Instant3D: Revolutionizing Reconstruction with AI

Shivam Doshi, Prakhar Bhardwaj, Vigknesh Rajan, Mukundan Chariar

Team Name: Phoenix
Computer Vision for Engineers (24-678)
Fall 2023

Department of Mechanical Engineering
Carnegie Mellon University

Problem Statement

The project, titled "Instant3D: Revolutionizing Reconstruction with AI," presents a transformative approach to address challenges in traditional 3D reconstruction methods. By harnessing the power of ubiquitous smartphone cameras and advanced algorithms like Structure from Motion (SfM) and Neural Radiance Fields (NeRF), the project eliminates the prohibitive costs and computational demands associated with existing techniques. This innovation not only streamlines the 3D modeling process but also democratizes the technology, paving the way for widespread applications in robotics, urban planning, and the preservation of historical artifacts. The methodology, developed to be both user-friendly and cost-effective, marks a significant step towards a more sustainable and inclusive revolution in the digitization of the physical world.

In the realm of computer vision and 3D reconstruction, the challenge lies in creating highly detailed and accurate three-dimensional models of real-world scenes from a collection of 2D images. The traditional approach involves Structure from Motion (SfM) techniques for camera pose estimation and sparse point cloud generation, followed by mesh reconstruction. However, this process often results in limitations such as sparse representation and geometric inaccuracies. Furthermore, the emerging Neural Radiance Fields (NeRF) technology offers a promising avenue for synthesizing realistic 3D scenes directly from images, but the computational cost and training complexity are significant hurdles.

The problem at hand is to develop an integrated solution that leverages both Structure from Motion and instant Neural Radiance Fields techniques to achieve a more robust and efficient 3D reconstruction. This involves addressing challenges such as optimizing the accuracy of camera pose estimation, enhancing the density and quality of the reconstructed point clouds, and streamlining the integration of instant NeRF for more detailed and visually appealing reconstructions. The project aims to advance the state-of-the-art in 3D reconstruction by combining the strengths of SfM and instant NeRF, ultimately providing a tool that can generate high-fidelity 3D models from 2D image collections in a more efficient and accurate manner

Introduction

3D reconstruction is a process in computer vision and image processing that involves capturing the shape and appearance of real objects. This process can be used to create 3D models from 2D images or video. It's a complex field that combines elements of photogrammetry, computer vision, and computational geometry.

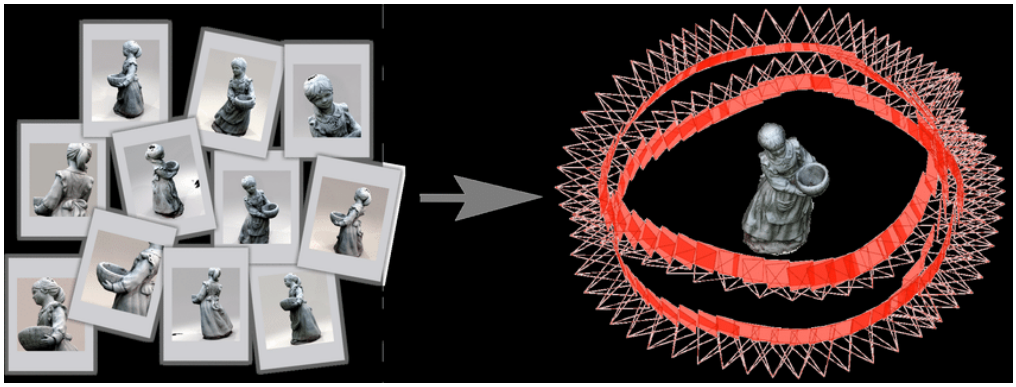


Figure 1: 3D Reconstruction

Fundamentals of 3D Reconstruction

- **Image Acquisition:** The first step is capturing images or videos of the object or scene from multiple angles. This can be done using standard cameras, specialized 3D scanners, or even drones for aerial photography.
- **Feature Detection and Matching:** Algorithms detect and match features across the different images. These features are specific points or patterns in the images that can be easily identified and tracked.
- **Estimation of Camera Parameters:** It involves determining the position and orientation of the camera for each image, which is crucial for accurate 3D reconstruction.
- **Triangulation and Point Cloud Generation:** Using the camera parameters and feature correspondences, the 3D coordinates of the points are estimated, typically resulting in a point cloud, which is a collection of points in 3D space representing the surface of the object or scene.
- **Creation of Meshes and Textures:** The point cloud is then converted into a mesh, a collection of vertices, edges, and faces that define the shape of the object. Textures from the original images are mapped onto the mesh to create a realistic 3D model.

