ISSN No:-2456-2165

Neural Radiance Fields-Comprehensive Survey

Anay Dongre
Information Technology, Marathwada Mitra Mandal's College of Engineering

Abstract:- Neural Radiance Fields (NeRF) is a machine learning model that can generate high-resolution, photorealistic 3D models of scenes or objects from a set of 2D images. It does this by learning a continuous 3D function that maps positions in 3D space to the radiance (intensity and color) of the light that would be observed at that position in the scene.

To create a NeRF model, the model is trained on a dataset of 2D images of the scene or object, along with their corresponding 3D positions and orientations. The model learns to predict the radiance at each 3D position in the scene by using a combination of convolutional neural networks (CNNs) and a differentiable renderer.

I. INTRODUCTION

Neural Radiance Fields (NRF) is a powerful deep learning technique for generating high-quality 3D images and videos. It is a type of generative model that is built on top of the traditional Radiance Fields (RF) model, which is a method for rendering 3D objects and scenes in computer graphics. NRF combines the advantages of the RF model, such as its ability to handle complex and realistic scenes, with the advantages of deep neural networks, such as their ability to learn from data and generalize to new situations. There are several types of Neural Radiance Fields (NRF) models, each with its own specific architecture and implementation details. Some of the most common types of NRF models include:

A. Single-View NRF

This type of NRF model generates 3D images and videos from a single 2D image or video. It uses an encoder-decoder architecture to extract features from the input data and generate the output image or video.

B. Multi-View NRF

This type of NRF model generates 3D images and videos from multiple 2D images or videos. It uses an encoder-decoder architecture to extract features from the input data and generate the output image or video.

C. Object-Level NRF

This type of NRF model generates 3D images and videos of objects, rather than entire scenes. It uses an encoder-decoder architecture to extract features from the input data and generate the output image or video of the object.

D. Volumetric NRF

This type of NRF model generates 3D images and videos by representing the scene as a set of volumetric data, such as a 3D grid of voxels. It uses an encoder-decoder architecture to extract features from the input data and generate the output image or video.

E. Hierarchical NRF

This type of NRF model generates 3D images and videos by using a hierarchical architecture that allows the model to generate images and videos at different levels of detail. This can be useful for handling large and complex scenes.

It's worth noting that these types of NRF models are not mutually exclusive and some NRF models can combine some of these types.

II. BACKGROUND

The traditional RF model represents an object or scene as a function that maps the 3D position of a point in the scene to the radiance or light intensity at that point. The function is represented using a set of predefined basis functions, such as spherical harmonics, that are used to approximate the function. The RF model can be used to generate high-quality 3D images and videos by rendering the scene from different viewpoints and lighting conditions. However, the traditional RF model has some limitations, such as the need for a large number of basis functions to represent the function accurately and the difficulty of handling complex and realistic scenes.

The NRF model overcomes these limitations by using deep neural networks to learn the basis functions from data. The NRF model is trained on a dataset of 3D objects or scenes, and it learns to generate new images and videos that are similar to the training data. The model consists of two main components: a canonical network that maps the input data to a set of parameters, and a deformation network that maps the parameters back to the output image or video.

III. ARCHITECTURE

The NRF model is composed of two main components: a canonical network and a deformation network. The canonical network is typically a convolutional neural network (CNN) that extracts features from the input data, and the deformation network is typically a transposed convolutional neural network (DeconvNet) that generates the output image or video.

ISSN No:-2456-2165

The canonical network takes an input 3D scene and outputs a set of parameters that represent the scene. The deformation network takes these parameters as input and generates the output image or video.

