Neural Rendering for 3D Reconstruction and View Synthesis: An Overview

Krish Rewanth Sevuga Perumal*
University of California, San Diego
ksevugaperumal@ucsd.edu

Manas Sharma*
University of California, San Diego
m7sharma@ucsd.edu

Ritika Kishore Kumar*
University of California, San Diego
rkishorekumar@ucsd.edu

Sanidhya Singal*
University of California, San Diego
ssingal@ucsd.edu

Abstract

Computer graphics and computer vision have experienced remarkable advancements in 3D reconstruction and novel view synthesis, primarily propelled by the emergence of neural networks. In this survey paper, we provide a comprehensive overview of the state-of-the-art techniques in neural rendering for 3D reconstruction and view synthesis. We discuss datasets and metrics for evaluation, trace the evolution from classical methods to neural networks, explore advancements in image rendering, and focus on Neural Radiance Fields (NeRF). We cover NeRF fundamentals, efficiency, sparse data handling, dynamic scenes, composition, and application-specific NeRFs. We conclude by summarizing key findings and identifying future research directions.

Keywords: Neural rendering, 3D reconstruction, View synthesis, Neural networks, NeRF, Image rendering.

1 Introduction

The field of computer graphics and computer vision has witnessed remarkable advancements in recent years, leading to significant progress in 3D reconstruction and novel view synthesis. These advancements have been driven by the emergence of neural networks, which have revolutionized the way we perceive and generate realistic 3D scenes.

The applications of neural rendering for 3D reconstruction and view synthesis have gained prominence due to their significant impact across various domains. For instance, in robotics, accurate perception of the environment is crucial for autonomous navigation and object manipulation. The ability to reconstruct 3D scenes and synthesize new viewpoints allows robots to understand their surroundings, aiding in tasks such as object recognition, scene understanding, and path planning. The work by McCormac et al. [34] demonstrates the importance of this field for robotic applications.

The paper by Sitzmann et al. [47] explores the use of neural networks in computer graphics rendering, showcasing their potential to generate realistic images and improve the visual quality of virtual environments. Similarly, in medical imaging [38], the reconstruction of detailed 3D anatomical structures from medical scans is essential for diagnosis, treatment planning, and surgical simulation. These applications, along with others such as virtual reality, gaming, and visual effects, have motivated researchers to continually improve upon traditional methods and introduce neural networks for enhanced reconstruction and rendering capabilities.

*All authors contributed equally to this work.

In this survey paper, we provide a comprehensive overview of the state-of-the-art techniques in neural rendering for 3D reconstruction and view synthesis. We begin by delving into the different types of datasets and metrics used in these fields in Section 2. Section 3 traces the evolution of 3D reconstruction methods, starting from classical approaches based on computer vision techniques. We explore the fundamental principles behind these methods and their limitations, setting the stage for the introduction of neural networks as a transformative paradigm for constructing 3D representations.

In Section 4, we embark on a journey through the evolution of image rendering techniques. We first explore classical rendering methods and their underlying principles. Subsequently, we delve into the groundbreaking concept of neural rendering, which harnesses the power of neural networks to generate highly realistic images. We highlight the advantages of neural rendering over classical techniques, such as improved visual quality, enhanced realism, and the ability to handle complex scenes.

Section 5 focuses specifically on Neural Radiance Fields (NeRF), an influential framework in neural rendering. We provide a concise overview of NeRF, discussing its fundamental concepts and architectures. Additionally, we explore its efficiency and scalability, addressing challenges associated with processing large-scale scenes. We also delve into techniques that enable NeRF to operate effectively with sparse or few-shot data, as well as its capability to handle dynamic scenes and enable scene composition and manipulation. Furthermore, we discuss the emergence of application-specific NeRFs, which tailor the framework to meet the requirements of specific domains.

Finally, in the Conclusion section, we summarize the key findings of this survey paper and highlight the most significant advancements in neural rendering for 3D reconstruction and view synthesis. We identify current challenges and potential future directions for research, opening avenues for further exploration and innovation in this exciting field.

Overall, this survey paper serves as a comprehensive resource that provides researchers and practitioners with a broad understanding of the state-of-the-art techniques in neural rendering for 3D reconstruction and view synthesis. By consolidating the existing knowledge and highlighting the latest advancements, we hope to inspire further research and foster continued progress in this rapidly evolving field.

2 Datasets and Metrics

2.1 Training Datasets

For the task of 3D reconstruction and view synthesis, large-scale datasets [61] with diverse 3D scenes are required. Computer vision practitioners use the following two common types of datasets:

- **Synthetic datasets:** These datasets are generated by rendering 3D scenes using computer graphics techniques. They typically consist of collections of 3D models with corresponding camera viewpoints and lighting conditions.
- **Real-world datasets:** These datasets are obtained by capturing real-world scenes using techniques like photogrammetry or LiDAR scanning. They contain densely sampled 3D point clouds or mesh representations, along with corresponding camera poses.

ShapeNet [3] is a synthetic dataset which is widely used. It involves aggregating 3D models from various online repositories, including websites like 3D Warehouse and TurboSquid. The collected models are then cleaned, normalized, and annotated with semantic information.

SceneNet [15] is a dataset developed by researchers from Stanford University. It involves creating synthetic indoor scenes using a game engine, specifically Unity3D. The scenes are designed to mimic real-world indoor environments and included furniture, props, and architectural elements. High-quality rendering techniques are employed to capture accurate lighting and material properties. Ground truth camera poses, surface normals, and semantic segmentations are obtained by offline rendering.

Another widely used dataset is **BlendedMVS** [65]. It aims to combine the benefits of ShapeNet and SceneNet by blending ShapeNet objects into SceneNet scenes. The blending process involves aligning the object geometry and textures with the scene geometry and textures.

Meanwhile in **SUNCG** [50], the scenes are created by using a room layout generation tool and populating the rooms with furniture and objects from a large furniture database. The scenes are then rendered using a physically based renderer to capture realistic lighting and material properties.

While coming to the real-world datasets, the **DTU Multi-View Stereo (MVS)** dataset [23] consists of densely captured 3D scenes using a calibrated camera array. It provides high-resolution images along with corresponding camera poses and depth maps. DTU MVS offers a diverse range of scenes, including both indoor and outdoor environments, making it suitable for training NeRF models to handle real-world scenes.

Another benchmark dataset is **Tanks and Temples** [26] which is specifically designed for evaluating 3D reconstruction and scene synthesis methods, including NeRF. It contains challenging real-world scenes captured using a handheld camera or drone. The dataset provides RGB images, camera poses, and sparse point clouds or depth maps for the scenes.

2.2 Evaluation Metrics

Evaluating the quality and accuracy of 3D reconstruction and view synthesis methods is crucial to assess their performance and effectiveness, and to account for improvements. To this end, various metrics have been developed to quantitatively measure the fidelity, alignment, and perceptual similarity between the synthesized or reconstructed results and the ground truth data. In this section, we discuss some commonly used evaluation metrics in the field of 3D reconstruction and view synthesis.

Mean Squared Error (MSE) [18] measures the average squared pixel-wise difference between the synthesized/reconstructed image and the ground truth image. It quantifies the overall discrepancy between the generated and reference images. **Root Mean Squared Error (RMSE)**, on the other hand, is the square root of MSE and provides a more interpretable measure of the average per-pixel error. Lower MSE and RMSE values indicate better alignment between the synthesized and the ground truth images.

Peak Signal-to-Noise Ratio (PSNR) is a widely used metric that measures the ratio between the maximum possible pixel intensity and the mean squared error. It provides a quantitative measure of the image fidelity. Higher PSNR values indicate better image quality and less distortion. However, PSNR is known to correlate poorly with human perception and may not capture all aspects of image quality.