Techniques used in 3D Recontruction

- **Structure from Motion (SfM):** This technique creates 3D structures from 2D image sequences. It's often used for smaller-scale projects and can be done with standard photographic equipment.

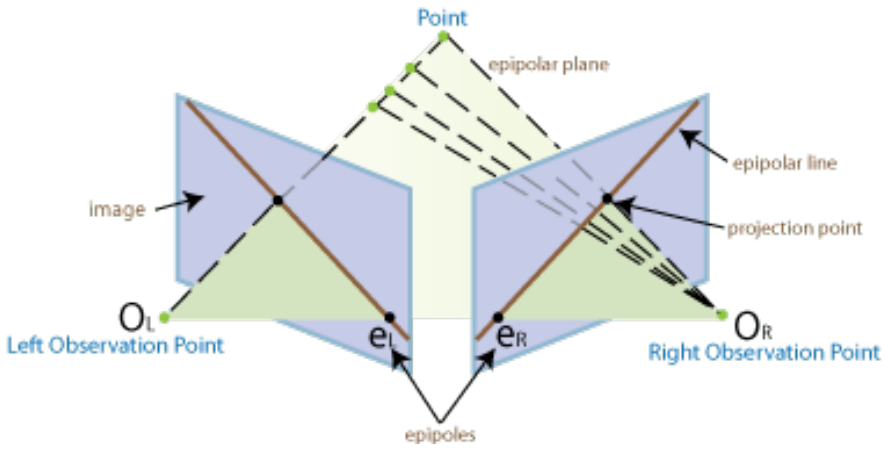


Figure 2: Structure from Motion

- **Stereoscopic Vision:** This approach mimics human binocular vision by using two cameras to capture images from slightly different angles, allowing for depth perception and 3D reconstruction.
- **Estimation of Camera Parameters:** These methods use laser light to capture highly accurate 3D data and are commonly used in topography, archaeology, and architecture.
- **LIDAR and Laser Scanning:** Using the camera parameters and feature correspondences, the 3D coordinates of the points are estimated, typically resulting in a point cloud, which is a collection of points in 3D space representing the surface of the object or scene.
- **Neural Radiance Fields (NeRF):** A recent development in 3D reconstruction, NeRF uses deep learning to create highly detailed and photorealistic 3D models from a set of 2D images. It models the volumetric scene function and can synthesize novel views of a scene.

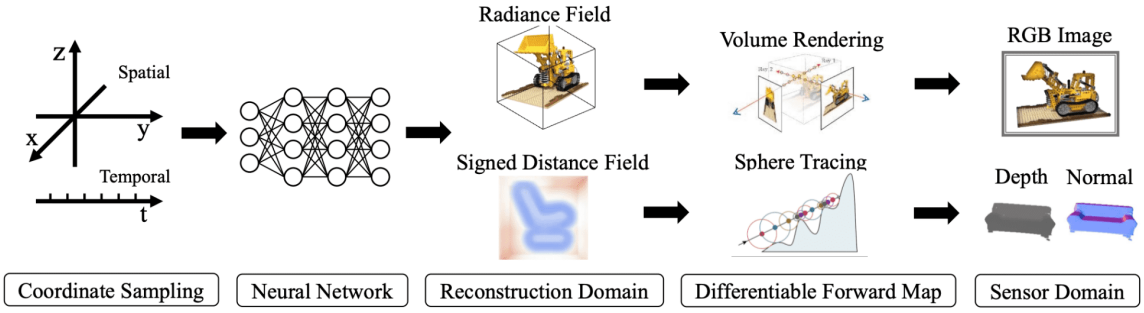


Figure 3: Neural Radiance Fields

Previous Work

3D Recontruction

The development of 3-D reconstruction has been intertwined with the evolution of computer vision (CV) for decades. In the 1980s, numerous CV algorithms were introduced, focusing on various

aspects of processing three-dimensional data, including modeling, acquisition, merging, and recognition. A significant advancement during this period was the use of discrete Markov Random Field (MRF) models for optimization in global image search. These models played a pivotal role in finding disparity maps and obtaining depth information from images in subsequent years. [1]

Moving into the 1990s, there were notable improvements in tracking and image segmentation techniques. These advancements paved the way for the development of multi-view stereo algorithms crucial for 3-D reconstruction. The 2000s saw the introduction of more sophisticated algorithms for complex global optimization, broadening the application of 3D processing to fields like medicine, archaeology, and robotics. [2]

In recent times, the integration of machine learning techniques with CV technology has demonstrated significant potential. Notably, showcases the use of 3-D vision combined with machine learning to train robots in grasping previously unseen objects at specific points, marking a leap forward in the application of artificial intelligence in CV technology.

