Novel View Synthesis: A Review and Taxonomy Focused on Neural Radiance Fields (NeRF)

22281188 Jiawei Jiang[1]
[1]Beijing Jiaotong University;
*Student Email: 22281188@bjtu.edu.cn

Abstract

Novel View Synthesis (NVS), the generation of images from new viewpoints based on existing input views, is a cornerstone problem in computer vision and graphics with wide-ranging applications. This paper presents a comprehensive survey charting the evolution of NVS methodologies. This paper traces the historical progression through four distinct eras: early techniques rooted in geometric principles and Multi-View Stereo (MVS) (c. 2000-2010); the advent of regression-based deep learning approaches (c. 2010-2017); the rise of powerful generative models like GANs and Transformers (c. 2018-2022); and the current wave of diffusion model-based synthesis (c. 2021-present). Subsequently, this paper provide an in-depth analysis of Neural Radiance Fields (NeRF), a paradigm-shifting technique that has recently dominated the field. This paper discuss the specific application requirements driving NeRF research, delineate key challenges including model capacity allocation, noise handling, few-shot learning, and computational efficiency, and summarize corresponding solution strategies. Furthermore, I offer a structured taxonomy of contemporary NeRF-related works, categorizing them based on their focus on large-scale scenes, dynamic elements and complex lighting, few-shot reconstruction, and computational acceleration. This survey serves as a valuable resource for understanding the NVS landscape and navigating the rapidly expanding NeRF literature.

1. Introduction

Novel View Synthesis (NVS), the task of generating photorealistic images of a scene from arbitrary viewpoints given a set of input views, stands as a fundamental challenge in computer vision and computer graphics. Its applications are widespread, ranging from immersive virtual and augmented reality (VR/AR) experiences and robotics to visual effects and digital content creation. The core challenge lies in inferring the underlying 3D scene structure and appearance from limited, often sparsely sampled, visual data to render plausible novel perspectives.

Early approaches to NVS were predominantly geometry-based, relying heavily on classical Structure-from-Motion (SfM) and Multi-View Stereo (MVS) techniques to explicitly reconstruct 3D scene geometry (e.g., point clouds, meshes, or depth maps) [28, 9]. While foundational, these methods often struggle to produce highly photorealistic renderings, particularly in regions with complex non-Lambertian surfaces, intricate details, or significant occlusions. Furthermore, the quality of the synthesized views is critically dependent on the accuracy of the intermediate geometric reconstruction, which can be fragile.

The advent of deep learning ushered in a new era for NVS. Regression-based methods emerged, leveraging convolutional neural networks (CNNs) to directly predict novel view images [7]. Many subsequent works integrated learned components with 3D representations and differentiable rendering pipelines [15]. While initially often scene-specific, significant progress has been made towards few-shot NVS, enabling generalization across scenes from only one or a few input images, sometimes employing meta-learning or test-time optimization strategies [38].

Concurrently, generative models gained traction, particularly for challenging scenarios involving large viewpoint extrapolation where geometric priors might be weak or absent [33, 26]. These models excel at synthesizing plausible scene content by learning powerful priors from data, capable of filling in significant missing information [23, 25]. 3D Generative Adversarial Networks (GANs) also demonstrated capability in generating 3D objects, although often limited to object-centric setups and canonical poses [34]. However, maintaining long-range geometric and temporal consistency remained a challenge for purely generative approaches without strong 3D guidance.

A pivotal moment arrived with the introduction of Neural Radiance Fields (NeRF) [20]. By representing a scene as a continuous volumetric function optimized via neural net-

