BLOG ›

Reconstructing indoor spaces with NeRF

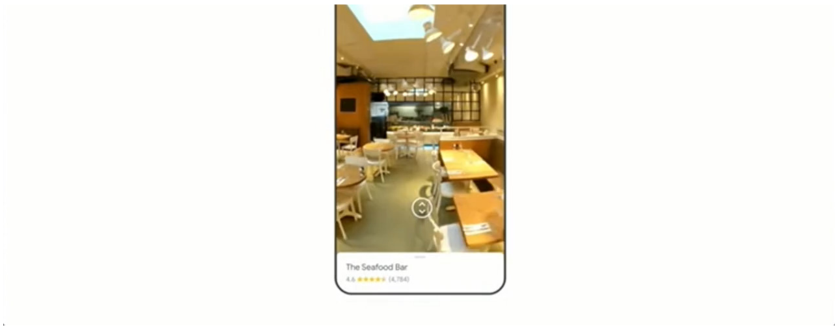
WEDNESDAY, JUNE 14, 2023

Marcos Seefelder, Software Engineer, and Daniel Duckworth, Research Software Engineer, Google Research

When choosing a venue, we often find ourselves with questions like the following: Does this restaurant have the right vibe for a date? Is there good outdoor seating? Are there enough screens to watch the game? While photos and videos may partially answer questions like these, they are no substitute for feeling like you're there, even when visiting in person isn't an option.

Immersive experiences that are interactive, photorealistic, and multi-dimensional stand to bridge this gap and recreate the feel and vibe of a space, empowering users to naturally and intuitively find the information they need. To help with this, Google Maps launched Immersive View, which uses advances in machine learning (ML) and computer vision to fuse billions of Street View and aerial images to create a rich, digital model of the world. Beyond that, it layers helpful information on top, like the weather, traffic, and how busy a place is. Immersive View provides indoor views of restaurants, cafes, and other venues to give users a virtual up-close look that can help them confidently decide where to go.

Today we describe the work put into delivering these indoor views in Immersive View. We build on neural radiance fields (NeRF), a state-of-the-art approach for fusing photos to produce a realistic, multi-dimensional reconstruction within a neural network. We describe our pipeline for creation of NeRFs, which includes custom photo capture of the space using DSLR cameras, image processing and scene reproduction. We take advantage of Alphabet's recent advances in the field to design a method matching or outperforming the prior state-of-the-art in visual fidelity. These models are then embedded as interactive 360° videos following curated flight paths, enabling them to be available on smartphones.

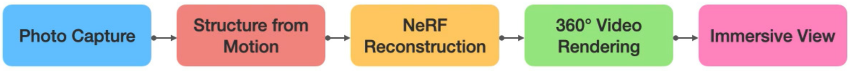


The reconstruction of The Seafood Bar in Amsterdam in Immersive View.

From photos to NeRFs

At the core of our work is **NeRF**, a recently-developed method for 3D reconstruction and novel view synthesis. Given a collection of photos describing a scene, NeRF distills these photos into a **neural field**, which can then be used to render photos from viewpoints not present in the original collection.

While NeRF largely solves the challenge of reconstruction, a user-facing product based on real-world data brings a wide variety of challenges to the table. For example, reconstruction quality and user experience should remain consistent across venues, from dimly-lit bars to sidewalk cafes to hotel restaurants. At the same time, privacy should be respected and any potentially personally identifiable information should be removed. Importantly, scenes should be captured consistently and efficiently, reliably resulting in high-quality reconstructions while minimizing the effort needed to capture the necessary photographs. Finally, the same natural experience should be available to all mobile users, regardless of the device on hand.



The capture, reconstruction, and presentation of indoor spaces is a process of five stages.

The Immersive View indoor reconstruction pipeline.

The first step to producing a high-quality NeRF is the careful capture of a scene: a dense collection of photos from which 3D geometry and color can be derived. To obtain the best possible reconstruction quality, every surface should be observed from multiple different directions. The more information a model has about an object's surface, the better it will be in discovering the object's shape and the way it interacts with lights.