Structure from Motion

The field of Structure-from-Motion (SfM) applied to unordered images has undergone significant advancements over time. Early systems focused on self-calibrating metric reconstruction laid the groundwork for more complex applications. These initial systems were pivotal in handling unordered Internet photo collections and urban scenes. This early success spurred the creation of larger-scale reconstruction systems, capable of processing from hundreds of thousands to recently, even a hundred million Internet photos. [5] Various SfM methodologies have emerged, including incremental, hierarchical, and global approaches, with incremental SfM becoming particularly prominent for reconstructing unordered photo collections. Despite its widespread adoption, the quest to develop a truly versatile, general-purpose SfM system remains an ongoing challenge in the field.

NeRF

Recent studies have explored the concept of representing 3D shapes in a continuous form using deep learning models. [3] These models function by transforming xyz coordinates into signed distance functions or occupancy fields, effectively mapping out 3D shapes. Initially, these methods relied heavily on having access to accurate 3D geometry, usually sourced from synthetic datasets like ShapeNet. However, newer approaches have evolved, overcoming this limitation by developing differentiable rendering functions. These advanced techniques enable the optimization of neural implicit shape representations using just 2D images, removing the dependency on pre-existing 3D shape data. When there's a comprehensive collection of viewpoints, it becomes possible to recreate photorealistic new views through straightforward interpolation of light field samples. This technique has been supported by various studies. [4] In scenarios where there's a sparser collection of views, significant advancements have been made in the field of computer vision and graphics. These advancements involve generating traditional geometry and appearance representations from the images captured. A notable methodology within this domain involves using mesh-based scene representations. These representations can either incorporate diffuse appearance models or can be tailored to change depending on the viewing angle.

Structure from Motion (SfM)

Structure from motion (SfM) is the process of estimating the 3-D structure of a scene from a set of 2-D images. It is the process of reconstructing 3D structure from its projections into a series of images taken from different viewpoints. SfM is used in many applications, such as 3-D scanning, augmented reality, and visual simultaneous localization and mapping (vSLAM).

SfM can be computed in many different ways. The way in which you approach the problem depends on different factors, such as the number and type of cameras used, and whether the images are ordered. If the images are taken with a single calibrated camera, then the 3-D structure and camera motion can only be recovered up to scale. up to scale means that you can rescale the structure and the magnitude of the camera motion and still maintain observations. For example, if you put a camera close to an object, you can see the same image as when you enlarge the object and move the camera far away.

