PNP: A Review Framework towards Efficient NeRF Study and Research

Jin Qi Yeo[1,2], Aik Beng Ng[1], Kan Chen[2], Pallavi Mohan[1], Frank Guan[2], Simon See[1]

[1] NVIDIA Corporation
[2] Singapore Institute of Technology
{jinqiy, aikbengn, pamohan, ssee}@nvidia.com, {kan.chen, frank.guan}@singaporetech.edu.sg

Abstract

Rapid advancements in techniques and applications of NeRF technologies have been seen in both academia and industry recently. It becomes apparent that there is a need to consolidate and organize the theoretical and practical knowledge on the end to end pipeline of using NeRF technology. To facilitate the study and research on NeRF technologies, we propose a novel review framework that identifies the key stages for NeRF related technology, namely: Pre-NeRF, NeRF, Post-NeRF, or "PNP" in short. Each stage has four attributes: An overview of what and how NeRF technology is adapted at that stage, the common challenges faced, exemplary toolkits that are used to handle the tasks, and how it is applicable to the next stage in the framework.

Introduction

Photorealistic digital content is broadly applied in domains such as film and arts, gaming, marketing, robotics, simulations, communications, virtual reality (VR) and augmented reality (AR). However, it is computationally expensive and manually intensive to produce digital assets. High quality 3D content that is reusable with maximum consistency and quality assurance poses additional challenges whilst the demand continues to grow. A promising solution is to leverage deep learning techniques into the pipeline of traditional 3D content creation, which usually begins with data capturing for references to create 3D models and rendering them to be used in various applications.

Neural Radiance Fields (NeRF) (Mildenhall et al. 2020) has achieved state-of-the-art view synthesis quality which garnered much attention in the computer vision community in the recent years. By encoding the spatial positions and corresponding view directions of input images into the weights of a Multi-Layer Perceptron (MLP) network, NeRF learns how to generate photorealistic images for real-world scenes. With the showcase of many impressive demonstrations, it has also inspired subsequent research works on the methodologies and applications based on this novel technique. A high level perspective of the end to end pipeline involving the NeRF technique begins from using captured images as input to an MLP network to render photorealistic novel views for 3D viewing. The overlapping similarities between both pipelines evidently shows that NeRF is an ideal deep learning based representative solution.

Initially introduced in 2020, the NeRF technique has progressed with remarkable advancements in the research field and continues to offer promising prospects for future developments. Numerous research works have been conducted to further develop the NeRF technique towards 3D perception aimed at 1) transcending its capabilities and limitations (Gafni et al. 2021; Srinivasan et al. 2021; Yu et al. 2021; Bergen and Adelson 1991; Barron et al. 2021), and 2) integrating it into project frameworks that solves real world industry use cases (Adamkiewicz et al. 2022; Ichnowski et al. 2021). Furthermore, NeRF platforms in the form of a mobile app[1], standalone SDK with GUI (Müller et al. 2022; Tancik et al. 2023; Perlman 2023) and professional software (Volinga 2023) have emerged in recent months, offering convenient and accessible use for research purposes and content creation for non-scientists.

It becomes apparent that with such rapid advancements of the NeRF technique in academia and industry, there is a need to consolidate and organize the theory and practical knowledge of the end to end pipeline of using NeRF technology. However, there is an absence of survey papers that decomposes and describes the usage of NeRF technology effectively, targeted at assisting non-technical users and novice researchers. Hence, this paper presents the concept of our review framework, "PNP" in short, towards NeRF study and research as a foundation and opens room for exploration at the individual stages, namely: Pre-NeRF, NeRF, Post-NeRF. Each stage has four attributes: An overview of what and how NeRF technology is adapted at that stage, the common challenges faced, exemplary toolkits being used to handle the tasks, and how it is applicable to the next stage.