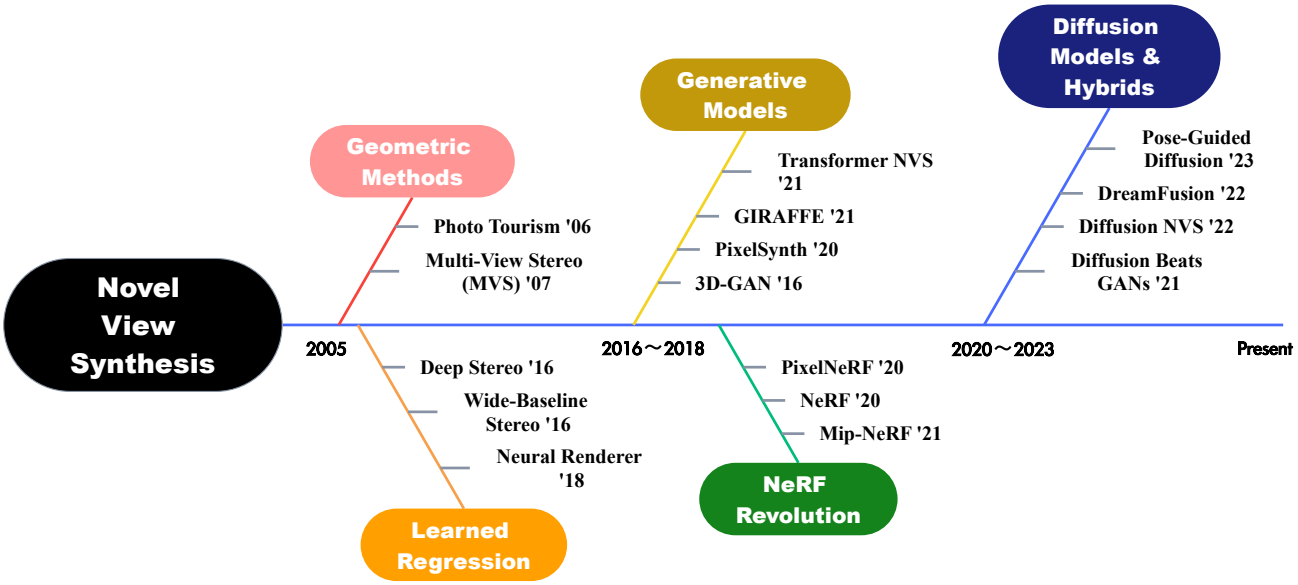

Figure 1. Timeline illustrating the evolution of Novel View Synthesis (NVS). Key phases include Geometric Methods, Learned Regression, Generative Models, the NeRF Revolution, and Diffusion Models & Hybrids.

works and rendered using differentiable volume rendering, NeRF achieved unprecedented photorealism and view consistency. This breakthrough spurred a massive wave of research, establishing NeRF as a cornerstone of modern NVS.

Most recently, diffusion models [6], initially demonstrating state-of-the-art performance in 2D image synthesis, have been successfully adapted for NVS and 3D-aware generation [24, 32]. These methods often combine the generative power of diffusion with geometric priors or integrate with NeRF-like representations, aiming to achieve both realism and multi-view consistency [16].

This paper provides a comprehensive survey of the evolution of Novel View Synthesis techniques. It traces the trajectory from early geometric foundations through the rise of deep learning-based regression and generative models, culminating in the current landscape dominated by Neural Radiance Fields and diffusion models. The primary focus of this paper is an in-depth analysis of NeRF-based approaches. This paper systematically discuss the application demands driving NeRF development, identify the core challenges encountered (e.g., foreground/background ambiguity, handling dynamic elements, optimizing from limited samples, computational cost), review proposed solution strategies (e.g., decomposition, appearance modeling, priors, architectural optimizations), and provide a structured taxonomy of the burgeoning NeRF literature. This includes works focusing on large-scale unbounded scenes [2, 31], handling dynamic scenes and complex illumination [19, 30], reconstructing from few input views [5, 12, 36], and accelerating training and inference [37, 8, 21]. By contextualizing NeRF within the broader history of NVS and organizing its variants, this survey aims to offer a valuable resource for researchers and practitioners navigating this rapidly evolving field.

2. Neural Radiance Fields (NeRF): Principles and Advancements

As introduced, the trajectory of Novel View Synthesis (NVS) took a sharp turn with the advent of Neural Radiance Fields (NeRF) [20]. This method provided a powerful paradigm shift away from explicit geometric representations towards continuous, implicit scene modeling optimized via analysis-by-synthesis. Its ability to generate highly photorealistic and view-consistent results from captured images established it as a cornerstone technology, sparking extensive follow-up research aimed at addressing its initial limitations and broadening its applicability. This section first reviews the core principles of NeRF and then delves into the major advancements categorized by the key challenges encountered in real-world scenarios, referencing the overview presented in Figure 2.