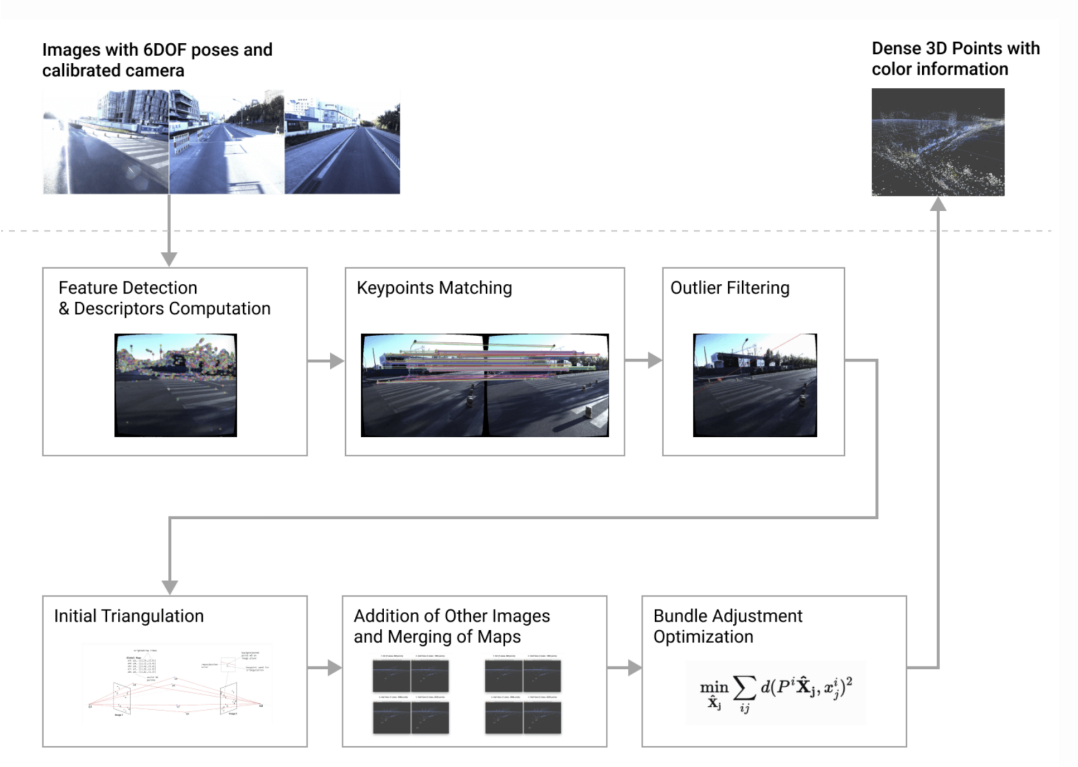


Figure 4: SfM Pipeline

Steps involved in SfM

- **Dataset Collection:** We took a series of overlapping photographs of the object or scene from different angles. The number and quality of these images can significantly affect the accuracy of the final 3D model. If possible, record the intrinsic parameters of the camera (like focal length and lens distortion).
- **Feature Detection:** Identify Key Points by using algorithms to detect distinctive features in each image. These features are often points that can be easily tracked across multiple images. The next step involves Feature Matching. This step involves identifying the same physical

point in the scene from different images.

1. Estimating Fundamental Matrix:
The fundamental matrix, denoted by F, is a 3×3 (rank 2) matrix that relates the corresponding set of points in two images from different views (or stereo images). But in order to understand what fundamental matrix actually is, we need to understand what epipolar geometry is! The epipolar geometry is the intrinsic projective geometry between two views. It only depends on the cameras' internal parameters (K matrix) and the relative pose i.e. it is independent of the scene structure.

2. Epipolar Geometry:
Let's say a point X in the 3D-space (viewed in two images) is captured as x in the first image and x' in the second. Can you think how to formulate the relation between the corresponding image points x and x'? Consider Fig. 2. Let C and C' be the respective camera centers which form the baseline for the stereo system. Clearly, the points x, x', and X (or C, C', and X) are coplanar, i.e., $C\vec{x} \cdot (\vec{C'C} - \vec{x'}) = 0$, and the plane formed can be denoted by π. Since these points are coplanar, the rays back-projected from x and x' intersect at X. This is the most significant property in searching for a correspondence.

3. The Fundamental Matrix F:
The **F** matrix is only an algebraic representation of epipolar geometry and can be understood both geometrically (constructing the epipolar line) and arithmetically (See derivation) (Fundamental Matrix Song). As a result, we obtain: $x_i'^T \mathbf{F} x_i = 0$ where $i = 1, 2, \ldots, m$. This is known as the epipolar constraint or correspondence condition (or Longuet-Higgins equation). Since **F** is a 3×3 matrix, we can set up a homogeneous linear system with 9 unknowns:

$$\begin{bmatrix} x_i' & y_i' & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = 0$$

$$x_i x_i' f_{11} + x_i y_i' f_{21} + x_i f_{31} + y_i x_i' f_{12} + y_i y_i' f_{22} + y_i f_{32} + x_i' f_{13} + y_i' f_{23} + f_{33} = 0$$

Simplifying for m correspondences

$$\begin{bmatrix} x_1 x_1' & x_1 y_1' & x_1 & y_1 x_1' & y_1 y_1' & y_1 & x_1' & y_1' & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_m x_m' & x_m y_m' & x_m & y_m x_m' & y_m y_m' & y_m & x_m' & y_m' & 1 \end{bmatrix} \begin{bmatrix} f_{11} \\ f_{21} \\ f_{31} \\ f_{12} \\ f_{22} \\ f_{32} \\ f_{13} \\ f_{23} \\ f_{33} \end{bmatrix} = 0$$