The next section introduces existing NeRF literatures which led to the emergence of NeRF platforms that generates NeRF renders seamlessly. Section 3 provides descriptions to each of the key stages of our framework represented with the four attributes. Lastly, a summary of this paper with discussions on future research directions.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] https://lumalabs.ai/

Related Works

Fundamentals of NeRF At a fundamental level, NeRF accepts 5D coordinates in the 3D space and the MLP network approximates the RGB and density values as radiance fields. In order to render a scene, for every pixel in an image, a target camera ray is projected into the 3D space and the samples along it are collected. Each sample is associated with a $\mathbf{p} = (x, y, z)$ position and its corresponding viewing direction $\mathbf{d} = (\theta, \phi)$ correlated to that camera ray. The acquired 5D values are inputs to the neural network, an Multilayer Perceptron (MLP), which predicts and outputs the corresponding colour $\mathbf{c} = (r, g, b)$ and density σ values which are then used to update the sample. This process is repeated for all the samples along that ray. These samples are composited together using volume rendering to compute a single color for that specific pixel. Eventually, the scene will be encoded within the weights of the trained NeRF model and is capable of rendering it from any viewpoint.

NeRF Literature Surveys One of the first survey papers on NeRF was published in December 2020 by Dellaert and Lin (Dellaert and Lin 2020), focused on the wide classification of NeRF research topics compiled from 50 preprints which were eventually published in top tier conferences. However, it does not provide much information on the process of training custom NeRF. Xie and team provided a comprehensive report in April 2022 (Xie et al. 2022) on the research directions of neural fields, providing background context and literature reviews from more than 250 papers. They proposed five categories of techniques used in neural fields: prior learning and conditioning, hybrid representations, forward maps, network architectures, manipulating neural fields, and also categorized the applications of neural fields in visual computing. Additionally, a web database collection[2] of NeRF research papers was developed and is maintained by the shared community, with functions such as search, filter of search results, a mindmap of the citations within a paper. Although this report is very comprehensive with illustrations to describe technical terminology, it is targeted at neural fields in general. Tewari et al. and team published a survey report in May 2022 on the state-of-the-art trends in neural rendering and some focus on the NeRF topic. Subsequently, Gao et al. (Gao et al. 2022) published a survey report fully focused on NeRF related research topics and proposed a taxonomy of the techniques in NeRF research papers, and a classification of NeRF applications. These survey reports provide very comprehensive knowledge on neural rendering, neural fields and recently, with more focus on NeRF topic. However, there is little focus on the pipeline of using NeRF to generate custom NeRF renders for individual or commercial use.

NeRF Platforms Documentations sometimes require certain level of technical knowledge to which can hinder the setup and installation for open source demonstrations. There are working projects published on Google Colab (Tancik et al. 2023) that allow interested users to get started with training and rendering custom NeRF models without the need for complex configuration setups. NeRF platforms, such as the LumaAI app[3] and professional software (Volinga 2023) have emerged, thus minimizing the complexity of manual setup and installation via the command line. The guide to using these is as simple as uploading a set of images to the NeRF platforms where the training will be done in the cloud to generate the final NeRF renders. However, they may require paid membership subscription with added features, and there are limitations when rendering real-world scenes due to expensive hardware.

Our Framework

With growing interest in the advancements of NeRF technology in academia and industry, it is essential to describe the usage and applications towards effectively using NeRF technology for study and research. We present our review framework that identifies the key stages for NeRF technology, namely: Pre-NeRF, NeRF, Post-NeRF, or "PNP" in short. As summarized in Figure1, there are four attributes to each stage, starting with an overview of what and how NeRF technology is adapted, followed by the general challenges and the common exemplary tools used by the community, and concluding with how the output in that stage is applicable and relevant to NeRF framework.

Pre-NeRF

There are varieties of existing datasets used in the experiments of NeRF research (Mildenhall et al. 2020, 2019; Knapitsch et al. 2017; Barron et al. 2022). Nonetheless, it will be most interesting if custom datasets (e.g., images, videos and 360° equilateral images or videos) can be used to train and render NeRF models.

Overview **Acquire datasets:** Capturing images for 3D reconstruction is an acquired skill, technically termed Photogrammetry, which plays a part in determining the render quality of the NeRF model. Currently, there is no set of hard rules on how to capture data that determines a good quality NeRF model. However, there are guidelines on the basics of photogrammetry[4, 5]. There are also experiences shared by users who have experimented with camera settings i.e., ISO, shutter speed, aperture, taking footages of outdoor environments[6] and indoor space[7]. A good rule of thumb is to take images with 2/3 of the scene overlapping each other, and the camera motion should create a trajectory path instead of pivoting at a point for video taking[8]. **Preprocess Images:** The acquired data needs to be converted into a format that the NeRF neural network can use to compute with, thus preprocessing of the images is required to recover the positions of

