NeRF: Geometry-aware 3D scene modelling

Yao Fang [1,†] **and Yuzhao Liu** [2,3,†]

[1]Academy of Fine arts, Jiangsu second Normal University, Nanjing, China.
[2]Academy of Information Engineering, China University of Geosciences Beijing, Beijing, China.

[3]1004221111@email.cugb.edu.cn
[†]These authors contributed equally.

Abstract. NeRF (Neural radiation field) as one of the most popular research fields, the method has also been extended to the field of viewpoint synthesis, and the use of NeRF to represent three-dimensional scenes or models has been proposed, so that it has been successfully applied to the field of viewpoint synthesis and achieved high-quality synthesis results. This paper reviews the principle of NERF and its advantages and disadvantages, focusing on the areas where NeRF needs to be improved. The research directions of NERF in different fields were introduced and analysed. For example, the construction of Block-NeRF in Google Maps makes unmanned technology go to a higher level, and the digital protection and restoration of ancient cultural relics. And the research directions that can be optimized and expanded by NERF in the future are analysed and proposed.

Keywords: NeRF, geometry-aware, 3D modelling, scene.

1. Introduction

NERF (Neural radiation field) uses implicit representation (using a function to describe scene geometry, the advantage is that it is suitable for large-resolution scenes, but the disadvantage is that it cannot generate photo-level virtual perspective) to encode volume density and color observation to achieve photo-level View synthesis effect. Treat the scene as a continuous field of opacity and color, the color is related to the viewing angle, and finally synthesize the picture through volume rendering [1]. This expression is continuous in 3D and can synthesize very high-quality pictures. However, rendering a NeRF needs to repeatedly calculate the output value of a large MLP at many 3D points, which cannot achieve real-time rendering. NERF has successfully solved the problem of invisible view synthesis. At the same time, 3D models are becoming increasingly popular in VR and AR applications. Training such 3D models using voxels or point clouds is challenging and requires sophisticated tools for proper color rendering. The advent of NeRF solved this problem.

NeRF's research work can be broadly divided into two categories according to the purpose of the research: analysis and optimization of the NeRF algorithm itself, and extension and extension based on the NeRF framework [2]. Among them, some studies have optimized the drawing efficiency and accuracy of NeRF and improved speed and quality, while some studies have improved the norm, and some have expanded and developed the research and development direction of NeRF to solve more complex problems. This paper reviews the principle of NeRF and its advantages and disadvantages, introduces and analyzes the research directions of NeRF in different fields, and analyzes and puts forward the research directions that can be optimized and extended by NeRF in the future.

2. Principle analysis of NeRF

NeRF is one of the hottest research fields at present, and the effect is very amazing. The problem it needs to solve is how to generate a picture under a new perspective given some captured pictures. Different from the traditional 3D reconstruction method that puts. The scene is expressed as an explicit expression such as point cloud, grid, voxel, etc. It has a unique way to model the scene as a continuous 5D radiation field and store it implicitly in the neural network [1,3]. It only needs to input sparse multi-angle images with poses A neural radiation field model is obtained through training, and clear photos from any viewing angle can be rendered according to this model. Generally speaking, it is to construct an implicit rendering process, whose input is the position o, direction d and corresponding coordinates (x, y, z) of the light emitted from a certain viewing angle, and send it into the neural radiation field Fθ to obtain the volume density and color, and finally get the final image through volume rendering.

The NeRF rendering process can be further refined as follows.

(1) Data acquisition: NeRF needs to acquire a set of input images from different angles, and each image needs to know its camera position and orientation. Usually, the images captured by the acquisition cameras are 2D images that need to be taken from different angles and enough images need to be taken to capture the 3D structure of the scene.

(2) Training the neural network: NeRF uses these input images and the corresponding camera parameters to train a neural network. This neural network is called a "radiance field" and consists of several fully connected layers, each consisting of multiple neurons. The input of this network is the coordinates of a 3D spatial point and the camera parameters, and the output is an RGB color and transparency, usually a 4-dimensional vector, with each component representing red, green, blue and transparency.