With $N \geq 8$ correspondences between two images, the fundamental matrix, F, can be obtained as follows: By stacking the above equation in a matrix A, the equation $Ax = 0$ is obtained. This system of equations can be solved by using the linear least squares method with Singular Value Decomposition (SVD), as explained in the Math module. When applying SVD to matrix A, the decomposition USV^T is obtained, with U and V being orthonormal matrices and S a diagonal matrix containing the singular values. Thus, the last column of V is the true solution, given that $\sigma_i \neq 0$ for all $i \in [1, 8]$, $i \in \mathbb{Z}$. However, due to noise in the correspondences, the estimated F matrix can be of rank 3, i.e., $\sigma_9 \neq 0$. So, to enforce the rank 2 constraint, the last singular value of the estimated F must be set to zero. If F has full rank, then it will have an empty null-space, i.e., it won't have any point that is on the entire set of lines. Thus, there wouldn't be any epipoles. See Fig. 3 for full rank comparisons for F matrices.

- **Camera Pose Estimation:** Calculate where each image was taken from (camera position) and the direction the camera was facing (camera orientation). This process typically involves solving complex mathematical equations and may use methods like bundle adjustment.

  The camera pose consists of 6 degrees-of-freedom (DOF): Rotation (Roll, Pitch, Yaw) and Translation (X, Y, Z) of the camera with respect to the world. Since the essential matrix E is identified, the four camera pose configurations: (C_1, R_1), (C_2, R_2), (C_3, R_3), and (C_4, R_4), where $C \in \mathbb{R}^3$ is the camera center and $R \in SO(3)$ is the rotation matrix, can be computed. Thus, the camera pose can be written as: $P = KR[I_{3\times3} - C]$.

  These four pose configurations can be computed from the E matrix. Let $E = UDV^T$ and $W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. The four configurations can be written as:

$$
\begin{aligned}
C_1 &= U(:,3) \text{ and } R_1 = UWV^T \\
C_2 &= -U(:,3) \text{ and } R_2 = UWV^T \\
C_3 &= U(:,3) \text{ and } R_3 = UWV^TV^T \\
C_4 &= -U(:,3) \text{ and } R_4 = UWV^TV^T
\end{aligned}
$$

  It is important to note that $\det(R) = 1$. If $\det(R) = -1$, the camera pose must be corrected, i.e., $C = -C$ and $R = -R$.

- **Triangulation:** For each matched feature across images, use triangulation to estimate its 3D position. The result is a sparse 3D representation of the scene or object, known as a point cloud.

  Given two camera poses and linearly triangulated points X, the locations of the 3D points that minimize the reprojection error (Recall Project 2) can be refined. Linear triangulation minimizes the algebraic error, but the reprojection error is a geometrically meaningful error that can be computed by measuring the error between measurements and projected 3D points:

$$
\min_x \sum_{j=1,2} \left(u_j - P_j^T 1 \tilde{\phi} P_j^T 3X\right)^2 + \left(v_j - P_j^T 2 \tilde{\phi} P_j^T 3X\right)^2
$$

  Where: - u_j and v_j are the measured image coordinates for the j-th camera. - P_j represents the projection matrix for the j-th camera. - $\tilde{\phi}$ is the homogeneous coordinate transformation.

- X represents the 3D point. - The sum is taken over all camera poses (in this case, $j = 1$ and $j = 2$). - The objective is to minimize the sum of squared differences between the measured image coordinates and the projected 3D points by adjusting the 3D point coordinates X.