2.1. NeRF Core Principles

NeRF represents a static scene as a continuous 5D function, typically implemented as a Multi-Layer Perceptron (MLP), F_Θ. This function maps a 3D location $\mathbf{x} = (x, y, z)$ and a 2D viewing direction $\mathbf{d} = (\theta, \phi)$ to a volume density σ and view-dependent RGB color $\mathbf{c}$ [20]:

$$F_\Theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma) \quad (1)$$

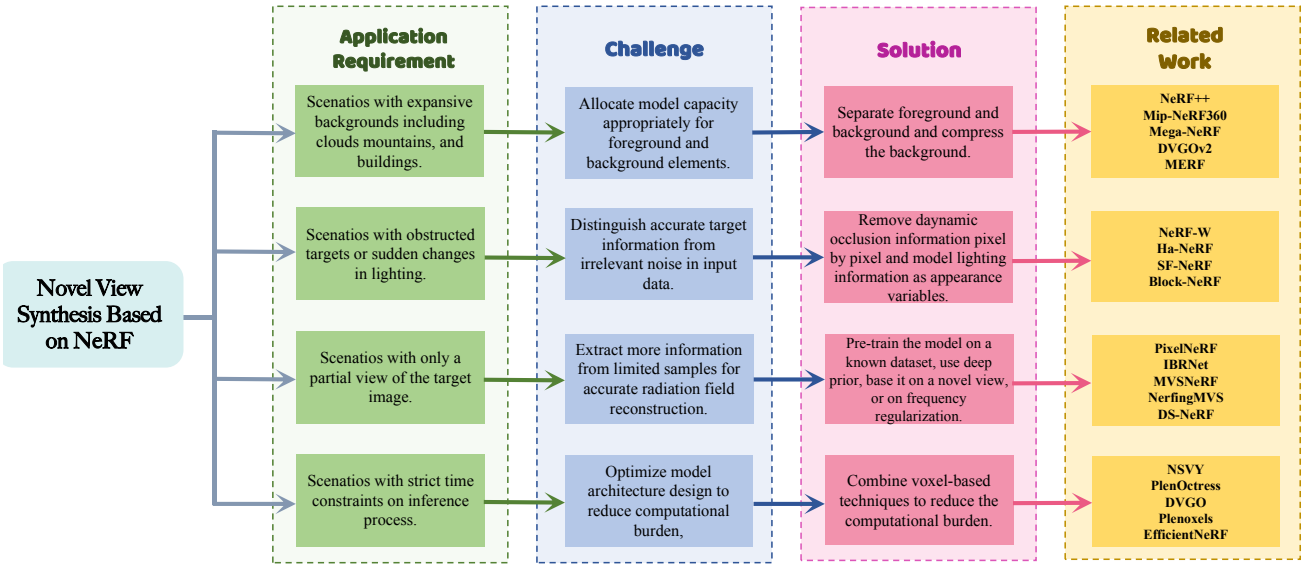

Figure 2. Overview of NeRF-based Novel View Synthesis, illustrating the relationship between application requirements, key challenges, common solution strategies, and related works across different research directions.

To handle the fine geometric and appearance details, the 5D input coordinates $(\mathbf{x}, \mathbf{d})$ are first mapped to a higher-dimensional feature space using a positional encoding $\gamma(\cdot)$, which helps the MLP learn high-frequency functions [20].