In addition, NeRF models place further assumptions on the camera and the scene itself. For example, most of the camera's properties, such as white balance and aperture, are assumed to be fixed throughout the capture. Likewise, the scene itself is assumed to be frozen in time: lighting changes and movement should be avoided. This must be balanced with practical concerns, including the time needed for the capture, available lighting, equipment weight, and privacy. In partnership with professional photographers, we developed a strategy for quickly and reliably capturing venue photos using DSLR cameras within only an hour timeframe. This approach has been used for all of our NeRF reconstructions to date.

Once the capture is uploaded to our system, processing begins. As photos may inadvertently contain sensitive information, we automatically scan and blur personally identifiable content. We then apply a structure-from-motion pipeline to solve for each photo's camera parameters: its position and orientation relative to other photos, along with lens properties like focal length. These parameters associate each pixel with a point and a direction in 3D space and constitute a key signal in the NeRF reconstruction process.

NeRF reconstruction

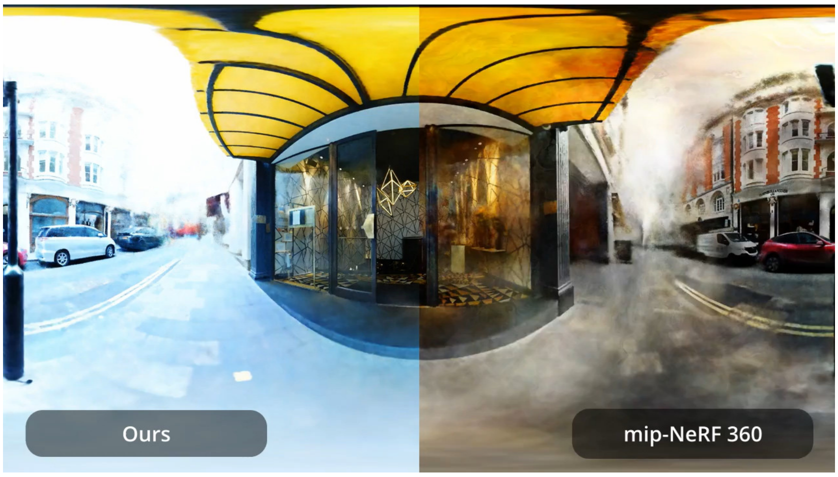
Unlike many ML models, a new NeRF model is trained from scratch on each captured location. To obtain the best possible reconstruction quality within a target compute budget, we incorporate features from a variety of published works on NeRF developed at Alphabet. Some of these include:

- We build on mip-NeRF 360, one of the best-performing NeRF models to date. While more computationally intensive than Nvidia's widely-used Instant NGP, we find the mip-NeRF 360 consistently produces fewer artifacts and higher reconstruction quality.
- We incorporate the low-dimensional generative latent optimization (GLO) vectors introduced in NeRF in the Wild as an auxiliary input to the model's radiance network. These are learned real-valued latent

phenomena such as lighting changes without resorting to cloudy geometry, a common artifact in casual NeRF captures.

- We also incorporate exposure conditioning as introduced in Block-NeRF. Unlike GLO vectors, which are uninterpretable model parameters, exposure is directly derived from a photo's metadata and fed as an additional input to the model's radiance network. This offers two major benefits: it opens up the possibility of varying ISO and provides a method for controlling an image's brightness at inference time. We find both properties invaluable for capturing and reconstructing dimly-lit venues.

We train each NeRF model on TPU or GPU accelerators, which provide different trade-off points. As with all Google products, we continue to search for new ways to improve, from reducing compute requirements to improving reconstruction quality.



A side-by-side comparison of our method and a mip-NeRF 360 baseline.

A scalable user experience

Once a NeRF is trained, we have the ability to produce new photos of a scene from any viewpoint and camera lens we choose. Our goal is to deliver a meaningful and helpful user experience: not only the reconstructions themselves, but guided, interactive tours that give users the freedom to naturally explore spaces from the comfort of their smartphones.

To this end, we designed a controllable 360° video player that emulates flying through an indoor space along a predefined path, allowing the user to freely look around and travel forward or backwards. As the first Google product exploring this new technology, 360° videos were chosen as the still resource intensive on a per-client basis (either on device or cloud computed), and relying on them would limit the number of users able to access this experience. By using videos, we are able to scale the storage and delivery of videos to all users by taking advantage of the same video management and serving infrastructure used by YouTube. On the operations side, videos give us clearer editorial control over the exploration experience and are easier to inspect for quality in large volumes.