Structural Similarity Index (SSIM) [58] is a comparison measure between the structural similarity of synthesized/reconstructed image and the ground truth image. It takes into account luminance, contrast, and structural information, providing a more comprehensive measure of image similarity. SSIM values range from 0 to 1, with higher values indicating better similarity. SSIM is more perceptually aligned compared to PSNR and is often considered a better metric for assessing image quality.

A less commlony used measure is **Frechet Inception Distance (FID)** [19]. It measures the quality and diversity of the synthesized images by computing the distance between feature embeddings of the generated images and the real images using an Inception Network. It considers both the distribution of features and the quality of individual images. Lower FID values indicate better image quality and diversity. FID is particularly useful when evaluating generative models, as it captures both visual fidelity and the diversity of generated samples.

These metrics provide quantitative measures to assess the quality, alignment, and perceptual fidelity of the output. Note that it is a common practice to employ multiple metrics in evaluation of a network due to the inherent limitations of individual metrics. By considering a combination of metrics, researchers can gain a more robust and nuanced understanding of the performance of their networks across different aspects of the reconstruction and synthesis process.

3 Advancements in 3D Reconstruction: From Classical Methods to Neural Networks

Construction of 3D representations from 2D images holds significant importance due to its applications in various domains, including computer vision, robotics, augmented reality, virtual reality, and medical imaging. By accurately reconstructing 3D geometry, it becomes possible to analyze, understand, and interact with the real world in a more immersive and detailed manner. Initially, classical methods based on computer vision techniques such as stereo matching and structure-from-motion were employed to perform 3D reconstruction. However, in recent times, there has been a remarkable shift in the approach towards utilizing deep learning and neural networks. Neural networks have shown immense potential in capturing complex patterns and relationships within images, leading to significant advancements in 3D reconstruction accuracy and efficiency. In this discussion, we explore both classical methods and deep learning-based methods for 3D reconstruction, highlighting their respective contributions and advancements.

3.1 Classical Methods based on Computer Vision Techniques

Classical 3D reconstruction techniques including Structure from Motion (SfM), stereo reconstruction, shape from shading, time of flight, and voxel-based methods, have been extensively used to capture 3D geometry and appearance from 2D observations.

SfM [17, 49] recovers the 3D structure of a scene by estimating camera poses and reconstructing sparse point clouds. One advantage of SfM is its ability to work with uncalibrated cameras, allowing for flexibility in camera setup. It is also a versatile technique that can handle different types of scenes. However, SfM is sensitive to camera motion and requires accurate feature matching across images, which can be challenging in certain scenarios.

Stereo reconstruction [46] infers depth information by analyzing pairs of images and computing disparity or depth maps. Stereo reconstruction provides accurate depth estimation, however, it requires a calibrated stereo camera setup, where precise knowledge of camera parameters is necessary. Additionally, stereo reconstruction is limited to pairwise images, which may restrict its applicability in dynamic or wide-ranging scenes.

Shape from shading techniques [20] estimate surface normals and 3D shape by analyzing shading patterns in images. One advantage of shape from shading is its ability to recover fine surface details, making it valuable in generating realistic renderings. However, shape from shading is highly sensitive to lighting conditions, and accurate estimation can be challenging when dealing with complex scenes or non-uniform illumination. It also has limitations in handling shiny or reflective surfaces.

Time-of-flight [16] techniques utilize sensors to measure the time taken for light to travel, enabling depth estimation and 3D scene reconstruction. These methods offer fast depth acquisition and real-time capabilities, making them suitable for applications such as gesture recognition, robotics, and virtual reality. However, time-of-flight techniques be susceptible to ambient light interference, which may affect their performance in certain environments.

Voxel-based methods [7] represent 3D scenes by dividing space into voxels and estimating their occupancy or attributes based on input data. Voxel-based representations are efficient for large-scale scenes and allow for volumetric data analysis. However, they can be memory-intensive, requiring substantial computational resources. Voxel-based methods may also have limitations in terms of spatial resolution, particularly when dealing with fine details or complex geometries.

Recently, **COLMAP** (Structure-from-Motion and Multi-View Stereo) [45] has gained popularity for constructing 3D representations. COLMAP is a software package that combines both SfM and MVS techniques for 3D reconstruction. It takes a collection of 2D images as input and estimates camera poses, sparse point clouds, and dense depth maps. Its popularity can be attributed to its robustness, efficiency, and ease of use, making it a valuable tool for researchers and practitioners in the field of computer vision, especially 3D reconstruction.

While classical 3D reconstruction techniques have been extensively used in diverse domains such as 3D modeling, augmented reality, robotics, medical imaging, and computer graphics, they have

certain limitations. These methods often rely on strong assumptions which can restrict their applicability in real-world scenarios. Moreover, they may struggle with handling occlusions, textureless regions, or complex scenes with large-scale structures.

To overcome these limitations, researchers have turned to neural networks as a powerful tool in the field of 3D reconstruction. By leveraging large-scale datasets and advanced learning algorithms, neural networks can learn complex representations and directly infer 3D geometry from 2D images. Deep learning approaches such as convolutional neural networks (CNNs) and generative adversarial networks (GANs), have shown promise in tasks like depth estimation, 3D shape reconstruction, and scene understanding. These neural network-based methods excel at handling challenging scenes, generalizing across different conditions, and capturing fine-grained details.

3.2 Neural Networks for Constructing 3D Representations

The last decade has seen a substantial increase in the utilization of neural networks for 3D reconstruction. Authors of ShapeNet [3] argue that dataset-driven approaches are essential for developing novel architectures. Accordingly, they introduce the ShapeNet dataset with several innovative features, including the integration of semantic and geometric annotations for each 3D model. Leveraging this dataset, many influential papers have made significant contributions to the field. For instance, Wang et al. (2018) proposed Pixel2Mesh [55], a method for generating 3D mesh models from single RGB images. Dosovitskiy et al. (2017) [9] presented a technique called Learning to Generate Chairs, Tables, and Cars with Convolutional Networks. Additionally, Maturana and Scherer (2015) introduced VoxNet [33], a real-time object recognition system based on a 3D Convolutional Neural Network.

Flynn et al. (2015) introduced **DeepStereo** [10], a method that leverages a large dataset of street view images to generate novel views of a scene. Unlike full 3D reconstruction, the primary focus of DeepStereo was on generating new views. The key advantage of this approach lies in training a neural network on an extensive amount of data, enabling the model to generalize effectively and produce accurate predictions for diverse scenes and viewpoints. In addition to view synthesis, the model also estimates the depth of the scene, which is vital for rendering the novel views with precision.

A significant contribution in the field was made by Choy et al. (2016) with the development of the unified framework, **3D-R2N2** [5]. This framework offers a comprehensive solution for both single and multi-view 3D reconstruction tasks. A key novelty of the paper lies in the introduction of a memory-efficient representation for 3D shapes using a 3D CNN architecture. This approach enables effective encoding and processing of volumetric data while reducing memory requirements compared to alternative methods. Notably, the focus of this work is on voxel-based representations rather than mesh models or point clouds, differentiating it from other approaches in the field.

Another noteworthy advancement in the field was by Wang et al. with their contribution, **Pixel2Mesh** [55]. In this paper, they leveraged both Graph Convolutional Networks (GCNs) and CNNs to predict a coarse 3D shape representation, followed by a graph-based optimization process for refinement to generate 3D mesh models from single RGB images. The paper presents a mesh refinement module that iteratively refines the generated mesh model. This helps to improve the quality and accuracy of the generated 3D meshes, enhancing the realism and detail of the final output.

The work by Sitzmann et al. in **DeepVoxels** [48] introduces DeepVoxels, a neural network architecture that predicts a 3D occupancy grid representation of a scene from 2D images. The authors introduce learning persistent 3D feature embeddings, which encode rich information about 3D shapes and structures. They are able to generate high quality reconstruction by combining the advantages of both voxel-based and mesh-based representations. Doing so enables the models to produce accurate reconstructions while preserving fine geometric details. It leverages training only the 2D images, which are more in abundance as compared to the annotated 3D datasets. The proposed representation also enables efficient rendering of 3D scenes by allowing the explicit modeling of occlusion and visibility. This allows for faster rendering compared to traditional voxel-based methods.