[2]https://neuralfields.cs.brown.edu/index.html

[3]https://lumalabs.ai/

[4]https://towardsdatascience.com/the-ultimate-guide-to-3d-reconstruction-with-photogrammetry-56155516ddc4

[5]https://colmap.github.io/tutorial.html

[6]https://neuralradiancefields.io/what-are-the-best-camera-settings-to-take-a-nerf/

[7]https://l1nq.com/straitstimes

[8]https://everypoint.io/wp-content/uploads/2022/09/How-to-Capture-Images-for-3D-Reconstruction.pdf

	Stages		
	PreNeRF	**NeRF**	**PostNeRF**
Overview	Data capturing of images, videos	Training of the AI model	Rendering the trained NeRF
Challenges	• To recover the 5D coordinates of the camera poses • What are the number of images required to achieve optimal 3D reconstruction? • What are the image resolutions? • Blurry images should be filtered out	To accelerate in training speed	• To render in real-time • How to utilize these trained NeRF (leads back to pre-trained NeRF)
Exemplary toolkits	COLMAP, KIRI Engine, Capturing Reality	InstantNGP, NeRF, NeRFStudio	Unreal, Blender, NVIDIA Omniverse, InstantNGP, NeRFStudio
Application	The 2D images and its corresponding camera poses are inputs to NeRF model	• Fundamentals of view-synthesis • Efficient, faster training/rendering speed • Higher quality images	• 3D contents in Asset store • City view maps • NeRF model with geometry background • Geometry with NeRF background

Figure 1: Summary of our proposed "PNP" framework

the images and the camera parameters. Data in video form must be extracted into image frames before they can be used as input to the MLP network. and this can be done with free media player software like VLC Media Player[9] and FFMPEG[10]. An additional step to ensure the quality of images would be to manually filter out the images that are blurry, or have too much overlap with the previous image.

Challenges Common issues with datasets are too few images, poor resolutions and blurriness in images, and inaccurate camera poses. It also depends on the image processing software used as there may be a learning curve and the processing speed to achieve camera parameters can be quite slow. Hence, this stage may pose as a bottleneck to some non-technical users and it is recommended to follow the guidelines of photogrammetry (e.g., no motion blurs, smooth connecting camera motions, no mislabelled camera data).

Exemplary Toolkits The conversion process to recover the 5D coordinates: $\mathbf{p}$, $\mathbf{d}$ is supported by photogrammetry software that considers the data type and the capturing device. Open source library COLMAP (Schonberger and Frahm 2016) is widely used in numerous NeRF literature as it supports images, videos and 360° data captured with any hardware. LumaAI[11] has its own image extraction pipeline to support its propriety website and mobile app. Alternatively, data captured with mobile phones can be extracted with mobile apps such as Polycam[12], KIRI Engine[13] and Record3D[14] which greatly simplify the process of image extraction to seamlessly import into NeRF platforms. Industry associates may utilize professional desktop-based software like Metashape[15] and Reality Capture[16] which require some level of expertise to use.

Applications Acquiring adequate number of images with optimal quality is necessary for the training of NeRF model in the next stage.

NeRF

The "NeRF" terminology is often vaguely used as a verb, a noun and an adjective in different circumstances, and in this case, it emphasizes on the techniques involved in the training of the NeRF network which can be modified to achieve different outputs.

Overview Given the images and their corresponding 5D coordinates, the next is to sample coordinates along the specified ray and feed them into the MLP network to produce RGB colour and density values. The density value predicted by the NeRF network at every sample point effectively provides valuable information on how much 3D data there is which can be used for occlusion or insertion of objects into the scene. Volume rendering is then used to composite these samples and render the colour pixel of that ray, which is used to compare to the ground truth pixel as a loss to update the NeRF network. Finally, this scene can be rendered anywhere as it is now stored as the weights of the neural network. **Image Quality Metrics:** Image quality degradation occurs during image acquisition and processing due to distortions i.e., noise, blurring, compression, ringing. Metrics used to evaluate and benchmark the qualitative comparisons of the original and reconstructed images for novel view synthesis via NeRF technique are widely adopted in most NeRF literatures (Mildenhall et al. 2020; Tancik et al. 2022; Hedman et al. 2021; Barron et al. 2022), including: Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) (Wang et al. 2004) and Learned Perceptual Image Patch (LPIPS) (Zhang et al. 2018). PSNR