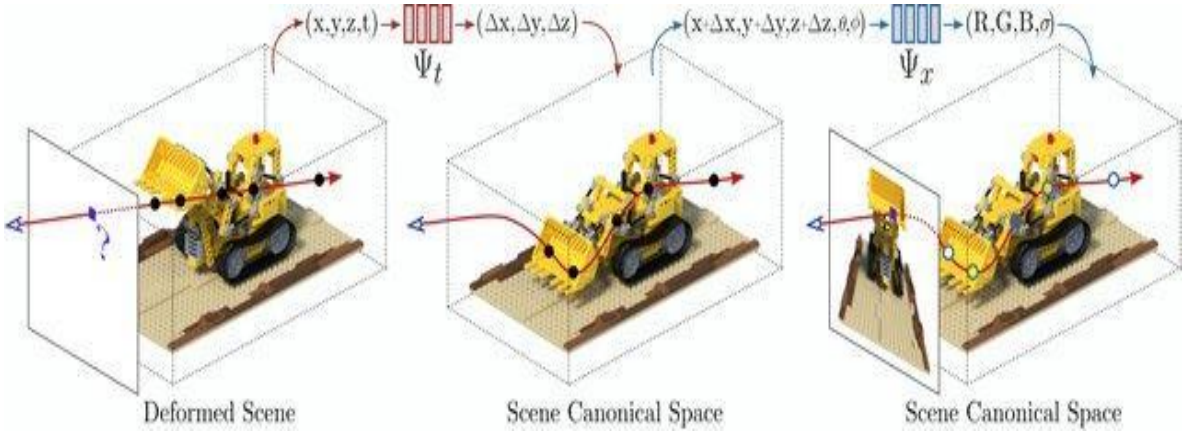

Fig 1: D-NeRF Model. The proposed architecture consists of two main blocks: a deformation network Ψt mapping all scene deformations to a common canonical configuration; and a canonical network Ψx regressing volume density and view-dependent RGB color from every camera ray.

Given a point x and viewing direction d at time instant t we first transform the point position to its canonical configuration as $\Psi t : (x, t) \rightarrow \Delta x$. Without loss of generality, we chose $t = 0$ as the canonical scene $\Psi t : (x, 0) \rightarrow 0$. By doing so the scene is no longer independent between time instances, and becomes interconnected through a common canonical space anchor.

Then, the assigned emitted color and volume density under viewing direction d equal to those in the canonical configuration $\Psi x : (x + \Delta x, d) \rightarrow (c, \sigma)$. We propose to learn Ψx and Ψt using a sparse set of T RGB images {It, Tt} T t=1 captured with a monocular camera, where It ∈ R H×W×3 denotes the image acquired under camera pose Tt ∈ R 4×4 SE(3), at time t.

A. Canonical Network

In the Neural Radiance Field (NRF) model, the canonical network is responsible for mapping the input data, which is a 3D scene, to a set of parameters that represent the scene. The canonical network is typically implemented as a convolutional neural network (CNN) that extracts features from the input data.

The architecture of the canonical network typically consists of several convolutional layers, each followed by a non-linear activation function, such as the rectified linear unit (ReLU), that is used to introduce non-linearity into the model. Each convolutional layer applies a set of filters to the input data, which are learned during the training process, to extract features from the data. As the input data is passed through multiple convolutional layers, the features become increasingly complex and abstract, capturing more and more information about the input data.

The final layer of the canonical network is typically a fully connected layer that maps the feature maps to the parameters of the scene. The number of parameters depends on the complexity of the 3D scene and the specific task.

The canonical network is trained to extract the most relevant information from the input data and represent it in a compact and efficient way in the form of a set of parameters. These parameters are then passed to the deformation network, which generates the output image or video based on these parameters. The goal of the training process is to learn the parameters that can be used to reconstruct the input data with high accuracy and realism.

B. Deformation Network

In the Neural Radiance Fields (NRF) model, the deformation Network network is responsible for taking the parameters generated by the canonical network and mapping them back to the output image or video. The deformation network is typically implemented as a transposed convolutional neural network (DeconvNet) that generates the output image or video.

The architecture of the deformation network typically consists of several transposed convolutional layers, each followed by a non-linear activation function, such as the rectified linear unit (ReLU), that is used to introduce non-linearity into the model. A transposed convolutional layer, also known as a fractionally-strided convolution or a deconvolutional layer, is a convolutional layer that is designed to increase the spatial resolution of the data, unlike regular convolutional layers which reduce the spatial resolution of the data.

The final layer of the deformation network is typically a convolutional layer that generates the output image or video. The output image or video should be similar to the input image

or video, but with some variations based on the parameters passed by the canonical network.

The deformation network is trained to generate high-quality images and videos that are consistent with the input data and the parameters passed by the canonical network. The deformation network is typically trained together with the canonical network and the discriminator network in a GAN setup, where the discriminator network is used to distinguish between the generated images and the real images.

Ψt(x, t) = Δx, if t 6= 0
0 , if t = 0

Directly feeding raw coordinates and angles to a neural network results in low performance. Thus, for both the canonical and the deformation networks, we first encode x, d and t into a higher dimension space. We use the same positional encoder where γ(p) =< (sin(2lπp), cos(2lπp)) >L 0 . We independently apply the encoder γ(·) to each coordinate and camera view component, using L = 10 for x, and L = 4 for d and t.