These aforementioned papers showcase diverse approaches to 3D reconstruction, highlighting the advancements made in terms of accuracy and efficiency. They tackle various aspects of the problem,

ranging from depth estimation and voxel-based representations to mesh generation. One common theme among these papers is the utilization of CNNs to learn the relationship between visual features and depth information. The ShapeNet dataset played a crucial role in training deep CNNs with high-quality 3D data, enabling researchers to explore novel architectures and techniques. The combination of different neural network architectures, such as CNNs, GCNs, and memory-efficient models, has pushed the boundaries of 3D reconstruction capabilities. These advancements pave the way for further applications, such as rendering novel views, which can benefit from the accurate and detailed 3D representations generated by these methods.

4 Evolution in Image Rendering Techniques

Rendering views from 3D models involves generating 2D images or animations that depict a specific viewpoint of a 3D scene. This process finds applications in varied fields, including computer graphics, virtual reality, gaming, architectural visualization, and product design. Initially, classical rendering methods were employed to achieve this task. These methods used various algorithms and techniques, such as wireframe, ray tracing, and rasterization, to simulate the interaction of light with the 3D objects and produce high-quality renderings. However, these approaches often required extensive computation and were time-consuming.

In recent years, there has been a shift towards leveraging neural networks for image synthesis from 3D representations. One approach involves creating a 3D representation from a set of 2D images, where the neural network learns to infer the underlying 3D structure. Once the 3D representation is obtained, another network can synthesize new views or images from different perspectives using the inferred 3D information. To further streamline the process, researchers have developed single end-to-end networks that can directly take 2D inputs and generate synthesized views. This integration of neural networks into the rendering pipeline has significantly enhanced the efficiency and quality of image synthesis from 3D representations.

4.1 Classical Rendering Methods

Classical rendering methods [52] have evolved over the years to generate visually appealing and realistic 2D images or animations from 3D scenes. Early techniques such as wireframe rendering [11] provide basic visual representations by depicting objects using lines and edges, but they lack surface details and a realistic appearance. However, as advancements in rendering techniques unfolded, surface rendering methods such as Gouraud shading [14] and Phong shading [41] emerged, revolutionizing the quality of rendered images.

Gouraud Shading [14] proposed by Henri Gouraud, aimed to achieve smooth shading on curved surfaces. This technique involves determining vertex normals which represent the surface orientation at each vertex of a polygonal mesh. By interpolating the colors across the vertices using these normals, Gouraud shading produces an approximation of the shading across the entire surface, resulting in a smoother appearance.

Phong Shading [41] presents an improved illumination model for computer-generated images. This technique considers ambient, diffuse, and specular reflection components to compute shading. By utilizing the surface normal, light sources, and viewer position, Phong shading determines the intensity of light reflected from the surface. Through the use of mathematical formulae, Phong shading achieves more realistic lighting effects in rendered images, enhancing the overall visual quality.

As the field progressed, **Ray Casting** technique gained prominence [59]. This method involves casting rays from the viewpoint through each pixel on the screen and determining the intersection with the 3D scene to determine the pixel's color. Ray casting considers complex lighting effects, shadows, and reflections, resulting in more accurate and visually appealing renderings.

Further, the introduction of the **Radiosity** method [6] addressed the challenge of simulating global illumination caused by indirect lighting effects. This technique aims to capture the inter-reflection of light between surfaces. By dividing the scene into small patches and placing a virtual cube known as a hemi-cube, around each patch, the Radiosity method accurately models complex lighting

interactions. Through the calculation of form factors and solving the radiosity equation, this method provides realistic renderings of scenes with intricate lighting conditions.

In recent years, with advancements in computer graphics, physically-based rendering techniques such as **Path Tracing** and **Bidirectional Path Tracing** have gained attention [30, 48]. These techniques utilize ray tracing algorithms to simulate the behavior of light in a more realistic manner. They account for effects like refractions, caustics, and subsurface scattering, resulting in highly accurate and visually impressive renderings.

4.2 Neural Rendering

Neural rendering represents a breakthrough in the field of 3D reconstruction, utilizing deep learning algorithms to generate realistic renderings directly from 3D data. Compared to classical rendering methods, neural rendering offers significant improvements in capturing complex lighting and shading effects, as well as producing high-fidelity reconstructions. Traditional methods often struggle with handling novel viewpoints and lighting conditions, while neural rendering excels in generalizing to unseen scenarios. By training on large datasets of real-world images and corresponding 3D models, neural networks learn to understand intricate lighting phenomena and subtle surface details, enabling them to produce visually stunning and realistic renderings with exceptional accuracy. These networks are designed to inherently learn and understand the underlying 3D structure from the 2D images, enabling the generation of novel views without explicitly constructing a 3D model.

RenderNet proposed by Thu Nguyen-Phuoc et al. in 2019 [37] provides a technique to produce high quality 2D renderings. RenderNet employs multiple neural networks which include a geometry network and a shading network. The geometry network models the scene's geometry by processing the 3D coordinates and normals. The shading network focuses on modeling the appearance and material properties, capturing the lighting and shading effects in the rendering. The main novel aspect of RenderNet paper is that the convolutional networks are differentiable.

Neural Volumes by Lombardi et al. [31] further made huge advancements in the field of 3D reconstruction by combining volumetric representations and neural networks. The approach is a two step process: an encoder-decoder network which transforms input images into a 3D volume representation, and a differentiable ray-marching operation that enables an end-to-end training. This model exhibits a lot of robustness to input variations such as changes in viewpoint, lighting conditions, and object appearances. The neural network model presented in the paper demonstrates good generalization capabilities, meaning it can effectively reconstruct various objects from different categories without requiring specific fine-tuning.

SynSin proposed by Wiles et al. [60] makes 3D reconstructions possible with just a single 2D image. They introduce a novel differentiable point cloud renderer that is used to transform a latent 3D point cloud of features into the target view. The major advantage with the process is that a single image at test time is enough to produce a 3D representation in real time. The paper also introduces a few-shot learning extension to SynSin, enabling the generation of novel views from a single image with limited additional supervision.

In **Learning to Stylize Novel Views** proposed by Hsin-Ping Huang et al. in 2021 [22], the authors provide an end-to-end neural network for the task of stylizing or transforming input images into new artistic or aesthetic renderings. They first construct the point cloud by back-projecting the image features to the 3D space and then develop point cloud aggregation modules to gather the style information of the 3D scene. Finally, they project the transformed features to 2D space to obtain the novel views. The trained model demonstrates the ability to generalize to unseen scenes, which is another major advantage of the paper.

Overall, each paper contributes unique ideas and techniques for neural rendering. RenderNet introduced the concept of using differentiable convolutional networks to model geometry and shading in the rendering process. This allowed for more accurate and realistic renderings by capturing intricate lighting and shading effects. Lombardi et al.'s work combined volumetric representations and neural networks, demonstrating robustness and generalization capabilities. SynSin made a breakthrough by enabling 3D reconstruction from a single 2D image using a differentiable point cloud renderer, which significantly reduced the data requirements and allowed

for real-time generation of 3D representations. Learning to Stylize Novel Views focused on transforming input images into artistic renderings through an end-to-end neural network, showcasing the ability to generalize to unseen scenes.