[9]https://www.videolan.org/vlc/
[10]https://ffmpeg.org/
[11]https://lumalabs.ai/
[12]https://poly.cam/
[13]https://www.kiriengine.com/
[14]https://record3d.app/
[15]https://www.agisoft.com/
[16]https://www.capturingreality.com/

measures the mean-squared-error (MSE) of pixel values between two images: original and reconstructed. It is expressed in terms of logarithmic decibel scale because many signals have a very wide dynamic range. The higher the PSNR, the better the quality of the reconstructed image as the error is smaller. SSIM evaluates the similarity in structural information i.e., luminance, contrast, and structure between two images: the original and the reconstructed and provides a value of 0 to 1 where 1 indicates full similarity. LPIPS measures perceptual similarity using learned convolutional features trained on human perception.

Challenges In general, the NeRF model makes a lot of assumptions during training, causing floaters and artefacts in the rendered scenes. Lighting between input images is assumed to be unchanged. The ability to handle only static scenes results in dynamic objects, such as a person walking or a car driving by, being rendered into the scene. These assumptions are challenging to solve for real-world rendered scenes. Therefore, there are much ongoing research works to resolve these with a more robust NeRF model architecture.

Exemplary Toolkits An interactive GUI integrated with InstantNGP (Müller et al. 2022) features many controls to explore deeper with the neural graphics primitives, NeRF is one of them, that is being generated. It enables viewing the training in real-time, saving the NeRF as a .ingp file which can be reloaded from the saved point, rendering with DLSS mode to improve on the rendering quality. The VR mode, a recent addition, allows the user to experience and interact with the NeRF model while wearing a VR headset. Similarly, NeRFStudio (Tancik et al. 2023) allows training in real-time to be viewed on a web viewer locally and remotely. These GUIs minimize the complexity of using command lines for non-technical users to experience generating NeRF models. Alternatively, training NeRF in Python using command line are most commonly used.

Applications Tuning the NeRF network architecture with various techniques to best address the assumptions for NeRF, typically leads to improvements in the view synthesis, efficient and faster training/rendering speed, and better quality reconstructed images.

Post-NeRF

Overview Videos are the most common and best option for viewing the NeRF renders in high quality. Another option is to crop these NeRF models within a bounding box, render and export them as meshes or point clouds.

Challenges Other than accelerated training speed for NeRF models, rendering in real-time is another ongoing development that would be very captivating as it creates a great immersive experience for VRAR.

Exemplary Toolkits NeRF renders can be viewed as high quality videos on media players. The NeRF SDKs (Müller et al. 2022; Tancik et al. 2023) and mobile app[17] support camera settings such as adding of in-scene cameras, changing the focal lengths of each camera to create cinematic zoom effects, creating camera trajectory paths with varied speeds to create slow and fast motion effects. Furthermore, they have built-in extensions within their pipelines to export NeRF renders to 3D development engines like Blender[18], Unreal Engine[19] and NVIDIA Omniverse[20].

Applications Initially, view synthesis via NeRF technique were commonly used for 3D viewing of object centric scenes with small objects, which later developed into large scenes for indoor spaces (Wei et al. 2021) and outdoor environments (Tancik et al. 2022; Barron et al. 2022). It has expanded to practical uses in many industries. Commercial advertising utilized with NeRF technology captures the moment in time to promote products[21]. Consumers' purchasing experience is elevated when searching and viewing in e-stores for true-to-life products that are cost-efficient to produce, allowing them to make informed and strategic planning before deciding on the purchase[22]. NeRF renders can be integrated with virtual backgrounds and provide basic user interactions such as selection and editing (Jambon et al. 2023). Likewise, NeRF rendered scenes can be used as a background in virtual environments.

Future Discussions and Conclusion

The rapid advancements in both academia and industry have seen an emergence in the applications and techniques of the NeRF technology, and it becomes apparent that there is a need to consolidate and organize the theory and practical knowledge on the end to end pipeline of using NeRF technology. However, there is an absence of survey papers that decomposes and describes its usage effectively, targeted towards NeRF study and research in this topic of growing interest. Our review framework identifies three key stages and their attributes towards NeRF related technologies.