Rendering involves casting camera rays $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ for each pixel, sampling points along these rays, querying the MLP F_Θ (using the positionally encoded inputs) to get $(\mathbf{c}, \sigma)$ at each sample point, and then approximating the volume rendering integral [13, 20] to compute the final pixel color $\hat{C}(\mathbf{r})$. This is calculated via the discrete summation:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i, \tag{2}$$

where T_i represents the accumulated transmittance along the ray up to sample i, defined as:

$$T_i = \exp\left(- \sum_{j=1}^{i-1} \sigma_j \delta_j \right). \tag{3}$$

Here, σ_i and $\mathbf{c}_i$ are the density and color of the i-th sample, and δ_i is the distance between adjacent samples. A hierarchical sampling strategy with "coarse" and "fine" networks is used to concentrate samples in relevant regions of space [20]. The entire model F_Θ is optimized end-to-end by minimizing the photometric reconstruction loss between rendered pixel colors $\hat{C}(\mathbf{r})$ and ground truth pixel colors from input images.

An early, significant improvement was Mip-NeRF [1], which addressed aliasing artifacts caused by NeRF's ray-point sampling at varying scales. By instead reasoning about conical frustums along rays and using integrated positional encodings, Mip-NeRF achieved better anti-aliasing and detail, becoming a strong baseline upon which many subsequent works were built.

Despite these foundational successes, applying NeRF effectively often requires overcoming specific hurdles related to scene properties, data capture conditions, and computational constraints.

2.2. Challenge: Handling Unbounded Scenes

Problem: Capturing scenes "in the wild" often involves vast, distant backgrounds (sky, landscapes) that extend effectively to infinity. Vanilla NeRF and Mip-NeRF, primarily designed for bounded object-centric scenes or forward-facing captures using coordinate normalization (NDC), struggle to represent these unbounded 360° environments effectively, facing challenges in parameterization and efficient sampling [39, 2]. Allocating model capacity appropriately between the near-field foreground and far-field background is crucial.

Advancements: A key solution involves parameterization techniques that map infinite exterior regions into a bounded domain. NeRF++ [39] pioneered a two-component approach, using standard coordinates inside a unit sphere for the foreground and an inverted sphere parameterization for the background, with separate MLPs for each. Mip-NeRF 360 [2] proposed an elegant scene contraction function based on squared distance that maps distant coordinates towards the surface of a sphere of radius 2, allowing a single Mip-NeRF model to represent the entire unbounded scene. It also introduced a distortion-based

regularizer to encourage compact representations. This contraction technique became highly influential. Zip-NeRF [3], for example, integrated this contraction with efficient grid-based representations (inspired by Instant NGP [21]) and improved sampling, achieving state-of-the-art quality and speed for unbounded scenes.

2.3. Challenge: Modeling Dynamic Scenes and Appearance Variations

Problem: Real-world data often violates NeRF's core assumption of a static scene. Transient elements like pedestrians or vehicles, as well as appearance changes due to variable lighting conditions, shadows, or camera auto-exposure, can lead to significant blurring, ghosting, or inconsistencies in the reconstruction [19, 30]. Effectively distinguishing the underlying static scene from these dynamic factors is necessary.

Advancements: To address appearance variations and non-static elements, NeRF-W (NeRF in the Wild) [19] introduced the idea of augmenting the NeRF architecture. It learns a per-image latent embedding vector to capture appearance variations (lighting, exposure) and adds a separate "transient" MLP branch that predicts color and density for dynamic objects, allowing the static branch to focus on the consistent scene structure. Block-NeRF [30] scaled these ideas to city-scale scenes by decomposing the environment into individually trained NeRF blocks, each conditioned on appearance embeddings and time, while explicitly detecting and modeling transient objects (like vehicles) often using semantic segmentation priors.

2.4. Challenge: Reconstruction from Few or Sparse Views

Problem: Optimizing NeRF typically requires dense image capture (dozens to hundreds of views) covering the scene well. When only a few input views are available (few-shot NVS), the reconstruction problem becomes severely ill-posed, often leading to overfitting, geometric inaccuracies, and unrealistic novel views [38, 22]. Extracting robust 3D information from limited samples is the core difficulty.