(3) Implicit sampling: In order to render an image, many points are sampled in 3D space and then the color and transparency of each point is calculated by a neural network. These sampled points are generated by a method called "implicit sampling". Specifically, NeRF uses a ray of light to pass through the 3D space and calculates the color and transparency of the intersection of this ray with the objects in the scene; NeRF samples this ray uniformly, i.e., distributes a certain number of sample points evenly over the ray, each of which corresponds to a 3D spatial point and the corresponding camera parameters [4]. These 3D spatial points and camera parameters are used as input to the neural network, and the color and transparency are computed by forward propagation, and finally the color and transparency of all the sampled points are combined to obtain the color and transparency of this light.

(4) Light projection: When we render an image, we actually project the light from each pixel into 3D space, and then implicitly sample the color and transparency at that point to produce the image. This process is called "ray projection". For each pixel, NeRF first calculates the corresponding ray, then implicitly samples this ray to obtain the color and transparency, and finally combines the color and transparency to obtain the color of the pixel.

(5) Post-processing: The rendered image usually requires some post-processing, such as denoising and color correction. Common post-processing methods are bilateral filtering and color mapping

(6) Acceleration methods: NeRF rendering is very slow because it requires ray projection and implicit sampling of each pixel point, and for a 1280 x 720 image, millions of neural network forward propagations need to be computed. To speed up this process, there are techniques that can be used, such as spatial indexing structures, such as the use of k-d trees, in order to find ray-object intersections faster and speed up the implicit sampling process.

$$(x,y,z,\theta,\phi)\rightarrow\blacksquare\ \blacksquare\ \blacksquare\rightarrow(RGB\sigma)F\Theta \qquad (1)$$

x,y,z are the 3D coordinate positions of the space points θ and ϕ are Perspective directions

Specifically, in the process of calculating color c = (r, g, b), this rendering function looks like here. When the opaque points in front of a ray accumulate to a certain degree to obscure the points behind, the points behind will not play a role even if they are opaquer. So NeRF uses formulas (2) and (3) to ignore the effect of the points behind.

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma\big(r(t)\big)c(r(t), d)dt \tag{2}$$

T(t) is the cumulative transparency of the ray on the path from t_n to t. Its specific form is formula (3)

$$T(t) = exp(-\int_{t_n}^{t_f} \sigma(r(s))ds) \tag{3}$$

It is usually not possible to calculate the integral in a neural network, so it can be expressed in terms of $\sum$ by discretizing the integration process. This is the formula used in the calculation of the data for the actual rendering [5].

$$t_i \sim u[t_n + \frac{i-1}{N}(t_f - t_n), \ t_n + \frac{i}{N}(t_f - t_n)] \tag{4}$$

The $\mathrm{t_i}$ sampling point can be expressed as formula (5)

$$C(r) = \sum_{i=1}^{N} T_i \, (1 - exp(-\sigma_i\delta_i))c_i, \ where \ T_i = exp(-\sum_{j=1}^{\mathrm{i}-1} \sigma_j\delta_j) \tag{5}$$

NeRF is an elegant representation capable of expressing complex 3D scenes for downstream tasks such as synthesizing new perspectives and content generation. NeRF was first applied in the direction of new viewpoint synthesis. Due to its super strong ability to implicitly express 3D information, it has developed rapidly in the direction of 3D reconstruction. Its core point is to use a neural network to unambiguously use a complex static scene modeling. After the network training is completed, clear scene pictures can be rendered from any angle.

In summary, NeRF provides a new tool for the field of computer vision and graphics by learning the 3D structure and colors in a scene and can generate high-quality images from different perspectives.

3. The fields where NeRF can be applied:

3.1. Application of Block-NeRF in map modeling

NERF can be applied to reconstruct large-scale environments in areas such as autonomous driving and aerial surveying. For example, establish a large-scale high-fidelity map for robot positioning and navigation. For example, researchers at Google AI and Google's own self-driving company Waymo practiced a new idea. They tried to use 2.8 million Street View photos to reconstruct the entire 3D environment of downtown San Francisco. Since high-definition maps are required to train an automatic driving simulation system, and the automatic driving system usually needs to re-simulate the previously encountered scenes for evaluation. However, any deviation from the recorded path may alter the vehicle's trajectory [6]. Therefore, a high-fidelity simulation along the path is necessary. This simulation includes view rendering, which not only synthesizes basic views but also utilizes scene-conditioned NeRF. This approach can modify environmental conditions, such as ambient lighting, to simulate different weather and lighting exposure. Moreover, appearance embeddings can be manipulated to interpolate between different conditions observed in the training data, such as cloudy versus clear skies or day and night. This method can significantly enhance the simulated scene.

