Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/45/20241016

An overview of Neural Radiance Fields

Xiaoju Liu

School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611730, China

2020010905010@std.uestc.edu.cn

Abstract. Synthesizing controllable, photo-realistic images and videos is one of the fundamental goals of computer graphics. Neural rendering is a rapidly emerging field in image synthesis that allows a compact representation of scenes, and by utilizing neural networks, rendering can be learned from existing observations. Neural Radiance Fields (NeRF) implement an effective combination of Neural Fields and the graphics component Volume rendering. It achieves the first photo-level view synthesis effect using an implicit representation. Unlike previous approaches, NeRF chooses Volume as an intermediate representation to reconstruct an implicit Volume. Although the advantages of NeRF are apparent, there are many drawbacks in the original version of NeRF: it is slow to train and render, requires a large number of perspectives, can only represent static scenes, and the trained NeRF representation does not generalize to other scenes. This report focuses on optimizing the shortcomings mentioned above of NeRF by scholars in the last three years and analyzes the solutions to the problems of NeRF from several perspectives.

Keywords: Neural Rendering, Neural Radiance Fields, Deep Learning, 3D Reconstruction.

1. Introduction

In recent years, with the gradual development of unmanned vehicles, drones, AR, VR, and other fields, the application field of 3D vision technology is becoming increasingly extensive, and the progress of these technologies is also driving the continuous advancement of three-dimensional vision technology. The focus of 3D-Reconstraction technology is how to obtain the depth information of the target scene or object. Traditional 3D-Reconstraction technology is mainly divided into two kinds passive and active. Among them, the process of passive three-dimensional reconstruction is: first, by taking pictures of the object from different angles to obtain RGB images of the object to be reconstructed, and then the initial point cloud of the camera pose and model is derived from it. Finally, the 3D model is obtained through a series of operations such as depth estimation, dense reconstruction of the point cloud, and mesh optimization of the resulting data[1]. Active 3D-Reconstraction refers to the use of light or energy sources such as lasers, acoustic waves, and electromagnetic waves emitted to the target object by receiving the returned light waves to obtain the depth images of the object, and the depth information is processed to reconstruct the 3D model of the target object[2]. Compared with NeRF, the traditional 3D reconstruction has some disadvantages, such as the final reconstructed model may have holes, texture overlap, and loss of many details due to voxel resolution limitations. In 2020, NeRF received a lot of attention from the academic community and became the best paper of ECCV 2020 due to its advantages,

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/45/20241016

such as the ability to synthesize new perspectives at the photo level, the richer details of the reconstructed model, the best result of synthesizing complex scene views, the absence of voids, and detail restoration, etc. This technique has been favored by scholars since the day it was proposed. A large number of researchers have explored and improved it, which has made it an important force in the field of 3D vision within two years.

1.1. Neural Fields

In the field of neural networks, a neural field is a mathematical framework that utilizes a neural network as a parameter, either partially or entirely. Within the realm of vision, a neural field is utilized to simulate the generation of target scalars such as depth, color, or other dimensions using spatial coordinates, time, or camera pose as inputs. To achieve the above functions, a Multi-Layer Perceptron (MLP) network [3] should be involved to simulate the target equation.

1.2. Volume Rendering

Volume rendering means generating two-dimensional images using the volume data obtained from three-dimensional images. Its most common application directions are rendering non-rigid objects, such as clouds, smoke, and jelly, which may be less dense or non-solid[4]. In fields such as medicine, geology, and meteorology, 3D data is commonly collected and stored in volume format. These data must be rendered into 2D images to be readily understood by humans. Volume rendering involves generating images or videos by simulating the propagation and absorption of light within the scene. This process can be accomplished by using one of several methods, including ray casting, ray marching, and ray tracing. NeRF has extended this technology to also include the rendering of solid objects.