These advancements have paved the way for the development of combined end-to-end networks for 3D reconstruction and view synthesis. Some notable networks are Occupancy Networks [35], DeepSDF [39] and Neural Radiance Fields (NeRF) [36]. NeRF has gained particular prominence among these approaches. NeRF stands out for its ability to model volumetric scene functions using deep learning. By learning the radiance at every point in the scene, NERF can generate high-fidelity renderings that capture intricate lighting, shading, and material properties. NeRF builds upon the advancements made by previous end-to-end networks and takes them further by explicitly modeling the radiance of the scene. This allows for the generation of photo-realistic renderings and opens up new possibilities in areas such as computer graphics, virtual reality, and augmented reality. In the following section, we will explore NeRF in more detail and delve into its underlying principles and applications in the field of neural rendering.

5 Neural Radiance Fields

Much of the breakthrough in the field of 3D reconstruction and novel view synthesis can be attributed to the seminal work of Mildenhall et al. who introduced **Neural Radiance Fields (NeRF)** [36] as an application of MLP-based scene representations to a single-scene. It achieved state-of-the-art visual quality, producing impressive demonstrations and inspiring many subsequent works. Over time, various implementations of NeRF have emerged [13, 53], each with its unique contributions. This section is dedicated to a few influential NeRF publications published in the top-tier computer vision conferences such as CVPR, ICCV, and ECCV. We present a taxonomy outlining these publications and their respective advancements. Note that each section is arranged chronologically along with critical analyses and possible future directions and applications.

5.1 A Brief Overview of NeRF

NeRF is a framework that represents a 3D scene as a radiance field approximated by a neural network. It maps a 5D input vector, comprising in-scene coordinates (x, y, z) and viewing angles (θ, ϕ), to a 4D output space (c, σ), where $c = (r, g, b)$ represents color and σ represents volume density. This can be mathematically expressed as $F(x, \theta, \phi) \rightarrow (c, \sigma)$, where F denotes the neural network modeling the radiance field. The neural network approximation of the function F in the NeRF model consists of two stages. In the first stage, the input x is processed to generate the volume density σ and a high-dimensional feature vector. In the second stage, the feature vector is concatenated with the viewing direction d and fed into an additional Multi-Layer Perceptron (MLP), which produces the color representation c.

5.2 Architectures and Fundamentals

This section focuses on the foundational principles and the core architecture of NeRF. In 2021, Barron et al. introduced **Mip-NeRF** [1] utilizing cone tracing instead of the traditional ray tracing. Here, a cone, approximated using a multivariate Gaussian, is casted from the camera center along the viewing direction to generate a pixel, resulting in much smoother lines than the baseline NeRF model. In Mip-NeRF, scenes are discretized into multiple 2D planes (mipmaps) along the viewing direction. This discretization significantly reduces the memory and computation burden, making it more scalable for large and complex scenes. Each mipmap is associated with its own radiance field and density, allowing for efficient and adaptive sampling. The resulting multiplane representation in Mip-NeRF enables accurate rendering of scenes with complex geometry while maintaining computational feasibility.

However, Mip-NeRF introduces a loss of continuity in the scene representation (due to the discretization into mipmaps), which results in potential artifacts and inaccuracies during rendering, especially in areas where mipmaps transition or overlap. Additionally, the layered nature of Mip-NeRF may lead to limited depth perception as compared to NeRF, as the discrete planes may not capture the same level of depth information as a continuous volume representation.

Neural Implicit Surfaces (NeuS) by Liu and Wang et al. (2021) [56] is a variation of the NeRF model that performs volume rendering using signed distance functions (SDF) to define scene geometries. NeuS replaces the density output with an MLP that directly outputs the signed distance function value. The signed distance function represents the distance between a point in space and the nearest surface, with positive values inside the surface and negative values outside. To construct the density $\rho(t)$ required for volume rendering, NeuS employs the sigmoid function $\phi(.)$ and its derivative $d\phi/dt$, which represents the logistic density distribution. The density is computed as $\rho(t) = \max(-\frac{d\phi}{dt}(\frac{f(r(t))}{\phi(f(r(t)))}, 0)$, where $f(r(t))$ represents the SDF value at the point $r(t)$ along the ray being traced. This formulation ensures that the density is positive and accounts for the surface geometry during volume rendering.

Several subsequent improvements were made on the NeuS model. **HFNeuS** [57] enhances the reconstruction quality by separating low-frequency details into a base SDF and high-frequency details into a displacement function. This separation allowed for better representation of fine surface details. **Geo-NeuS** [12] introduces multi-view constraints, including a multi-view geometry constraint supervised by sparse point clouds and a multi-view photometric consistency constraint, to further improve the accuracy of the SDF-based reconstruction.

Ref-NeRF by Verbin et al. (2021) [54] introduces modifications to the parameterization of NeRF radiance by utilizing a directionless MLP that outputs not only density but also diffuse color, specular color, roughness, and surface normal. The diffuse and specular colors are combined to create the final color representation. Ref-NeRF employs a parameterization technique for the directional vector using spherical harmonics, allowing for precise modeling of specular reflections based on the local normal vector. Through these advancements, Ref-NeRF outperforms benchmarked methods and excels in accurately representing specular reflections and highlights on various datasets, showcasing its ability to enhance the realism of rendered images with reflective surfaces.

However, Ref-Nerf has several shortcomings. First, it requires more computational resources as the process of evaluating the integrated directional encoding is slightly slower than a standard positional encoding. Additionally, calculating normal vectors by backpropagating through the spatial MLP takes about 25% more time compared to Mip-NeRF. Secondly, Ref-NeRF's way of representing outgoing light based on reflection direction doesn't consider how light reflects between surfaces or non-distant light sources. Because of this, Ref-NeRF doesn't show as much improvement compared to Mip-NeRF in situations involving these factors.

In 2022, Xu et al. proposed **PointNeRF** [63] which incorporates a feature point cloud as an intermediate stage in the rendering process. A CNN is employed to generate depth and surface probability from this point cloud. The utilization of the point cloud network enables efficient skipping of empty spaces, leading to notable improvements in rendering speed.

Deng et al. introduced **Depth-supervised NeRF (DS-NeRF)** [8] in 2022. It incorporates depth supervision from sparse 3D point clouds computed during standard structure-from-motion (SfM) pre-processing, which provides additional information about the scene geometry. The method combines color and depth supervision, utilizing volume rendering, photometric loss, and a KL divergence term. The NeRF architecture remains unchanged, but the implicit surface reconstruction loss enhances accuracy. It achieves faster training (2-3x speedup) and improved results, even with fewer training views. If external depth estimates are provided, the performance is further improved.

Note that DS-NeRF depends on the output of SfM pre-processing. A poor SfM fit can result in a bad depth measurement. Also, the authors assume that the uncertainty of depth is modeled by a Gaussian distribution, which is a simplified assumption in our opinion.

Overall, these methods contribute to the advancement of NeRF by addressing various aspects such as smoothness, scene representation, surface geometry, specular reflections, rendering efficiency and accuracy. While each method brings valuable improvements, they also come with their own limitations and considerations. Further research and exploration in these areas can lead to even more refined and versatile NeRF-based approaches.

5.3 Efficiency and Scalability

Here, we discuss papers exploring techniques to improve the speed, scalability, and efficiency of NeRF, such as optimization methods, acceleration structures, or parallelization strategies.

Neural Sparse Voxel Fields (NSVF) by Liu and Gu et al. (2020) [29] utilizes a sparse voxel grid representation, enabling efficient storage and processing. It obtains feature representations by interpolating learnable features stored at voxel vertices which are passed to an implicit neural network that models voxel occupancy and color. NSVF achieves view-dependent rendering by projecting rays into the sparse voxel grid. It supports high-resolution reconstruction through adaptive voxel grid refinement. It uses a sparse voxel intersection-based point sampling for rays, which is more efficient than the hierarchical approach of vanilla NeRF. NSVF is scalable for large-scale scenes using an octree-based structure and performs 10x faster than NeRF. However, storing feature vectors on a potentially dense voxel grid can require significant memory resources, particularly for very large-scale scenes. Additionally, the iterative re-sampling of smaller voxels may not fully capture intricate geometric features or subtle variations in the scene. Insufficient voxel resolution may lead to loss of details, while overly high resolution can increase computational demands and memory requirements.