C. Volume Rendering

The volume rendering network is a specific type of network that is used to generate images and videos of volumetric data. The architecture of the volume rendering network typically consists of several layers, such as a ray-casting layer that traces the path of light through the volume, and a compositing layer that combines the contributions of light from different voxels to generate the final image. The volume rendering network may also include additional layers such as a shading layer, which applies lighting and shading effects to the image, and a post-processing layer, which applies final adjustments such as gamma correction or denoising. The volume rendering network is trained to learn the properties of the volumetric data, such as the color and opacity of the voxels, and to generate high-quality images and videos that are consistent with the input data.

We now adapt NeRF volume rendering equations to account for non-rigid deformations in the proposed 6D neural radiance field. Let x(h) = o+hd be a point along the camera ray emitted from the center of projection o to a pixel p. Considering near and far bounds hn and hf in that ray, the expected color C of the pixel p at time t is given by:

C(p, t) = Z hf hn T(h, t)σ(p(h, t))c(p(h, t), d)dh, (2)

where p(h, t) = x(h) + Ψt(x(h), t), (3)

[c(p(h, t), d), σ(p(h, t))] = Ψx(p(h, t), d), (4)

and T(h, t) = exp − Z h hn σ(p(s, t))ds! . (5)

The 3D point p(h, t) denotes the point on the camera ray x(h) transformed to canonical space using our Deformation Network Ψt, and T(h, t) is the accumulated probability that the ray emitted from hn to hf does not hit any other particle. Notice that the density σ and color c are predicted by our Canonical Network Ψx.

IV. LEARNING THE MODEL

The parameters of the canonical Ψx and deformation Ψt networks are simultaneously learned by minimizing the mean squared error with respect to the T RGB images {It} T t=1 of the scene and their corresponding camera pose matrices {Tt} T t=1. Recall that every instant is only acquired by a single camera. At each training batch, we first sample a random set of pixels {pt,i} Ns i=1 corresponding to the rays cast from some camera position Tt to some pixels i of the corresponding RGB image t. The training loss we use is the mean squared error between the rendered and real pixels:

$$\mathcal{L} = \frac{1}{N_s} \sum_{i=1}^{N_s} \left\| \hat{C}(p,t) - C'(p,t) \right\|_2^2$$

where Cˆ are the pixels' ground-truth color

V. COMPARING RESULTS

To evaluate our model we compare against current top-performing techniques for view synthesis, detailed below:

A. Neural Volumes (NV)

Neural Volumes is a method for generating 3D images and videos using deep learning. It is a type of generative model that is based on the idea of representing 3D scenes as a set of volumetric data, such as a 3D grid of voxels. The main goal of Neural Volumes is to generate high-quality images and videos of 3D scenes that are consistent with the input data.

B. Scene Representation Networks (SRN)

Scene Representation Networks (SRNs) are a type of deep learning model that are used to generate 3D images and videos of scenes. The main goal of SRNs is to generate high-quality images and videos that are consistent with the input data and that can capture the geometric and semantic properties of the scene.

C. Ground Truth

Ground truth refers to the correct or true 3D representation of an object or scene. In NRF-based models, the goal is to train a model to generate high-quality 3D images and videos that are as close as possible to the ground truth representation. The ground truth in NRF can be obtained from a variety of sources such as 3D models, real-world scans, or synthetic scenes.

D. Layered Light Field Fusion (LLFF)

It is a method used to generate high-resolution images from a set of low-resolution images. It is a type of image upsampling method that utilizes the information in multiple low-resolution images to generate a single high-resolution image.