1.3. Neural rendering

To understand Neural Radiance Fields, it is important to first comprehend the concept of neural rendering. Neural rendering, especially in 3D, builds upon classic computer graphics principles by learning to render and/or represent a scene using real-world images. These images can be a collection of unordered sets or structured multi-view images and videos. Neural rendering mimics the physical process that takes place when a camera captures a scene, but there is a key difference: the training process separates the camera capture process (such as projection and image formation) from the 3D scene's representation [5]. This separation has several benefits, including maintaining a high level of 3D consistency during image synthesis for viewpoint synthesis. To achieve this separation, 3D neural rendering techniques rely on image formation models commonly used in computer graphics, such as rasterization, point stitching, and volume integration.

1.4. Neural Radiance Fields

Neural Radiance Fields is a deep learning model oriented to 3D implicit spatial modeling, a deep learning model also known as MLP. The task to be performed by NeRF is Novel View Synthesis, which is defined as a series of captures of a scene in a known viewpoint (including the captured images, as well as each image's corresponding internal and external parameters) without the intermediate 3D reconstruction process, and synthesizing the images in the new viewpoint based only on the pose intrinsic and images. Under the NeRF-based representation, the 3D space is represented as a set of learnable and continuous radiation fields, which are learned from the input viewpoint and position to obtain the density as well as color[6]. NeRF marks a breakthrough in MLP-based scene characterization for single-scene, realistic novel view synthesis.

The NeRF method requires nearly 200 forward predictions of the MLP depth model for each pixel when producing raw images. Despite the small size of a single computation, the computational effort to complete the rendering of the whole image on a pixel-by-pixel basis is substantial. Second, the time required for NeRF to train for each scene is also slow.

DOI: 10.54254/2755-2721/45/20241016

2. Methodology

In this paper, by reading academic journals, conference papers, reports, and abstracts on Neural Radiance Fields in the past three years, we meticulously read and analyzed the literature involving the improvement of the defects of Neural Radiance Fields, selected the articles that addressed the problems of NeRF, and categorized them according to the specific problems they addressed The articles are classified and summarized according to the specific problems they address, and their advantages and disadvantages are analyzed. By comparing the functionalities of the NeRF improvement measures, we propose general directions to comprehensively improve the performance of NeRF and make up for its shortcomings.

3. Existing solutions for increasing training and rendering speed

To address the slow rendering problem of NeRF, Matthew Johnson et al. proposed FastNeRF [7], a system that decomposes NeRF into two neural networks: a location-dependent network that produces a deep radiosity map and an orientation-dependent network that produces weights. The inner product of the weights and the deep radiosity map estimates the colors in the scene as seen at the specified location and from the specified direction. FastNeRF can be efficiently cached, significantly improving the efficiency of test time while maintaining the visual quality of NeRF.

NeRF uses a rendering procedure that samples the scene with a single ray per pixel and thus may produce overly blurred or mixed renderings when testing images observe scene content at different resolutions. This situation is usually caused by the inconsistent resolution of multiple images corresponding to the same scene (the relative sampling frequency changes due to different camera distances, thus causing signal distortion). If multiple rays per pixel are to be rendered, this results in a significant increase in computation, as rendering each ray requires hundreds of queries to the multi-layer perceptron. Mip-NeRF is an extension of NeRF by adopting the vertebral form arrangement for modeling, which is anti-aliased compared to NeRF by arranging 3D points in the form of rays for modeling. Ultimately, Mip-NeRF has the advantage of being faster, smaller, and more accurate than NeRF, and is more suitable for handling multiscale data [8].

Moreover, since standard volume rendering does not have a mandatory constraint, incorrect geometry can occur when the number of input views is insufficient. With this in mind, Depth-supervised NeRF (DS-NeRF) takes advantage of the sparse 3D points generated by structure-from-motion (SFM) and uses them as a "free supervision", adding a loss to encourage ray termination. The depth distribution is matched to a given 3D key point and incorporates depth uncertainty [9]. Research shows that DS-NeRF can render better images and can improve training speed by 2-3 times.

Although the NeRF method is capable of achieving excellent view synthesis, it has a very large requirement for the number of views used for training, which can run into hundreds of views, limiting its application in reality. PixelNeRF uses a CNN Encoder to propose image features, thus making 3D points generalizable and supporting a small number of inputs. It is even possible to get 3D features from just one input image [10]. Although the model still requires some input camera pose and viewpoint, it also extends the range of applications available for NeRF.

Chen et al. [11] gave MVSNeRF, which provided a solution to the problem from another direction in 2021. By building a generalized deep neural network, the network can reconstruct the radiance fields from only three nearby input views by fast network inference. This is done by using planar swept cost volumes which had a wide application in multi-view stereo for geometry-aware scene inference and combining it with physically-based volume rendering for neural radiation field reconstruction.

NeRF makes it difficult to perform 3D reconstruction of large-scale scenes due to the excessive time required to optimize individual scenes. To address this problem, Zhang et al. [12] give a new solution called NeRFusion, which combines NeRF and TSDF-based fusion techniques for efficient rendering and 3D reconstruction of large-scale images. The truncated symbolic distance function (TSDF) is parametric and often used to represent 3D structures that facilitate neural network computation, as well as computer storage. It can also play an important role in 3D reconstruction. This study is an inspiration for the development of large-scale 3D reconstruction, and also greatly improves the speed and quality of

DOI: 10.54254/2755-2721/45/20241016

3D reconstruction. In addition, BlockNeRF [13], which is an improved method for the 3D reconstruction of large outdoor scenes, is mainly used for large outdoor scenes.

4. NeRF's generalizability problem and solution

NeRF methods need to be retrained for a new scenario and cannot be directly extended to scenarios that have not been seen before, which contradicts one's goal of pursuing generalizability.

Trevithick, A and Yang B proposed a new solution called General Radiance Field (GRF) to this problem [14]. They designed a new neural network, the main working mechanism of which is to build an internal representation of each point in three-dimensional space by using a set of two-dimensional images with camera poses and intrinsic features as input, and then rendering the corresponding appearance and geometry of the point from any position to model the three-dimensional geometry as a general radiance field. This method can produce high-quality and realistic new views for new objects, new categories, and challenging real-world scenarios.

Other methods, such as IBRnet [15] and pixelNeRF, have similar core ideas to GRF, combining convolution with NeRF. However, this kind of generalization is still relatively preliminary, cannot achieve ideal results in complex scenes, and requires more in-depth research and improvement.

5. Representation of dynamic scenes

The NeRF method only considers static scenes and cannot be extended to dynamic scenes. However, the application scenarios of 3D reconstruction for dynamic scenes are very rich in reality, and the application of NeRF to dynamic scenes is also a popular direction of current research. The main objective of the research in this direction is to obtain a method that allows fast and realistic rendering of videos.

The original research selection methods were based on implicit neural representations to optimize scene representations using body rendering techniques to produce free point-of-view videos. D-NeRF recovered motions of dynamic scenes using implicit neural representations to achieve photo-level realistic rendering [16]. However, this class of methods is difficult to recover motions of complex scenes, and they take anywhere from a few hours to several days to train a model. Moreover, rendering an image takes minutes. Image-based rendering techniques can have better results compared to implicit neural representation, as it renders dynamic scenes by pre-training the model to avoid retraining at every moment. IBRNet can treat each frame as a separate scene, thus eliminating the need to recover the motions of the scene [15]. However, IBRNet still needs a few minutes to render an image.

6. Results

By changing the rendering method and other operations on NeRF, the rendering speed of images can be significantly improved, and the integration of NeRF with generalized deep neural networks, TSDF, and other technologies can enable the 3D reconstruction of large-scale scenes. With improvements, the current techniques can increase the rendering efficiency up to 10 times more than the original one. In terms of the generalizability of NeRF, there is no good solution to enable it to extend the rendering model to large-scale scenes, and more relevant research is needed. The authors believe that it can be combined with a Generative adversarial network (GAN) to try to generalize its functionality to more scenes.

Since dynamic scenes exist in many different scenes and need to be rendered one by one, the current NeRF improvement solutions are poor for rendering them, and further work is urgently needed.

