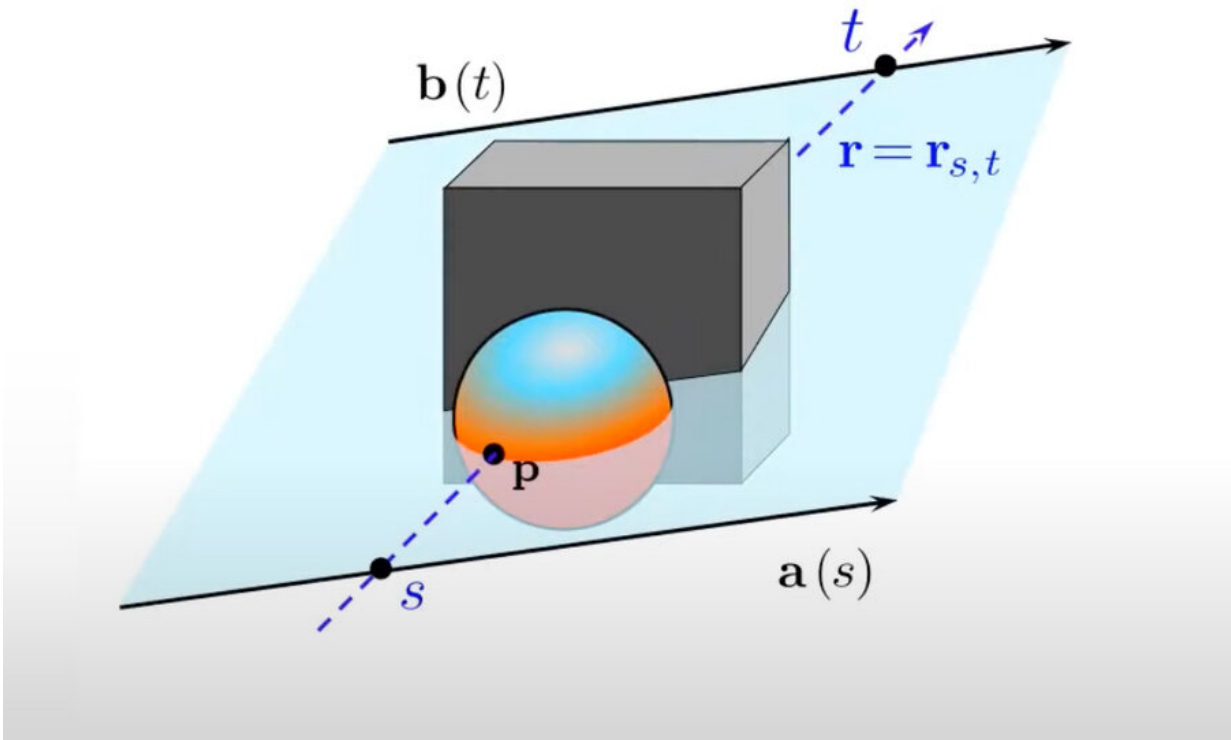
Given an image of a 3D scene and a light ray, a light field network can compute rich information about the geometry of the underlying 3D scene. Credit: Massachusetts Institute of Technology

"With other methods, when you do this rendering, you have to follow the ray until you find the surface. You have to do thousands of samples, because that is what it means to find a surface. And you're not even done yet because there may be complex things like transparency or reflections. With a light field, once you have reconstructed the light field, which is a complicated problem, rendering a single ray just takes a single sample of the representation, because the representation directly maps a ray to its color," Sitzmann says.

The LFN classifies each camera ray using its "Plücker coordinates,"

which represent a line in 3D space based on its direction and how far it is from its point of origin. The system computes the Plücker coordinates of each camera ray at the point where it hits a pixel to render an image.

By mapping each ray using Plücker coordinates, the LFN is also able to compute the geometry of the scene due to the parallax effect. Parallax is the difference in apparent position of an object when viewed from two different lines of sight. For instance, if you move your head, objects that are farther away seem to move less than objects that are closer. The LFN can tell the depth of objects in a scene due to parallax, and uses this information to encode a scene's geometry as well as its appearance.

But to reconstruct light fields, the neural network must first learn about the structures of light fields, so the researchers trained their model with many images of simple scenes of cars and chairs.

"There is an intrinsic geometry of light fields, which is what our model is trying to learn. You might worry that light fields of cars and chairs are so different that you can't learn some commonality between them. But it turns out, if you add more kinds of objects, as long as there is some homogeneity, you get a better and better sense of how light fields of general objects look, so you can generalize about classes," Rezchikov says.

Once the model learns the structure of a light field, it can render a 3D scene from only one image as an input.

Rapid rendering

The researchers tested their model by reconstructing 360-degree light fields of several simple scenes. They found that LFNs were able to render scenes at more than 500 frames per second, about three orders of magnitude faster than other methods. In addition, the 3D objects

rendered by LFNs were often crisper than those generated by other models.

An LFN is also less memory-intensive, requiring only about 1.6 megabytes of storage, as opposed to 146 megabytes for a popular baseline method.

"Light fields were proposed before, but back then they were intractable. Now, with these techniques that we used in this paper, for the first time you can both represent these light fields and work with these light fields. It is an interesting convergence of the mathematical models and the neural network models that we have developed coming together in this application of representing scenes so machines can reason about them," Sitzmann says.

In the future, the researchers would like to make their model more robust so it could be used effectively for complex, real-world scenes. One way to drive LFNs forward is to focus only on reconstructing certain patches of the light field, which could enable the model to run faster and perform better in real-world environments, Sitzmann says.

"Neural rendering has recently enabled photorealistic rendering and editing of images from only a sparse set of input views. Unfortunately, all existing techniques are computationally very expensive, preventing applications that require real-time processing, like video conferencing. This project takes a big step toward a new generation of computationally efficient and mathematically elegant neural rendering algorithms," says Gordon Wetzstein, an associate professor of electrical engineering at Stanford University, who was not involved in this research. "I anticipate that it will have widespread applications, in computer graphics, computer vision, and beyond."

 More information: Vincent Sitzmann et al, Light Field Networks: