Notes: This article is only to showcase my preliminary research plan and my interest and research in the field of 3D reconstruction to you. I would prefer to be involved in practical projects, rather than just doing well on public datasets. If you have a suitable project, I can modify some of the research plan content.

Neural Radiation Field for One or Several Images Based on NeRF

Zisong Wang, University of Chinese Academy of Sciences
E-mail: zswanggogo@hotmail.com

Abstract

Currently, the use of Neural Radiance Fields (NeRF) for implicit three-dimensional (3D) reconstruction has shown promising results, surpassing traditional geometric reconstruction methods in small-scale scenes. However, NeRF still has some research gaps, particularly in terms of its generalization capabilities. The NeRF network cannot directly extend to unseen scenes. This paper primarily discusses the strengths and limitations of traditional geometric methods and NeRF-based deep learning methods for 3D reconstruction. Additionally, recent works addressing the research gaps of NeRF, specifically in large-scale scenes, dynamic scenes, training acceleration, and generalization abilities, are introduced. Finally, I briefly introduced the research content and plan that I expected during my doctoral studies.

Introduction

The current objective of image-based 3D reconstruction is to infer the 3D geometry of objects within a scene using multiple 2D images and generate novel viewpoints in the form of 2D images. In earlier research, geometric methods were commonly employed, utilizing images captured by calibrated cameras and mathematical projections from 3D to 2D to address the ill-posed inverse problem. With the emergence of NeRF [1], 3D models can be generated with higher quality, including more detailed object representations and lighting information. Moreover, NeRF achieves this without requiring sensor data and solely relies on 2D posed images as supervision, implicitly representing complex 3D scenes. Currently, in small-scale scenes, NeRF-based approaches have gradually surpassed geometric methods.

In my opinion, there are three research gaps in NeRF-based approaches:

1, In large-scale scenes, NeRF lacks effective solutions due to factors such as network parameterization, image lighting, and distortion. Some recent works address this by employing scene segmentation and fusion techniques, such as BlockNeRF [2].

2, The training and inference costs of NeRF are high, requiring significant

computational resources and time. Several methods have been proposed to accelerate NeRF, for example, FastNeRF [3] and R2L [4].

3, NeRF exhibits poor generalization capabilities and requires retraining for each new scene. The network tends to overfit to specific scenes and cannot directly extend to unseen scenes. When only a few viewpoints are available (typically one or a few images), the random sampling nature of NeRF makes it challenging to obtain a continuous radiance field through training. Currently, research on NeRF generalization is limited, mainly focusing on generating object images for 3D reconstruction, but achieving ideal results for complex scene objects remains challenging. Representative works in this area include pixelNeRF [5], IBRNet [10], MVSNeRF [11], and NeRFDiff [6].

Regarding the three research gaps mentioned above, I aspire to conduct more in-depth research on the generalization of NeRF during my doctoral studies. The preliminary research question I intend to address is the construction of a neural radiance field for one or several images based on NeRF. This research primarily aims to enhance performance in two key aspects:

1, Improving Model Clarity: The first aspect involves enhancing the clarity of the generated model, which entails elevating three evaluation metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

2, Generalization to Complex Scenes: The second aspect pertains to extending the current generalization capabilities. Currently, generalization efforts mostly focus on utilizing the generated model for three-dimensional reconstruction of individual objects. I aim to ensure that my model can also achieve favorable results when dealing with scenarios involving multiple mutually occluding objects or when background information is present. This broader applicability of my research will make it more practical and relevant for real-world projects.

Review

The current research on 3D reconstruction can be categorized into two main areas: traditional geometric-based methods and deep learning-based methods. Traditional methods, based on visual geometry, are relatively mature. On the other hand, deep learning-based methods, especially those utilizing NeRF, have shown promising results.

Traditional 3D reconstruction methods can be classified as active and passive methods. In active methods, depth information of objects is known. Common approaches include: 1, Moiré fringe method [7], which uses interference images to reconstruct 3D contours. However, this method requires objects with regular textures. 2, Structured light methods, which project modulated light patterns onto the scene and capture the deformed patterns

to calculate depth information. This method is widely used, such as in Microsoft Kinect's depth camera, based on structured light principles. However, it is influenced by strong lighting conditions and projection distance.

Passive methods primarily utilize multi-view stereo (MVS) [8] techniques. MVS estimates the depth information of each pixel by matching pixel correspondences across multiple viewpoints, resulting in high-resolution 3D models. It is important to note that before the reconstruction process, camera positions and 3D points need to be determined through Structure from Motion (SfM) [12] algorithms to establish the relationship between world points and image points.

With the success of NeRF [1] in the field of 3D reconstruction using deep learning, there has been a surge of research based on NeRF in the past two years. NeRF can be summarized as using a Multilayer Perceptron (MLP) neural network to implicitly learn a static 3D scene. To train the network, a large number of images with known camera parameters are provided for a static scene. The trained neural network can then render images from arbitrary viewpoints. However, NeRF has several limitations. One limitation is its applicability primarily to small scenes, especially static scenes, with limited capability to handle dynamic scenes. Additionally, NeRF has high computational costs and suffers from overfitting, resulting in poor generalization. To overcome these limitations, researchers have proposed various solutions.