7. Conclusion

NeRF has a broad application space in reverse rendering, controllable editing, digital human body, multi-modality, and image processing. However, it still has many shortcomings and can be optimized in many ways. One of the obvious aspects of NeRF is that the exploration of low-level semantics, such as denoising, image recovery, etc., is not particularly well-developed. The author believes that NeRF has not yet had a relatively successful commercial application, and many of the excellent effects in the paper

DOI: 10.54254/2755-2721/45/20241016

cannot be implemented in reality. Future research can focus on combining technology with practical applications to promote the implementation of technology. The main direction can include modeling the human body, especially the human face, and conducting research on multi-modality, such as the conversion between text and images, just like what CLIP-NeRF [17] does.

References

[1] Lin CH, Kong C, Lucey S. Learning efficient point cloud generation for dense 3d object reconstruction. *InProceedings of the AAAI Conf. on Artificial Intelligence 2018 Apr 27* (Vol. 32, No. 1).

[2] Mi Q, Gao T. 3D reconstruction based on the depth image: A review. *InInnovative Mobile and Internet Services in Ubiquitous Computing: Proceedings of the 16th Int. Conf. on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS-2022) 2022 Jun 16* (pp. 172-183). Cham: Springer International Publishing.

[3] Xie Y, Takikawa T, Saito S, Litany O, Yan S, Khan N, Tombari F, Tompkin J, Sitzmann V, Sridhar S. Neural fields in visual computing and beyond. *InComputer Graphics Forum 2022 May* (Vol. 41, No. 2, pp. 641-676).

[4] Drebin RA, Carpenter L, Hanrahan P. Volume rendering. *ACM Siggraph Computer Graphics. 1988 Jun 1*;22(4):65-74.

[5] Tewari A, Fried O, Thies J, Sitzmann V, Lombardi S, Sunkavalli K, Martin‐Brualla R, Simon T, Saragih J, Nießner M, Pandey R. State of the art on neural rendering. *InComputer Graphics Forum 2020 May* (Vol. 39, No. 2, pp. 701-727).

[6] Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*. 2021 Dec 17;65(1):99-106.

[7] Garbin SJ, Kowalski M, Johnson M, Shotton J, Valentin J. Fastnerf: High-fidelity neural rendering at 200fps. *InProceedings of the IEEE/CVF Int. Conf. on Computer Vision 2021* (pp. 14346-14355).

[8] Barron JT, Mildenhall B, Tancik M, Hedman P, Martin-Brualla R, Srinivasan PP. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *InProceedings of the IEEE/CVF Int. Conf. on Computer Vision 2021* (pp. 5855-5864).

[9] Deng K, Liu A, Zhu JY, Ramanan D. Depth-supervised nerf: Fewer views and faster training for free. *InProceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition 2022* (pp. 12882-12891).

[10] Yu A, Ye V, Tancik M, Kanazawa A. pixelnerf: Neural radiance fields from one or few images. *InProceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition 2021* (pp. 4578-4587).

[11] Chen A, Xu Z, Zhao F, Zhang X, Xiang F, Yu J, Su H. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. *InProceedings of the IEEE/CVF Int. Conf. on Computer Vision 2021* (pp. 14124-14133).

[12] Zhang X, Bi S, Sunkavalli K, Su H, Xu Z. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. *InProceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition 2022* (pp. 5449-5458).

[13] Tancik M, Casser V, Yan X, Pradhan S, Mildenhall B, Srinivasan PP, Barron JT, Kretzschmar H. Block-nerf: Scalable large scene neural view synthesis. *InProceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition 2022* (pp. 8248-8258).

[14] Trevithick A, Yang B. Grf: Learning a general radiance field for 3d representation and rendering. *InProceedings of the IEEE/CVF Int. Conf. on Computer Vision 2021* (pp. 15182-15192)..

[15] Wang Q, Wang Z, Genova K, Srinivasan PP, Zhou H, Barron JT, Martin-Brualla R, Snavely N, Funkhouser T. Ibrnet: Learning multi-view image-based rendering. *InProceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition 2021* (pp. 4690-4699).

DOI: 10.54254/2755-2721/45/20241016

[16] Pumarola A, Corona E, Pons-Moll G, Moreno-Noguer F. D-nerf: Neural radiance fields for dynamic scenes. *InProceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition 2021* (pp. 10318-10327).

[17] Wang C, Chai M, He M, Chen D, Liao J. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. *InProceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition 2022* (pp. 3835-3844).