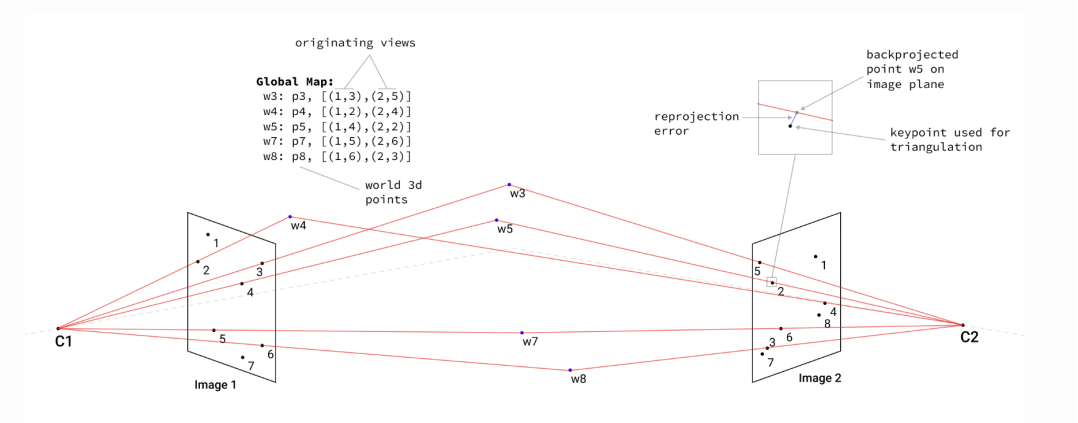


Figure 5: Example of initial image pair triangulation.

- **Bundle Adjustment:** Optimize and adjust the 3D points and camera parameters to minimize re-projection error (the difference between the observed image points and the projected points from the 3D model).

  Minimize the loss function:

$$\min_{\hat{X}_j} \sum_j \left(\sum_i d\left(P_i \hat{X}_j, x_{ij}\right)^2 \right) \tag{1}$$

  where

  - $d(a, b)$ is the geometric distance between two points;
  - $\hat{X}_j$ is an estimated 3D point in a world space;
  - P_i is a projection matrix for camera i;
  - x_{ij} are 2D coordinates of a keypoint in image i that corresponds to the 3D point $\hat{X}_j$;
  - $P_i \hat{X}_j$ is a backprojection of point $\hat{X}_j$ to image i.

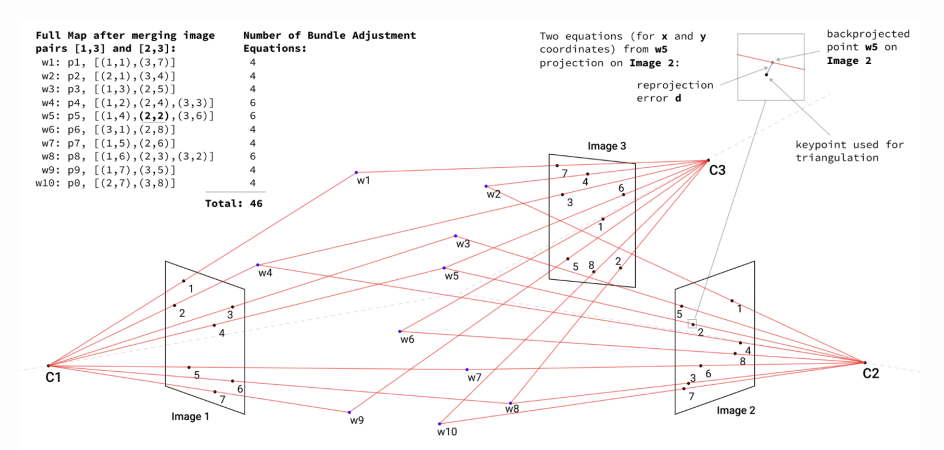


Figure 6: Bundle adjustment for 3 pair images

- **Dense Reconstruction:** Increase the density of the 3D point cloud to create a more detailed 3D model. This involves generating new points that were not originally identified in the feature detection phase.

NeRF: Neural Radience Fields:

Neural Radiance Fields, commonly referred to as NeRF, represent a significant advancement in the field of computer vision and graphics. This approach centers on the use of Neural Radiance Fields (NeRF), which is a method for creating highly detailed and realistic 3D representations of scenes from a set of 2D images. The core idea is to represent a static scene as a continuous 5D function, mapping every point in space (x, y, z) and direction (θ, ϕ) to a specific radiance (color) and density. This function is modeled using a deep fully-connected neural network (also known as a multilayer perceptron or MLP), which is trained to regress from a 5D coordinate to a single volume density and view-dependent RGB color.

Steps involved in NeRF

- **Generating a Sampled Set of 3D Points by Marching Camera Rays Through the Scene:** This step involves simulating the path of light rays as they would travel from a camera into the scene. Imagine a virtual camera positioned at a specific viewpoint. From this viewpoint, rays (analogous to lines of sight) are projected through each pixel of the image plane into the scene. As these rays penetrate the scene, they intersect with various 3D points in space. The process of calculating these intersections is called "ray marching." By sampling points along these rays, you effectively create a series of coordinates in 3D space that the neural network will later use to estimate color and density.