While we had considered capturing the space with a 360° camera directly, using a NeRF to reconstruct and render the space has several advantages. A virtual camera can fly anywhere in space, including over obstacles and through windows, and can use any desired camera lens. The camera path can also be edited post-hoc for smoothness and speed, unlike a live recording. A NeRF capture also does not require the use of specialized camera hardware.

Our 360° videos are rendered by ray casting through each pixel of a virtual, spherical camera and compositing the visible elements of the scene. Each video follows a smooth path defined by a sequence of keyframe photos taken by the photographer during capture. The position of the camera for each picture is computed during structure-from-motion, and the sequence of pictures is smoothly interpolated into a flight path.

To keep speed consistent across different venues, we calibrate the distances for each by capturing pairs of images, each of which is 3 meters apart. By knowing measurements in the space, we scale the generated model, and render all videos at a natural velocity.

The final experience is surfaced to the user within Immersive View: the user can seamlessly fly into restaurants and other indoor venues and discover the space by flying through the photorealistic 360° videos.

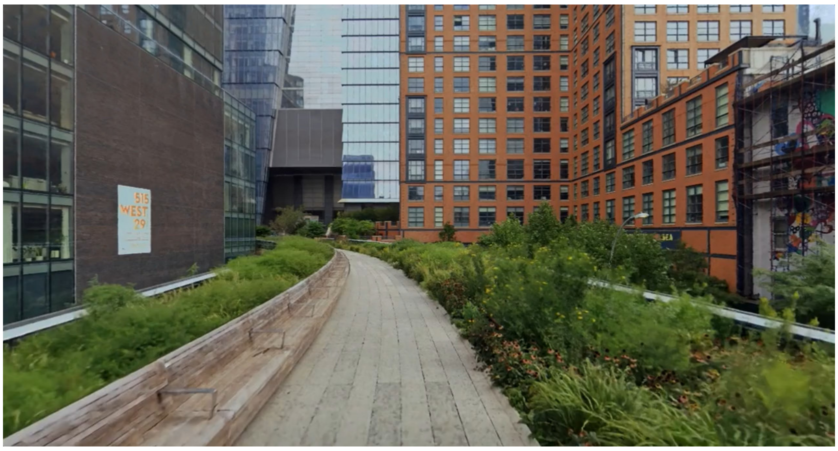
Open research questions

We believe that this feature is the first step of many in a journey towards universally accessible, AI-powered, immersive experiences. From a NeRF research perspective, more questions remain open. Some of these include:

1. Enhancing reconstructions with scene segmentation, adding semantic information to the scenes that could make scenes, for example, searchable and easier to navigate.

world and change how users could experience the outdoor world.

3. Enabling real-time, interactive 3D exploration through neural-rendering on-device.



Reconstruction of an outdoor scene with a NeRF model trained on Street View panoramas.

As we continue to grow, we look forward to engaging with and contributing to the community to build the next generation of immersive experiences.

Acknowledgments

This work is a collaboration across multiple teams at Google. Contributors to the project include Jon Barron, Julius Beres, Daniel Duckworth, Roman Dudko, Magdalena Filak, Mike Harm, Peter Hedman, Claudio Martella, Ben Mildenhall, Cardin Moffett, Etienne Pot, Konstantinos Rematas, Yves Sallat, Marcos Seefelder, Lilyana Sirakovat, Sven Tresp and Peter Zhizhin.

Also, we'd like to extend our thanks to Luke Barrington, Daniel Filip, Tom Funkhouser, Charles Goran, Pramod Gupta, Santi López, Mario Lučić, Isalo Montacute and Dan Thomasset for valuable feedback and suggestions.

Previous posts

Labels: Augmented Reality Computational Photography Google Maps

JUN 13, 2023

Enabling delightful user experiences via predictive

→

JUN 10, 2023

Imagen Editor and EditBench: Advancing and

→

JUN 8, 2023

Evaluating speech synthesis in many languages with

→