As a variant of the neural radiation field, Block-NeRF can exhibit large-scale environments. Specifically, when scaling NeRF to render city-scale scenes that span multiple blocks, it is critical to break down the scene into individually trained NeRFs. This decomposition method decouples render time from scene size, allowing rendering to scale to arbitrarily large environments and allowing updates to each slice of environment. Block-NeRF incorporates architectural changes to make NeRF robust to data collected over months under different environmental conditions (Figure 1). This method adds appearance embedding, learning pose refinement, and controlled exposure to each individual NeRF, and introduces an appearance alignment procedure between adjacent NeRFs so that they can be seamlessly combined.

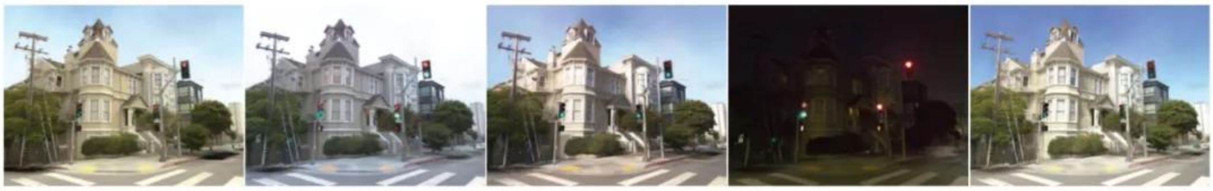

Figure 1. Application of Block-NeRF in map modelling.

3.2. NERF's restoration and preservation of cultural relics

NERF can be applied to the restoration and preservation of cultural relics. Starting from images, reconstruct virtual cultural heritage. For example, the community digital human DAO team recommended a photo-digital scene reconstruction project combining implicit functions and neural rendering published at the Siggraph2022 conference. The author of the paper said introduces a new method for the efficient and accurate surface reconstruction of images from Internet photo collections to generate 3D models in the presence of different lighting conditions (Figure 2). The author of the paper said also conducted an intensive comparison between the generated model and the model included in the Heritage-Recon cultural relic model library, and the fit is very high [7]. This not only enhances the accuracy of the restoration of cultural relics, but also enables the reconstruction of most of the cultural relics that have disappeared physically but still have photos. Because it is a virtual cultural relic, there is no need to worry about the damage to the cultural relics caused by touching, which solves the problem of not being able to see the details of the cultural relics clearly through the glass in the museum.

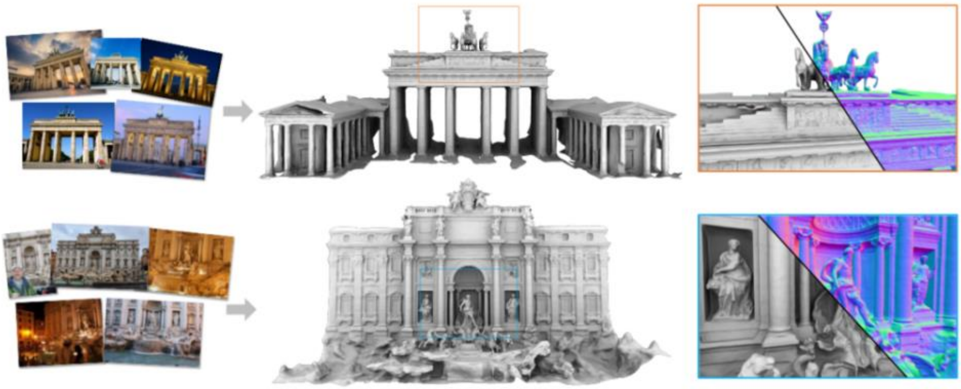

Figure 2. NERF's restoration and preservation of cultural relics.