- **Producing an Output Set of Densities and Colors:** For each sampled 3D point, its spatial coordinates (x, y, z) and the 2D direction of the ray (θ, ϕ) that intersects it are fed into a trained neural network. The neural network, which has learned a mapping from these 5D coordinates to color and density, outputs these values for each point. The color represents the RGB value that the point contributes to the final image, while the density indicates how opaque that point is (i.e., how much it obstructs light).

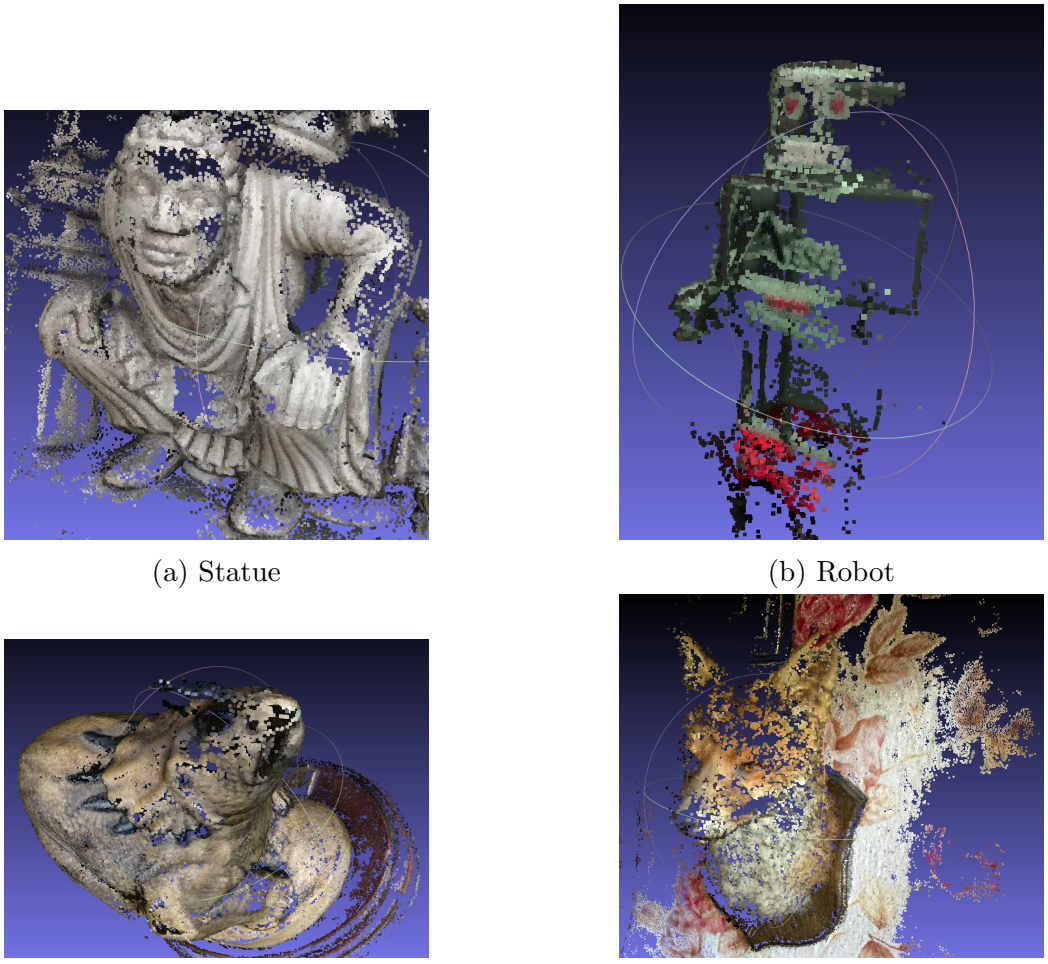
- **Accumulating Densities and Colors into a 2D Image Using Classical Volume Rendering Techniques:** Volume rendering is a technique used to create a 2D image from 3D scalar fields, like the densities and colors output by the neural network. This process involves integrating information along each ray. For each ray, starting from the near end (closest to the camera) and moving towards the far end, the contributions of each sampled point are accumulated. The color and opacity at each point along a ray are combined to compute the color of the corresponding pixel in the image. Points closer to the camera can obscure those further away, depending on their density. This accumulation process results in a 2D image that represents the scene as seen from the chosen viewpoint, with realistic lighting, shadows, and occlusions.