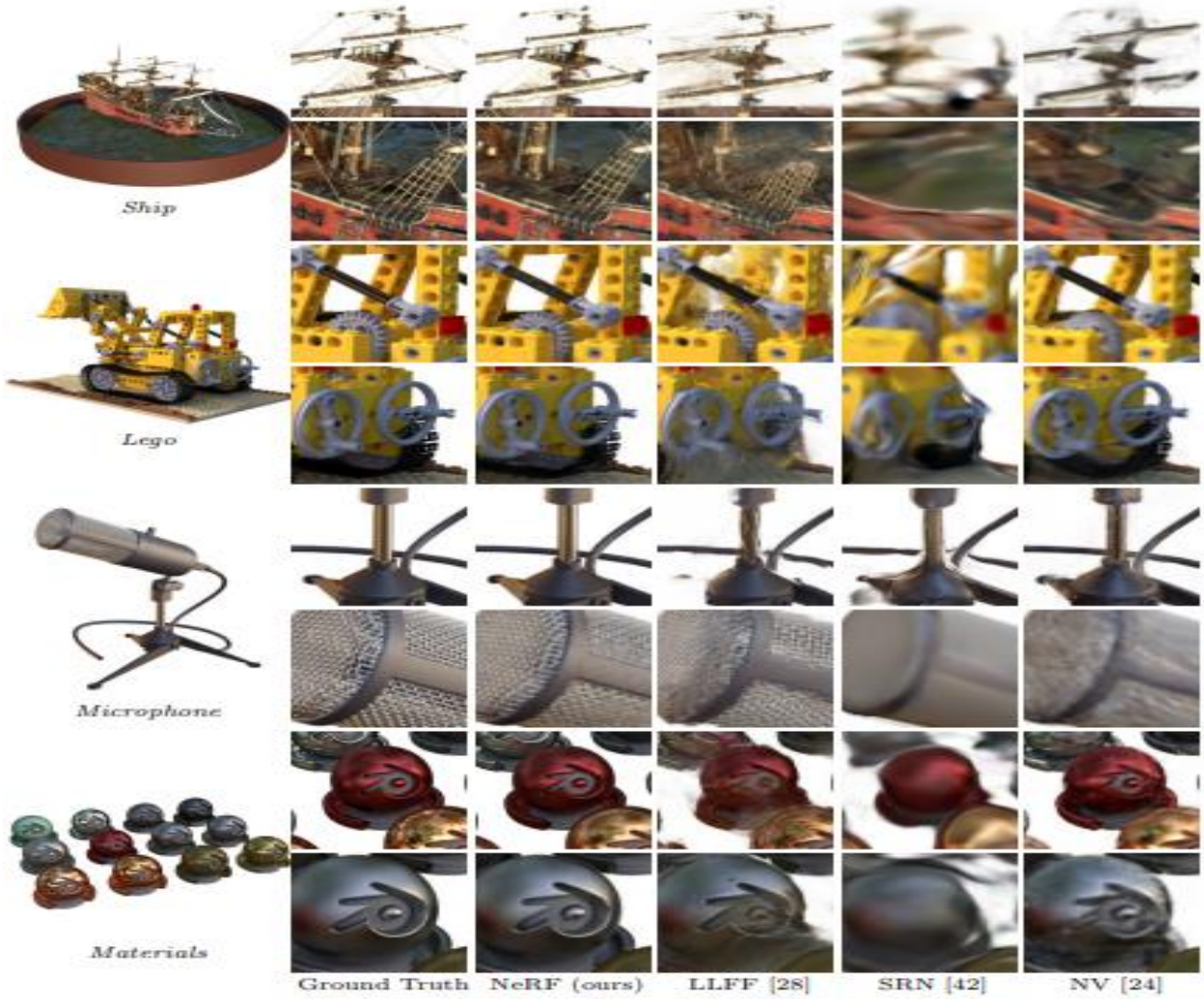

Fig 2. Comparing various methods with NeRF

VI. CONCLUSION

In comparison with other methods for generating 3D images and videos, such as Scene Representation Networks (SRN) and Neural Volumes, NRF has the ability to generate high-quality 3D images and videos that are highly realistic and consistent with the input data. Additionally, it can handle large and complex scenes, and it can generate images and videos of novel objects and scenes, and handle different lighting conditions. It is worth noting that, while NRF has many advantages over other methods, it also has some limitations. For example, it can be computationally expensive and may require large amounts of data to train. Additionally, the architecture and training process of NRF is complex and requires a good understanding of the underlying algorithms and technologies.

REFERENCES

[1]. Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM* 65.1 (2021): 99-106.

[2]. Pumarola Peris, Albert, et al. "D-NeRF: neural radiance fields for dynamic scenes." *Proceeding of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Institute of Electrical and Electronics Engineers (IEEE), 2021.

[3]. Antonio Agudo and Francesc Moreno-Noguer. Simultaneous pose and non-rigid shape with particle dynamics. In CVPR, 2015.

[4]. Aayush Bansal, Minh Vo, Yaser Sheikh, Deva Ramanan, and Srinivasa Narasimhan. 4d visualization of dynamic events from unconstrained multi-view videos. In CVPR, 2020. 2, 3

[5]. Adrien Bartoli, Yan Gerard, Francois Chadebecq, Toby ´ Collins, and Daniel Pizarro. Shape-from-template. T-PAMI, 37(10), 2015.
[6]. Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. X-fields: Implicit neural view-, lightand time-image interpolation. TOG, 39(6), 2020.
[7]. Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In SIGGRAPH, 2001
[8]. "An Introduction to NeRF: Neural Radiance Fields." *Ai-contentlab*, 31 Dec. 2022, www.ai-contentlab.com/2022/12/an-introduction-to-nerf-neural-radiance.html?m=1.