3.3. Application of Head-NeRF in face modeling

With the rise of the meta-universe, the digital virtual human industry has also risen, which can be adjusted to the favourite face through the software HeadNeRF and applied to personal avatar or game image. In terms of effect, HeadNeRF can render the face head of the high-definition image level in real time, and can directly edit and adjust the various semantic attributes of the rendering result, such as identity, expression, and color appearance. Thanks to the introduction of the NeRF structure, HeadNeRF also supports direct editing and adjustment of the rendering perspective to achieve excellent rendering consistency.

Since the shape of most people's faces and the distribution of their facial features are similar, the face itself is suitable for parametric modelling. However, existing digital models based on traditional geometric representation of faces are still limited. The GCL Laboratory of the University of Science and

Technology of China has developed a new digital face head model construction method, HeadNeRF. HeadNeRF is a face rendering method that integrates NeRF and face digital models, its advantages lie in its effectiveness, and it can render picture-level face models in real time, and can directly edit people's facial features, face shape and facial texture (Figure 3) [8].

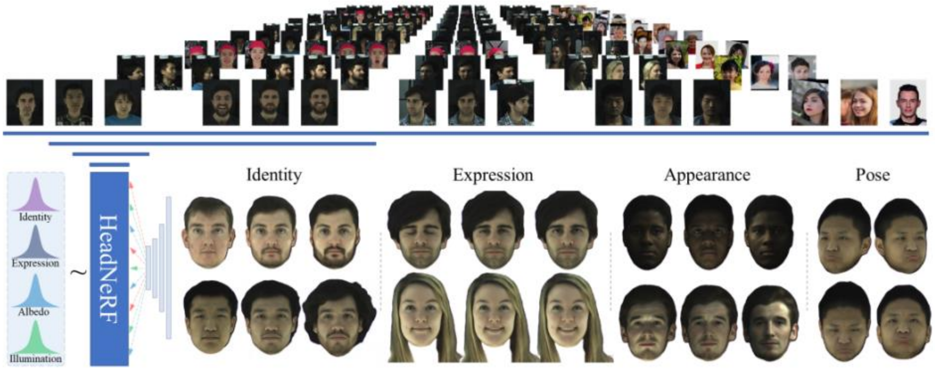

Figure 3. Application of Head-NeRF in face modelling.

3.4. Application of NeRF in video processing.

NeRF can be applied to video processing, such as building animatable 3D models from videos. For example, the cat in the picture, its model is constructed through the video of cat. For example, when people want to miss someone, they can also apply this technology to turn the person they miss into an animated 3D model.

BANMo is a tool for building animatable 3D neural models from multiple casual videos. With 2D cues from thousands of images integrated into a fixed canonical space, BANMo does not require pre-defined templates or pre-prepared multi-view cameras.

4. The advantages of NeRF

- The model detail texture is more fine

Because traditional 3D modelling will have many disadvantages, such as the final finished model may have problems such as overlapping textures, distortion, etc., and due to the limitation of voxel resolution, we may lose a lot of details [9]. However, NeRF can synthesize new perspectives at the photo level, and the generated model will be richer in detail, it forms a continuous volumetric scene function through a very small input view, achieving the best effect of comprehensive complex scene view, forming a model without voids and high degree of detail reproduction, and because of the novelty and interesting technology, many people study it, so that its development has become reasonable.

- The convenience of perspective synthesis

Perspective synthesis refers to the method of synthesizing observation (pictures) of a scene given to a specific perspective and synthesizing observations (pictures) from a new perspective. The intermediate 3D scene acts as an intermediary by generating high-quality virtual perspectives. How to represent this intermediate 3D scene is divided into "display representation" and "implicit representation", and then render this intermediate 3D scene to generate realistic perspective.

- The advantage of "implicit representation" over "display representation"

The "display representation" 3D scene includes Point Cloud, Voxel, Mesh, Volume, etc. It can explicitly model the scene, but because it is discrete, it will cause artifacts such as overlaps due to insufficient refinement, and more importantly It is because the amount of 3D scene expression information stored in it is extremely large, and the consumption of memory limits the application of high-resolution scenes [10]. "Implicit representation" of 3D scenes usually uses a function to describe the scene geometry, which can be understood as storing complex 3D scene expression information in the parameters of the function. Because it is often to learn a description function of a 3D scene, when expressing a large-resolution scene, its parameter amount is relatively small compared with the "display representation", and the "implicit representation" function is a continuous expression. The expression of the scene will be more refined.