The approach also overcomes challenges like the storage costs of voxel grids, and includes technical contributions like representing scenes as 5D neural radiance fields, a differentiable rendering procedure, and a positional encoding technique for handling high-frequency scene content. This method

surpasses existing view synthesis techniques, enabling the rendering of high-resolution, photorealistic views of real-world scenes.

Results

Originally our results were not satisfactory. SfM performs a good job of reconstruction, but the density of the obtained reconstruction is low, and some parts of the object are not continuous.



(a) Statue

(b) Robot

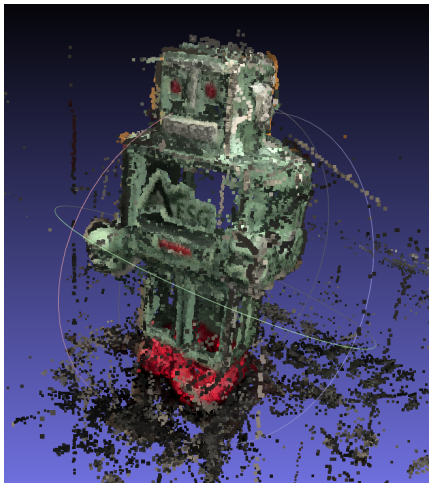
(c) Dino

(d) Fox

Figure 7: Bad reconstructions of objects

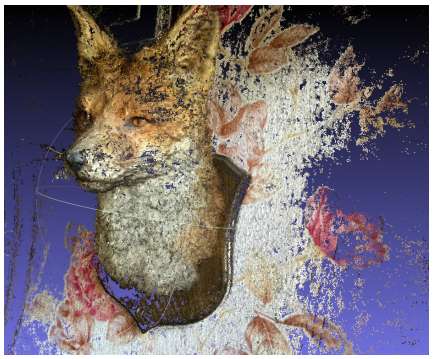
Once we upsample our images, we get much better results. The density of our reconstruction improves.

(a) Statue



(b) Robot

(c) Dino

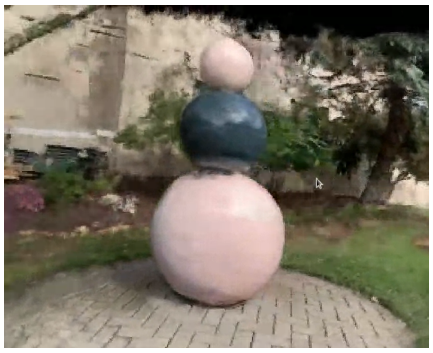


(d) Fox

Figure 8: Good reconstructions of objects

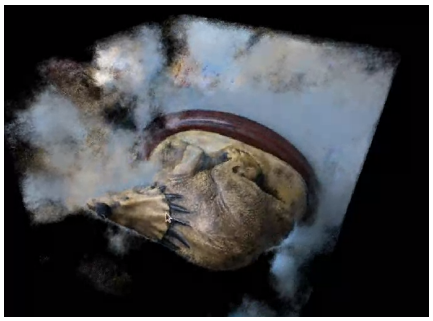
Upsampling improves the detection of features and makes our reconstructions more continuous. Even so, with SfM we cannot produce novel views of the object. We use NERF to produce these views. Nerf is successful at producing novel views of the object with the lighting conditions considered.

(a) Ball



(b) Ball view 2

Figure 9: Good reconstructions of objects



(a) Dino

(b) Dino view 2

Figure 10: Good reconstructions of objects

(a) Fox

(b) Fox view 2

Figure 11: Good reconstructions of objects

References

[1] Julius Butime, Iñigo Gutierrez, L Galo Corzo, and C Flores Espronceda. 3d reconstruction methods, a survey. In *Proceedings of the First International Conference on Computer Vision Theory and Applications*, pages 457–463, 2006.

[2] Zhiliang Ma and Shilong Liu. A review of 3d reconstruction techniques in civil engineering and their applications. *Advanced Engineering Informatics*, 37:163–174, 2018.

[3] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *CoRR*, abs/2003.08934, 2020.

[4] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[5] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35:151–173, 1999.