Kilo-NeRF by Reiser et al. (2021) [43] addresses the scalability issue of NeRF by leveraging a hierarchical structure with thousands of tiny MLPs to model local regions of the scene. It achieves efficient rendering through adaptive sampling, reduces memory consumption by sharing MLP parameters, and increases expressiveness for capturing complex geometry. It also extends NeRF to handle dynamic scenes by incorporating a temporally varying translation field. By employing early ray termination and empty space skipping, it achieves real-time rendering (practical NVS) of bounded medium-sized scenes, utilizing only about 100MB of storage.

However, there exists a scope for a few improvements. Kilo-NeRF shares the assumption of a bounded scene similar to NeRF. Furthermore, it cannot perform real time rendering of unbounded scenes due to memory constraints and its shallow MLP architecture. Finally, there is a trade-off between the number of MLPs and the size of each MLP, which needs to be tuned carefully for an optimal speedup.

In 2022, Hu et al. proposed a new model called **EfficientNeRF** [21]. It introduced advancements over baseline NeRF, including improved training speed using a density voxel grid, valid and pivotal sampling, and caching the trained scene. By improving efficiency, EfficientNeRF enhances the practicality of NeRF, reducing training time by over 88% and reaching rendering speeds over 200 FPS. These advancements make NeRF more accessible for real-world applications, promoting its use in various tasks.

However, we note that while EfficientNeRF achieves faster training and rendering speeds, there may be a trade-off in terms of accuracy compared to the baseline NeRF or more computationally expensive alternatives. Furthermore, caching the trained scene in a NeRF tree could require significant memory resources, limiting the scalability of the method for larger or more complex scenes. Finally, the use of pivotal sampling during the fine stage means that points away from the pivot points are not considered, potentially leading to the loss of fine details or the omission of important scene information.

Overall, these papers contribute valuable ideas to address the limitations of NeRF and enhance its speed, scalability, and efficiency. While they bring notable improvements, there are still trade-offs to consider, such as potential accuracy loss, memory requirements, and limitations in handling unbounded scenes. Future research may focus on finding a balance between efficiency and accuracy while further addressing scalability challenges to make NeRF more practical for a wider range of applications.

5.4 Few Shot or Sparse Data

These papers address the challenges of training NeRF with limited or sparse views, including approaches for handling incomplete data.

In 2020, Yu et al. introduced **PixelNeRF** [66] that performs well with one or few input images. Additionally, it also eliminates the need of separate models for different scenes, uses camera

coordinates instead of world coordinates, and handles multiple input images using average pooling. It leverages pretrained CNN layers and bilinear interpolation for feature extraction, allowing it to handle limited data scenarios. These advancements enhance the flexibility, generalization, and efficiency of PixelNeRF compared to vanilla NeRF.

PixelNeRF also has its own limitations, such as slow rendering speed and the need for manual parameter adjustment. Similar to NeRF, the rendering process in PixelNeRF is sluggish and runtime increases proportionally with the number of input views. In contrast, alternative approaches allow for quicker rendering and manipulation by generating meshes, unlike representations based on NeRF. Additionally, the manual fine-tuning of ray sampling bounds and positional encoding scales continues to present a challenge.

Multi-View Stereo NeRF (MVSNeRF) [4] offers advancements in efficient and generalizable radiance field reconstruction. Proposed by Chen and Xu et al. in 2021, it reconstructs radiance fields for view synthesis using only three nearby input views, enabling faster reconstruction. By leveraging plane-swept cost volumes and combining them with physically based volume rendering, MVSNeRF achieves geometry-aware scene reasoning and realistic view synthesis results. It demonstrates the ability to generalize across scenes, even for indoor scenes different from the training data. Additionally, MVSNeRF can be fine-tuned with dense images for faster per-scene reconstruction and higher rendering quality compared to traditional NeRF methods. Overall, MVSNeRF significantly reduces training time compared to baseline NeRF, achieving comparable results within just 15 minutes of training compared to hours required by traditional NeRF.

Despite its advancements, the method of MVSNeRF relies on accurate camera calibration and dense multi-view stereo reconstruction, which can be challenging and time-consuming in practical scenarios. Additionally, MVSNeRF assumes static scenes and does not handle dynamic objects well. It can struggle with scenes that exhibit large-scale occlusions or complex geometry. Lastly, the efficiency gains achieved by MVSNeRF may come at the cost of reduced reconstruction quality compared to more computationally expensive alternatives.

To summarize, the papers discussed present methods that address the challenges of training NeRF with limited or sparse views, offering techniques to handle incomplete data and enhance the flexibility and efficiency of NeRF-based approaches. They provide insights into improving generalization, efficiency, and reconstruction quality. However, there are still challenges to overcome, such as slow rendering time, manual parameter tuning, and the assumptions made about scene dynamics and reconstruction requirements. We believe that most models have successfully achieved the objective of handling few-shot scenarios with a small number of views (ranging from 2 to 10). So, the future research can move towards enhancing the flexibility and practicality of NeRF-based methods to achieve real-time and higher quality rendering.

5.5 Dynamic Scenes

The following papers specifically target dynamic or time-varying scenes in NeRF, aiming to capture changes over time or handle objects with non-static properties.

Introduced by Pumarola et al. in 2021, **D-NeRF** [42] advances NeRF models for dynamic scenes by handling articulated objects, complex human body movements, and using a single monocular camera and time parameter. It introduces a canonical network and deformation network, enabling the modeling of scene displacement over time. D-NeRF employs ray casting for view synthesis, ensuring accurate rendering of dynamic scenes. Furthermore, it enforces spatial and temporal consistency constraints, maintaining visual coherence. The multi-level radiance prediction captures fine details and global scene properties, and implicit occlusion handling prevents incorrect blending of foreground and background objects. These advancements enable D-NeRF to model, render, and synthesize dynamic scenes realistically and efficiently.

However we must note that such NeRF-based models can be computationally expensive during both training and rendering. Furthermore, variations in motion, occlusions, drastic changes in illumination, or scene deformations can affect the model's performance and result in artifacts or inaccurate renderings. Specifically, if an object enters or leaves a scenes, it can affect the deformation network's ability to map it to the canonical configuration.

In 2022, Liu et al. presented **DeVRF** [28] as an advancement in modeling dynamic scenes, offering fast convergence and high fidelity. DeVRF introduces a deformable voxel representation for non-rigid scenes and employs a static to dynamic learning paradigm for efficient training. With a novel data capture setup, it learns the 3D canonical space from static images and the 4D voxel deformation field from a few-view dynamic sequence. DeVRF achieves a significant speedup of two orders of magnitude (100x faster) compared to previous state-of-the-art methods while maintaining high-quality results. Its evaluation on synthetic and real-world dynamic scenes demonstrates its effectiveness in capturing various types of deformations. DeVRF addresses the challenges of training a large-parameter representation and offers practical applications in virtual reality and tele-presence.

DeVRF has some potential shortcomings to consider. Firstly, the voxel-based representation used in DeVRF can lead to high memory requirements, making it less efficient for scenes with intricate details or long temporal sequences. This could limit its scalability and practicality in real-world applications. Secondly, the static to dynamic learning paradigm relies on the availability of multi-view static images and few-view dynamic sequences, which may not always be easily obtainable or applicable in all scenarios. This dependence on specific data availability can restrict the generalizability of DeVRF to a wider range of dynamic scenes.