5. Where NeRF needs to be improved:

(1) Calculation speed is slow: First, there are few effective pixels, and the effective pixels of the generated 2D image are less than 1/3. Being able to quickly obtain effective pixels can improve the inference speed. When the NeRF method generates an image, each pixel requires nearly 200 forward predictions of the MLP depth model. Although the scale of a single calculation is not large, the amount of calculation required to complete the rendering of the entire image pixel by pixel is still very large. Second, there are few effective voxels, and 192 points are sampled. Only the point density σ near the surface is relatively large, and reasoning is not necessary for other points. Third, the network reasoning speed is slow, and NeRF needs to train very slowly for each scene. It needs 12 layers of fully connected network reasoning to get the color and density of 1 voxel. Optimizing this performance can also greatly speed up reasoning.

(2) Only for static scenes: The NeRF method only considers static scenes and cannot be extended to dynamic scenes. This problem is mainly combined with monocular video to learn the implicit representation of the scene from monocular video.

(3) Poor generalization: The NeRF method needs to be retrained for a new scene, and cannot be directly extended to unseen scenes, which is obviously contrary to the goal of people's pursuit of generalization.

(4) Requires a large number of views: Although the NeRF method can achieve excellent view synthesis results, it requires a large number (hundreds) of views for training, which limits its application in reality.

(5) There is currently no successful commercial application: Because the data from the experiments in the paper do not work well in real situations. If technology cannot be applied, it will become a castle in the air. Implicit neural representations are a new key to the future of vision, and the significance of their existence will take time and experiment to prove. But the several demos it presents so far are indeed our most ideal pursuit of vision and graphics.

6. Conclusion

This article introduces the background and principles of NeRF, summarizes its advantages and areas for improvement, and also analyzes the application of NeRF in various directions and its development direction. Although NeRF has many problems at the beginning, such as slow computing speed, large number of perspectives, poor normalization, etc., many people have proposed solutions to these problems and are studying and optimizing it and integrating it into new fields. For example, a large number of NeRF-based improvement work, such as portrait reconstruction, scene combination, the combination of dynamic video and static video to reconstruct moving objects, etc. In the current development of NeRF, the surface of the object has been successfully reconstructed, so will there be any subsequent attempts to reconstruct and analyze the internal framework of the object, not just the surface? In fact, NeRF itself is not complicated, but its generation effect is particularly good, which also inspired us to explore more visual and graphic intersectional inspiration.

References

[1] ZhuFang.3D scene characterization—a review of recent results of neural radiation fields. *J. Cryst. Growth* **5**:64-77

[2] Jo K, Shim G, Jung Setal. Cg-nerf: Conditional generative neural radiance fields. arXiv preprint 2021 arXiv:2112.03517.

[3] Deng K.Liu, A. Zhu, J Y & Ramanan, D. Depth-supervised nerf: Fewer views and faster training for free. 2022 In Conf. Comput Vis. Patt. Recogni. 12882-12891.

[4] Condorelli F, Rinaudo F, Salvadore F, etal. A comparison between 3D reconstruction using nerf neural networks and mvs algorithms on cultural heritage images. 2021, The Intern. Arch. Photo., 43: 565-570.

[5] Kosiorek A R, Strathmann H, Zoran D, et al. Nerf-vae: A geometry aware 3d scene generative model.2021, Inter. Conf. Mach. Learn., 5742-5752.

[6] Yang G, Vo M, Neverova N, et al. Banmo: Building animatable 3d neural models from many casual videos 2022 In Conf. Comput Vis. Patt. Recogni. 2863-2873.

[7] Wang B, Chen L, Yang B. DM-NeRF: 3D Scene Geometry Decomposition and Manipulation from 2D Images. arXiv preprint arXiv:2208.07227,2022.

[8] Nguyen-Phuoc T, Liu F, Xiao L. Snerf: stylized neural implicit representations for 3d scenes. arXiv preprint arXiv:2207.02363, 2022.

[9] Metzer G, Richardson E, Patashnik O, et al. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures, arXiv preprint arXiv:2211.07600, 2022.

[10] Hong Y, Peng B, Xiao H, et al. Headnerf: A real-time nerf-based parametric head model. 2022 In Conf. Comput Vis. Patt. Recogni 20374-20384.