For addressing the challenge of large-scale scenes, a promising research approach is to partition a large scene into multiple blocks and train individual NeRFs for each block. Then, an aligning appearance method is used to fuse different scene information between adjacent NeRF blocks [2]. This strategy allows for the efficient handling of extensive scenes by breaking them down into manageable components and ensuring smooth transitions between them. When dealing with dynamic scenes, DynIBaR [9] provides a solution by employing a volume-based rendering framework. It aggregates features from nearby views in a scene-motion-aware manner to synthesize new views. This approach is particularly valuable for capturing the complexity of scenes with moving elements, ensuring that dynamic aspects are adequately represented.

To address the issue of high computational costs, there are two notable research works: FastNeRF [3] employs a factorized architecture that independently caches the position-dependent and ray direction-dependent outputs. This approach results in a remarkable 3000-fold speedup compared to the original NeRF model. However, it's worth noting that the storage requirements for FastNeRF become higher due to this factorization. R2L [4] addresses the computational cost challenge by concatenating multiple point coordinates sampled along a ray into a single vector, which is then used as input to the neural network. This strategy achieves a 30-fold acceleration while maintaining the advantage of small storage requirements.

In the realm of addressing the generalization challenge: PixelNeRF [5] takes spatial

image features aligned with each pixel as input. It leverages convolutional neural networks to extract image features and incorporates these features into the NeRF network's input. This enables the model to learn prior knowledge about the scene. Consequently, after a single training, it can generate new viewpoint images for unknown scenes with minimal input. IBRNet [10] tackles the task of synthesizing new views for complex scenes by interpolating a sparse set of nearby views. Specifically, the model learns a universal view interpolation function that can be applied to novel scenes. MVSNeRF [11] leverages cost volumes generated by sweeping planes through a scene for geometric scene understanding. It combines this geometric information with physics-based volume rendering to reconstruct the neural radiance field. These methods represent significant strides in improving the generalization capabilities of NeRF and its variants.

Methodology

At present, there are still many shortcomings in research on the generalization of NeRF. Apart from the accuracy of the generative model, the current main methods mainly focus on research using generated single objects. However, the performance of NeRF in complex or real-world scenes is not satisfactory. Therefore, the main research direction during the doctoral stage is to make improvements based on the NeRF network and achieve generalization work for slightly more complex real-world scenes, such as indoor scenes.

Currently, there are two publicly available datasets that can be used for experimental purposes. They are ShapeNet and ScanNet. ShapeNet is a synthetic dataset that includes over 50 semantic categories of objects, including furniture, vehicles, animals, food, etc. The shape of each object is stored in the form of a 3D mesh, with each mesh composed of hundreds to thousands of triangles. ScanNet is a large real RGB-D multimodal dataset that contains over 2.5 million indoor scene images, 1500 scenes, along with corresponding camera poses, mesh models, semantic labels, instance labels, and CAD models. Depth images are captured at a resolution of 640x480, and RGB images are captured at 1296x968 resolution. 3D models are reconstructed using BundleFusion. For this dataset, some indoor scene images with simple backgrounds will be selected for the final model validation and testing. Evaluation metrics will include Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

Below is my tentative schedule. I hope to complete the study of relevant basic theoretical knowledge in my first year of doctoral studies, including but not limited to papers on reproduction and some explanatory research on models. At the same time, I hope to complete the writing of a review paper on 3D reconstruction in the first grade. During this period of research, I hope to propose some possible effective innovative points and conduct experimental verification. In my second and third year as a doctoral student, I hope to basically complete the experimental part and complete the writing of

relevant papers in the second half of the year. During this period, I hope that my model can be applied in practical projects, rather than just performing well on public datasets. In my fourth and fifth year of doctoral studies (if there is a fifth year of doctoral studies), I hope to improve the entire experiment and ultimately complete my doctoral studies.

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] Tancik, Matthew, et al. "Block-nerf: Scalable large scene neural view synthesis." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[3] Garbin, S.J., Kowalski, M., Johnson, M., Shotton, J., Valentin, J.: Fastnerf: Highfidelity neural rendering at 200fps. arXiv preprint arXiv:2103.10380 (2021)

[4] Wang, Huan, et al. "R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

[5] Yu, Alex, et al. "pixelnerf: Neural radiance fields from one or few images." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[6] Gu, Jiatao, et al. "Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion." International Conference on Machine Learning. PMLR, 2023.

[7] Witkin, Andrew P. "Recovering surface shape and orientation from texture." Artificial intelligence 17.1-3 (1981): 17-45.

[8] Furukawa, Yasutaka, and Jean Ponce. "Accurate, dense, and robust multiview stereopsis." IEEE transactions on pattern analysis and machine intelligence 32.8 (2009): 1362-1376.

[9] Li, Zhengqi, et al. "Dynibar: Neural dynamic image-based rendering." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[10] Wang, Qianqian, et al. "Ibrnet: Learning multi-view image-based rendering." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[11] Chen, Anpei, et al. "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[12] Schonberger, Johannes L., and Jan-Michael Frahm. "Structure-from-motion revisited." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[13] Gao, Kyle, et al. "Nerf: Neural radiance field in 3d vision, a comprehensive review." arXiv preprint arXiv:2210.00379 (2022).

[14] Tewari, Ayush, et al. "Advances in neural rendering." Computer Graphics Forum. Vol. 41. No. 2. 2022.