Neural 3D Video Synthesis from Multi-view Video by Li et al. (2022) [27] introduces advancements in video synthesis, including depth-aware view synthesis for accurate geometry, temporal coherence for smooth sequences, fine-grained 3D modeling for realistic details, and end-to-end learning for efficiency. It leverages multi-view data, incorporating depth information to synthesize views. Temporal coherence ensures consistency across frames, reducing artifacts. Fine-grained 3D modeling captures realistic textures, lighting, and appearances. The approach enables direct video synthesis from multi-view input, eliminating intermediate representations and manual preprocessing. By combining these advancements, it achieves high-quality, dynamic, and visually consistent video synthesis from multi-view data.

Despite its advancements, the aforementioned method for novel video synthesis may have some limitations. Firstly, the method relies heavily on the availability of accurate depth information from multi-view videos. In scenarios where depth estimation is challenging or inaccurate, the synthesized views may suffer from geometric distortions or inconsistencies. Secondly, the fine-grained 3D modeling approach, while enhancing visual quality, may also increase the computational complexity and memory requirements, limiting its scalability for real-time or large-scale applications. Lastly, the quality and realism of the synthesized videos may still fall short of real-world footage, particularly in complex scenes with dynamic objects, challenging lighting conditions, or intricate textures.

The papers discussed in this section introduce advancements in NeRF-based methods to handle dynamic or time-varying scenes. These works offer valuable insights into modeling, rendering, and synthesizing dynamic scenes realistically and efficiently. However, challenges such as computational complexity, memory requirements, data availability, and limitations in handling complex scenes remain to be addressed. Future research may focus on overcoming these challenges to further enhance the flexibility and practicality of NeRF-based methods in capturing and synthesizing dynamic scenes.

5.6 Scene Composition and Manipulation

This section focuses on techniques for composing or manipulating NeRF scenes, including object insertion, removal, scene editing, or interactive scene generation.

NeRF in the Wild (NeRF-W) by Martin-Brualla and Radwan et al. (2020) [32] introduces advancements that enable robust and scalable scene reconstruction from unstructured photo collections. It handles large-scale datasets with diverse lighting conditions and dynamic objects. It incorporates view selection, pose estimation, and image-specific priors to improve scene reconstruction. The model represents the scene as a combination of shared and image-dependent elements, removing transient objects. The authors address per-image appearance variations and transient objects using per-image embeddings. NeRF in the Wild achieved high-quality results on the Phototourism dataset [24], demonstrating its effectiveness in handling real-world scenarios.

NeRF-W has several shortcomings. First, it inherits the high computational complexity of traditional NeRF models, which makes training and rendering large-scale datasets time-consuming. Secondly, the model's ability to render areas rarely seen in the input images, such as the ground or occluded regions, is limited. Additionally, NeRF in the Wild relies on the distribution of camera viewpoints in the photo collection, making it sensitive to uneven or sparse coverage, which can lead to incomplete scene reconstructions. While it extends NeRF to handle dynamic scenes, it may struggle to capture rapid or complex motions accurately. Lastly, the inclusion of view selection, pose estimation, and per-image embeddings adds complexity to the training process.

Cui et al. presented **Object NeRF** in 2021 [64] as a voxel-based approach that combines the power of Neural Radiance Fields with instance segmentation labels to model and manipulate objects within a scene. It utilizes separate NeRF models for objects and the scene, both conditioned on interpolated voxel features. The method incorporates segmentation labels through a mask loss term, enhancing the fidelity of object representations. By editing the objects, the authors obtain background information from the scene NeRF and apply user-defined manipulations to the object's colors and densities. The aggregated colors and densities are then rendered using a volume rendering function. The advantages of Object NeRF include its ability to model objects and scenes simultaneously, and its improved performance over baseline NeRF and Neural Sparse Voxel Fields [29].

However, the method of Object NeRF heavily relies on the network's spatial smoothness to render unseen textures under objects due to the limited availability of observations. This may result in less accurate and unrealistic rendering of complex scene details. Secondly, optimizing camera poses and ray directions to mitigate pose noise and rolling shutter artifacts in real-world data poses a challenge. Thirdly, the current framework does not fully integrate a scene lighting model, which could impact the realism and visual quality of the rendered scenes. Future work could address these limitations by incorporating scene completion methods, improving optimization techniques, and integrating a more realistic scene lighting model.

CoNeRF by Kania et al. (2022) [25] introduces a framework for fine-grained control over 3D scenes by annotating a small number of masks that indicate the desired control areas. These masks serve as latent variables and are learned by the neural network, either through provided annotations or automatic discovery when annotations are absent. By leveraging the concept of spatial quasi-conditional independence of attributes, CoNeRF enables localized control over various appearance attributes of the scene. This approach surpasses the limitations of coarse-grain controls, such as materials, colors, or object placement, by allowing specific attribute control without restrictions to particular object classes or properties. The framework employs a few-shot learning framework that combines ground truth information with sparse 2D mask annotations. These annotations specify the regions of the scene that correspond to each attribute. By treating the attributes as latent variables, the framework can automatically extend the mask annotations to the entire input video, facilitating comprehensive attribute control throughout the scene.

In a nutshell, the above methods offer valuable advancements in enabling robust and scalable scene reconstruction, object modeling and manipulation, and fine-grained control over 3D scenes. However, computational complexity, accuracy in rendering unseen textures, optimization challenges, and the integration of realistic scene lighting models are some areas identified for improvemnet.

5.7 Application-specific NeRFs

The following papers demonstrate NeRF-based methods or adaptations for specific applications, such as NeRF for complex and large-scale urban environments.

Animatable NeRF by Peng et al. (2021) [40] modifies existing NeRF methods to handle the reconstruction and animation of human models. While NeRF is originally designed for static 3D scenes, this work extends it to handle non-rigidly deforming scenes, specifically focusing on animatable humans. The modifications made to NeRF include the introduction of a new motion representation called the neural blend weight field. This representation combines 3D human skeletons with blend weight fields, which are learned in the canonical space. By incorporating the skeleton-driven deformation framework, the blend weight fields provide an effective regularization during the learning of deformation fields.

Unlike previous approaches that optimize NeRF jointly with translational vector fields or $SE(3)$ fields, this paper proposes a new approach where the human skeleton is easy to track and does not need to be jointly optimized. This provides a more efficient and effective regularization on the learning of deformation fields. Furthermore, an additional neural blend weight field is learned at the canonical space, enabling explicit animation of the neural radiance field with input motions. This allows for the synthesis of novel scenes given input motions for animation. Overall, this paper extends NeRF by introducing the neural blend weight field representation and leveraging the skeleton-driven deformation framework to handle the reconstruction and animation of animatable human models, overcoming the limitations of previous approaches in representing dynamic scenes and enabling explicit motion synthesis.

Urban Radiance Fields by Rematas et al. (2021) [44] addresses the challenge of representing complex urban environments by leveraging an efficient hierarchical data structure that can handle large-scale scenes. It incorporates a multi-level radiance modeling approach, which enables the capture of both global illumination effects and fine-grained details in urban scenes. It incorporates three key features in NeRF: the use of LiDAR data for depth information, separate treatment of sky pixels, and compensation for varying exposure through affine color transformation estimates for each camera. Urban Radiance Fields also introduce additional loss terms, including a LiDAR-based depth loss, sight loss to concentrate radiance at the surface, and a segmentation loss for sky pixels. Finally, it incorporates semantic knowledge, and captures and represents the distinct characteristics of different urban elements, leading to more realistic and contextually-aware scene synthesis.

Despite advancements in efficiency compared to traditional NeRF, Urban Radiance Fields still involves computationally intensive processes for volumetric rendering and 3D reconstruction. The complexity increases with the size and complexity of the urban scene, potentially limiting real-time or interactive applications. The performance of Urban Radiance Fields can be sensitive to the quality and accuracy of the input data, including the panorama images and LiDAR point clouds. Moreover, collecting data for urban scenes using street-level panoramas and LiDAR can be challenging due to various factors such as occlusions, dynamic objects, and varying lighting conditions. This can lead to incomplete or noisy data, impacting the quality of the scene reconstruction.