Still in its early research literature review stage, the modular concept of our "PNP" framework allows for exploration at individual stages. By adding text-conditioned diffusion models to the Pre-NeRF stage, generative AI with NeRF can be newly created with text prompts. In contrast to this, if an image-conditioned diffusion model is added to the Post-NeRF stage, it enables editing of the rendered NeRF with text prompt. The lighting and weather conditions of NeRF scene can be implemented and adjusted by adding exposure controls and appearance embeddings to the NeRF stage. The NeRF technique in general has several challenges, including the trade-offs in memory footprint, training and render speed, and the quality of NeRF renders. There are multiple concurrent works demonstrating various techniques to improve on these and taking a further step to utilize NeRF in the generative AI domain which has huge and exciting prospects to look forward to.

[17] https://lumalabs.ai/

[18] https://www.blender.org/

[19] https://www.unrealengine.com/en-US

[20] https://www.nvidia.com/en-us/omniverse/

[21] https://youtu.be/34KeBnSwvmc

[22] https://blog.google/products/shopping/search-on-2022-shopping/

References

Adamkiewicz, M.; Chen, T.; Caccavale, A.; Gardner, R.; Culbertson, P.; Bohg, J.; and Schwager, M. 2022. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2): 4606–4613.

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5855–5864.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *CVPR*.

Bergen, J. R.; and Adelson, E. H. 1991. The plenoptic function and the elements of early vision. *Computational models of visual processing*, 1: 8.

Dellaert, F.; and Lin, Y.-C. 2020. Neural volume rendering: Nerf and beyond. *arXiv preprint arXiv:2101.05204*.

Gafni, G.; Thies, J.; Zollhofer, M.; and Nießner, M. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8649–8658.

Gao, K.; Gao, Y.; He, H.; Lu, D.; Xu, L.; and Li, J. 2022. NeRF: Neural Radiance Field in 3D Vision, A Comprehensive Review. *arXiv preprint arXiv:2210.00379*.

Hedman, P.; Srinivasan, P. P.; Mildenhall, B.; Barron, J. T.; and Debevec, P. 2021. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5875–5884.

Ichnowski, J.; Avigal, Y.; Kerr, J.; and Goldberg, K. 2021. Dex-nerf: Using a neural radiance field to grasp transparent objects. *arXiv preprint arXiv:2110.14217*.

Jambon, C.; Kerbl, B.; Kopanas, G.; Diolatzis, S.; Drettakis, G.; and Leimkühler, T. 2023. NeRFshop: Interactive Editing of Neural Radiance Fields. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(1).

Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Trans. Graph.*, 36(4).

Mildenhall, B.; Srinivasan, P. P.; Ortiz-Cayon, R.; Kalantari, N. K.; Ramamoorthi, R.; Ng, R.; and Kar, A. 2019. Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines. *ACM Transactions on Graphics (TOG)*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 405–421. Springer.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4): 1–15.

Perlman, J. 2023. GitHub - JamesPerlman/TurboNeRF: A render engine for NeRFs! https://github.com/JamesPerlman/TurboNeRF. Accessed: 2023-05-19.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Srinivasan, P. P.; Deng, B.; Zhang, X.; Tancik, M.; Mildenhall, B.; and Barron, J. T. 2021. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7495–7504.

Tancik, M.; Casser, V.; Yan, X.; Pradhan, S.; Mildenhall, B.; Srinivasan, P. P.; Barron, J. T.; and Kretzschmar, H. 2022. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8248–8258.

Tancik, M.; Weber, E.; Ng, E.; Li, R.; Yi, B.; Kerr, J.; Wang, T.; Kristoffersen, A.; Austin, J.; Salahi, K.; et al. 2023. Nerfstudio: A modular framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*.

Volinga. 2023. Volinga Creator. https://volinga.ai/. Accessed: 2023-05-19.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wei, Y.; Liu, S.; Rao, Y.; Zhao, W.; Lu, J.; and Zhou, J. 2021. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5610–5619.

Xie, Y.; Takikawa, T.; Saito, S.; Litany, O.; Yan, S.; Khan, N.; Tombari, F.; Tompkin, J.; Sitzmann, V.; and Sridhar, S. 2022. Neural Fields in Visual Computing and Beyond. *Computer Graphics Forum*.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelNeRF: Neural Radiance Fields from One or Few Images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4576–4585.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.