Street-view Neural Radiance Fields (S-NeRF) by Xie and Zhang et al. (2023) [62] is an architecture that aims to synthesize novel views of both large-scale background scenes and foreground moving vehicles jointly. The architecture comprises several key components and improvements. Firstly, the scene parameterization function and camera poses are enhanced to facilitate better learning of neural representations from street views. This improvement leads to more accurate and detailed representations of the environment. Secondly, S-NeRF leverages noisy and sparse LiDAR points during training to boost performance and address depth outliers effectively. By incorporating these points, the architecture learns a robust geometry and reprojection confidence, improving the overall accuracy of the synthesized scenes. One notable advantage of S-NeRF over other approaches is its capability to reconstruct moving vehicles. Unlike conventional NeRFs that struggle with capturing dynamic objects, S-NeRF extends the method to handle the rendering of moving vehicles in street scenes. This extension enables the generation of realistic and accurate representations of vehicles, enhancing the overall fidelity of the synthesized views. Furthermore, S-NeRF outperforms state-of-the-art rivals in terms of mean-squared error reduction in street-view synthesis. The experiments conducted on large-scale driving datasets such as nuScenes [2] and Waymo [51] demonstrate a significant improvement ranging from 7% to 40% reduction in mean-squared error. Additionally, S-NeRF achieves a 45% gain in peak signal-to-noise ratio (PSNR) for rendering moving vehicles. These advantages highlight S-NeRF's ability to produce high-quality street-view synthesis and render moving objects with improved accuracy compared to other existing methods.

Overall, the aforementioned methods showcase the versatility of NeRF-based approaches by adapting them to specific applications and addressing the challenges associated with each domain. They provide valuable contributions in terms of animating non-rigidly deforming scenes, representing complex urban environments, and synthesizing street views with moving objects. However, it is important to note that these methods may still have limitations, such as computational complexity, sensitivity to input data quality, and challenges in data collection. In the near future, more innovations in areas such as medical imaging, augmented reality, and robotics are expected as more computer vision practitioners adopt NeRF models.

6 Conclusion

This survey paper provides a comprehensive and in-depth exploration of the recent trends and state-of-the-art techniques in neural rendering for 3D reconstruction and novel view synthesis. We trace the evolutionary journey from classical methods to the emergence of neural networks as transformative tools in the field of computer graphics and computer vision. The advantages of neural rendering over traditional techniques, including improved visual quality, enhanced realism, and the ability to handle complex scenes, have been thoroughly highlighted.

We have provided a detailed examination of Neural Radiance Fields (NeRF), an influential framework, covering its fundamentals, efficiency, scalability, and innovative approaches for sparse or dynamic data handling. Furthermore, the discussion of application-specific NeRF variants has underscored the adaptability of neural rendering techniques to diverse domains and requirements.

Looking ahead, neural rendering remains an exciting and rapidly evolving field with various open challenges that warrant further research. We have identified and discussed multiple directions for future exploration, aiming to address these challenges and drive advancements.

References

[1] BARRON, J. T., MILDENHALL, B., TANCIK, M., HEDMAN, P., MARTIN-BRUALLA, R., AND SRINIVASAN, P. P. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 5835–5844.

[2] CAESAR, H., BANKITI, V., LANG, A. H., VORA, S., LIONG, V. E., XU, Q., KRISHNAN, A., PAN, Y., BALDAN, G., AND BEIJBOM, O. nuscenes: A multimodal dataset for autonomous driving, 2020.

[3] CHANG, A. X., FUNKHOUSER, T., GUIBAS, L., HANRAHAN, P., HUANG, Q., LI, Z., SAVARESE, S., SAVVA, M., SONG, S., SU, H., XIAO, J., YI, L., AND YU, F. Shapenet: An information-rich 3d model repository, 2015.

[4] CHEN, A., XU, Z., ZHAO, F., ZHANG, X., XIANG, F., YU, J., AND SU, H. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 14104–14113.

[5] CHOY, C. B., XU, D., GWAK, J., CHEN, K., AND SAVARESE, S. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction, 2016.

[6] COHEN, M. F., AND GREENBERG, D. P. A hemispherical representation for radiosity. *ACM SIGGRAPH Computer Graphics 20*, 4 (1986), 303–312.

[7] CURLESS, B., AND LEVOY, M. A volumetric method for building complex models from range images. *ACM Transactions on Graphics (TOG) 15*, 3 (1996), 303–334.

[8] DENG, K., LIU, A., ZHU, J.-Y., AND RAMANAN, D. Depth-supervised nerf: Fewer views and faster training for free. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 12872–12881.

[9] DOSOVITSKIY, A., SPRINGENBERG, J. T., TATARCHENKO, M., AND BROX, T. Learning to generate chairs, tables and cars with convolutional networks, 2017.

[10] FLYNN, J., NEULANDER, I., PHILBIN, J., AND SNAVELY, N. Deepstereo: Learning to predict new views from the world's imagery, 2015.

[11] FOLEY, J. D., VAN DAM, A., FEINER, S. K., AND HUGHES, J. F. *Computer graphics: principles and practice*. Addison-Wesley Professional, 1990.

[12] FU, Q., XU, Q., ONG, Y.-S., AND TAO, W. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction, 2022.

[13] GAO, K., GAO, Y., HE, H., LU, D., XU, L., AND LI, J. Nerf: Neural radiance field in 3d vision, a comprehensive review, 2023.

[14] GOURAUD, H. Continuous shading of curved surfaces. *IEEE Transactions on Computers 20*, 6 (1971), 623–629.

[15] HANDA, A., PATRAUCEAN, V., BADRINARAYANAN, V., STENT, S., AND CIPOLLA, R. Scenenet: Understanding real world indoor scenes with synthetic data, 2015.

[16] HANSARD, M., LEE, S. W., CHOI, O., AND CHRISTENSEN, H. I. Time-of-flight cameras: principles, methods and applications. *Springer Science+ Business Media 1*, 1 (2013), 3–56.

[17] HARTLEY, R., AND ZISSERMAN, A. *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[18] HECKBERT, P. S. Survey of interpolation methods. *ACM SIGGRAPH Computer Graphics 24*, 4 (1990), 291–342.

[19] HEUSEL, M., ET AL. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)* (2017).

[20] HORN, B. K. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. *Optical engineering 10*, 3 (1970), 319–322.

[21] HU, T., LIU, S., CHEN, Y., SHEN, T., AND JIA, J. Efficientnerf - efficient neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 12892–12901.

[22] HUANG, H.-P., TSENG, H.-Y., SAINI, S., SINGH, M., AND YANG, M.-H. Learning to stylize novel views, 2021.

[23] JENSEN, R., DAHL, A., VOGIATZIS, G., TOLA, E., AND AANÆS, H. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 406–413.

[24] JIN, Y., MISHKIN, D., MISHCHUK, A., MATAS, J., FUA, P., YI, K. M., AND TRULLS, E. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision 129*, 2 (Oct 2020), 517–547.

[25] KANIA, K., YI, K. M., KOWALSKI, M., TRZCIŃSKI, T., AND TAGLIASACCHI, A. CoNeRF: Controllable Neural Radiance Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2022).

[26] KNAPITSCH, A., PARK, J., ZHOU, Q.-Y., AND KOLTUN, V. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics 36*, 4 (2017).

[27] LI, T., SLAVCHEVA, M., ZOLLHOEFER, M., GREEN, S., LASSNER, C., KIM, C., SCHMIDT, T., LOVEGROVE, S., GOESELE, M., NEWCOMBE, R., AND LV, Z. Neural 3d video synthesis from multi-view video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 5511–5521.

[28] LIU, J.-W., CAO, Y.-P., MAO, W., ZHANG, W., ZHANG, D. J., KEPPO, J., SHAN, Y., QIE, X., AND SHOU, M. Z. Devrf: Fast deformable voxel radiance fields for dynamic scenes. *arXiv preprint arXiv:2205.15723* (2022).

[29] LIU, L., GU, J., ZAW LIN, K., CHUA, T.-S., AND THEOBALT, C. Neural sparse voxel fields. In *Advances in Neural Information Processing Systems* (2020), H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., pp. 15651–15663.

[30] LOMBARDI, S., LOMBARDI, A., SAJJADI, M., AND NOWOZIN, S. Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 5985–5994.

[31] LOMBARDI, S., SIMON, T., SARAGIH, J., SCHWARTZ, G., LEHRMANN, A., AND SHEIKH, Y. Neural volumes. *ACM Transactions on Graphics 38*, 4 (jul 2019), 1–14.

[32] MARTIN-BRUALLA, R., RADWAN, N., SAJJADI, M. S. M., BARRON, J. T., DOSOVITSKIY, A., AND DUCKWORTH, D. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 7206–7215.

[33] MATURANA, D., AND SCHERER, S. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2015), pp. 922–928.

[34] MCCORMAC, J., HANDA, A., DAVISON, A., AND LEUTENEGGER, S. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).

[35] MESCHEDER, L., OECHSLE, M., NIEMEYER, M., NOWOZIN, S., AND GEIGER, A. Occupancy networks: Learning 3d reconstruction in function space, 2019.

[36] MILDENHALL, B., SRINIVASAN, P. P., TANCIK, M., BARRON, J. T., RAMAMOORTHI, R., AND NG, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM 65*, 1 (dec 2021), 99–106.

[37] NGUYEN-PHUOC, T., LI, C., BALABAN, S., AND YANG, Y.-L. Rendernet: A deep convolutional network for differentiable rendering from 3d shapes, 2019.

[38] NIE, D., ZHANG, H., ADELI, E., AND LIU, L. 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. vol. 9901.

[39] PARK, J. J., FLORENCE, P., STRAUB, J., NEWCOMBE, R., AND LOVEGROVE, S. Deepsdf: Learning continuous signed distance functions for shape representation, 2019.

[40] PENG, S., DONG, J., WANG, Q., ZHANG, S., SHUAI, Q., ZHOU, X., AND BAO, H. Animatable neural radiance fields for modeling dynamic human bodies. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 14294–14303.

[41] PHONG, B. T. Illumination for computer generated pictures. *Communications of the ACM 18*, 6 (1975), 311–317.

[42] PUMAROLA, A., CORONA, E., PONS-MOLL, G., AND MORENO-NOGUER, F. D-nerf: Neural radiance fields for dynamic scenes. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 10313–10322.

[43] REISER, C., PENG, S., LIAO, Y., AND GEIGER, A. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 14315–14325.

[44] REMATAS, K., LIU, A., SRINIVASAN, P., BARRON, J., TAGLIASACCHI, A., FUNKHOUSER, T., AND FERRARI, V. Urban radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 12922–12932.

[45] SCHÖNBERGER, J. L., AND FRAHM, J.-M. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 4104–4113.

[46] SEITZ, S. M., CURLESS, B., DIEBEL, J., SCHARSTEIN, D., AND SZELISKI, R. A comparison and evaluation of multi-view stereo reconstruction algorithms. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2006), 519–528.

[47] SITZMANN, V., ZOLLHÖFER, M., DAVIDOVIC, D., FISHER, M., WANG, O., CHEN, W., AND WETZSTEIN, G. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[48] SITZMANN, V., ZOLLÖFER, M., AND WETZSTEIN, G. Deepvoxels: Learning persistent 3d feature embeddings. *ACM Transactions on Graphics (TOG) 38*, 4 (2019), 1–15.

[49] SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. Photo tourism: exploring photo collections in 3d. *ACM Transactions on Graphics (TOG) 25*, 3 (2006), 835–846.

[50] SONG, S., YU, F., ZENG, A., CHANG, A. X., SAVVA, M., AND FUNKHOUSER, T. Semantic scene completion from a single depth image, 2016.

[51] SUN, P., KRETZSCHMAR, H., DOTIWALLA, X., CHOUARD, A., PATNAIK, V., TSUI, P., GUO, J., ZHOU, Y., CHAI, Y., CAINE, B., VASUDEVAN, V., HAN, W., NGIAM, J., ZHAO, H., TIMOFEEV, A., ETTINGER, S., KRIVOKON, M., GAO, A., JOSHI, A., ZHANG, Y., SHLENS, J., CHEN, Z., AND ANGUELOV, D. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 2443–2451.

[52] TEWARI, A., FRIED, O., THIES, J., SITZMANN, V., LOMBARDI, S., SUNKAVALLI, K., MARTIN-BRUALLA, R., SIMON, T., SARAGIH, J., NIESSNER, M., PANDEY, R., FANELLO, S., WETZSTEIN, G., ZHU, J.-Y., THEOBALT, C., AGRAWALA, M., SHECHTMAN, E., GOLDMAN, D. B., AND ZOLLHÖFER, M. State of the art on neural rendering. *Computer Graphics Forum 39*, 2 (2020), 701–727.

[53] TEWARI, A., THIES, J., MILDENHALL, B., SRINIVASAN, P., TRETSCHK, E., YIFAN, W., LASSNER, C., SITZMANN, V., MARTIN-BRUALLA, R., LOMBARDI, S., SIMON, T., THEOBALT, C., NIESSNER, M., BARRON, J. T., WETZSTEIN, G., ZOLLHÖFER, M., AND GOLYANIK, V. Advances in neural rendering. *Computer Graphics Forum 41*, 2 (2022), 703–735.

[54] VERBIN, D., HEDMAN, P., MILDENHALL, B., ZICKLER, T., BARRON, J. T., AND SRINIVASAN, P. P. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 5481–5490.

[55] WANG, N., ZHANG, Y., LI, Z., FU, Y., LIU, W., AND JIANG, Y.-G. Pixel2mesh: Generating 3d mesh models from single rgb images, 2018.

[56] WANG, P., LIU, L., LIU, Y., THEOBALT, C., KOMURA, T., AND WANG, W. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS* (2021).

[57] WANG, Y., SKOROKHODOV, I., AND WONKA, P. Hf-neus: Improved surface reconstruction using high-frequency details, 2022.

[58] WANG, Z., ET AL. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (2004), 600–612.

[59] WHITTED, T. An improved illumination model for shaded display. *Communications of the ACM 23*, 6 (1980), 343–349.

[60] WILES, O., GKIOXARI, G., SZELISKI, R., AND JOHNSON, J. Synsin: End-to-end view synthesis from a single image, 2020.

[61] XIE, Y., TAKIKAWA, T., SAITO, S., LITANY, O., YAN, S., KHAN, N., TOMBARI, F., TOMPKIN, J., SITZMANN, V., AND SRIDHAR, S. Neural fields in visual computing and beyond. *Computer Graphics Forum 41*, 2 (2022), 641–676.

[62] XIE, Z., ZHANG, J., LI, W., ZHANG, F., AND ZHANG, L. S-nerf: Neural radiance fields for street views. In *ICLR 2023* (2023).

[63] XU, Q., XU, Z., PHILIP, J., BI, S., SHU, Z., SUNKAVALLI, K., AND NEUMANN, U. Point-nerf: Point-based neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 5428–5438.

[64] YANG, B., ZHANG, Y., XU, Y., LI, Y., ZHOU, H., BAO, H., ZHANG, G., AND CUI, Z. Learning object-compositional neural radiance field for editable scene rendering. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 13759–13768.

[65] YAO, Y., LUO, Z., LI, S., ZHANG, J., REN, Y., ZHOU, L., FANG, T., AND QUAN, L. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks, 2020.

[66] YU, A., YE, V., TANCIK, M., AND KANAZAWA, A. pixelnerf: Neural radiance fields from one or few images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 4576–4585.