Advancements: Solutions broadly fall into two categories: incorporating learned priors or applying regularization. (1) **Learned Priors:** Methods like PixelNeRF [38] leverage large-scale pre-training on multi-view datasets. They use a CNN to extract image features from input views, which then condition the NeRF MLP prediction based on the projected sample point location, enabling generalization to new scenes from one or few views. MVSNeRF [5] similarly uses pre-training but incorporates ideas from multi-view stereo by building a feature volume to guide the NeRF MLP. (2) **Regularization/Optimization Strategies:** Other methods focus on adding regularization terms during NeRF optimization for the specific sparse scene, without requiring extensive pre-training. DietNeRF [12] encouraged semantic consistency between rendered novel views and input views using features from pre-trained vision-language models like CLIP. RegNeRF [22] introduced regularizers that promote geometric smoothness by penalizing inconsistencies in rendered depth patches across different views. FreeNeRF [36] provided a simple yet effective frequency annealing strategy for the positional encoding, starting with low frequencies and gradually introducing higher ones, which significantly reduces overfitting in sparse settings.

2.5. Challenge: Accelerating Training and Inference

Problem: The need to query a potentially large MLP hundreds of times for every pixel ray makes both the training optimization and the rendering process computationally intensive and time-consuming (minutes to hours or days for training, seconds per frame for rendering) [21, 4]. This severely limits NeRF's use in interactive or real-time applications. Optimizing the architecture and query process is essential.

Advancements: A highly successful direction involves replacing the purely implicit MLP representation with hybrid approaches that leverage explicit data structures for faster querying. PlenOctrees [37] demonstrated real-time rendering by pre-baking a trained NeRF into an octree storing density and spherical harmonic coefficients, though the baking process itself was slow. Subsequent works focused on *fast optimization* of these explicit structures. DVGO [29] and Plenoxels [8] directly optimized density and appearance features (like spherical harmonics) stored on voxel grids (dense or sparse), largely removing the need for MLPs and achieving training times in minutes. TensoRF [4] proposed factorizing the 4D radiance field into compact tensor components (using techniques like CP or VM decomposition), significantly reducing memory footprint while enabling fast reconstruction. Perhaps most impactful was Instant NGP [21], which introduced multi-resolution hash grid encodings. This technique uses hash tables to store feature vectors corresponding to different resolution levels, queried via interpolation, allowing a very small MLP to reconstruct high-frequency details rapidly. This enabled high-quality NeRF training in seconds to minutes on a single GPU.

2.6. Summary and Outlook

The advancements discussed highlight the rapid maturation of NeRF. By addressing challenges related to scene scale, dynamics, data sparsity, and computational cost, NeRF variants have become increasingly practical and versatile. Research continues to push boundaries, exploring further improvements in fidelity, speed, editability, and generalization, solidifying NeRF's role as a central pillar in

Table 1. Performance comparison of NeRF acceleration methods on the NeRF-Synthetic dataset. Metrics include PSNR↑ (higher is better), SSIM↑ (higher is better), LPIPS↓ (lower is better), Rendering Speed(FPS)↑ (higher is better), and Training Time↓ (lower is better). Data adapted from He et al. [10].

Method	PSNR↑	SSIM↑	LPIPS↓	Speed (FPS)↑	Training Time↓
NeRF [20]	31.01	0.947	0.081	0.023	56 h
NSVF [18]	31.75	0.953	0.047	0.815	100 h
PlenOctrees [37]	31.71	0.958	0.053	167.68	58 h (incl. baking)
DVGO [29]	31.95	0.957	0.053	-	14.2 min
DVGOv2 [29]	32.76	0.962	0.046	-	6 min
Plenoxels [8]	31.71	0.958	0.049	-	11 min
ReLU Fields [14]	30.04	-	0.050	-	10 min
TensoRF [4]	33.14	0.963	0.047	-	17.6 min
PlenVDB [35]	31.90	-	-	20.75	12.4 min
EfficientNeRF [11]	31.68	0.954	0.028	238.46	6 h
Instant NGP [21]	32.11	0.961	0.053	-	5 min
NeRFAcc [17]	33.11	-	0.053	-	4.5 min

modern 3D computer vision and graphics.

3. Quantitative Comparison

To provide a clearer picture of the performance trade-offs among different NeRF advancements, this section presents quantitative results on standard benchmarks. Evaluating NeRF variants typically involves measuring image reconstruction quality using metrics like PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index), perceptual similarity using LPIPS (Learned Perceptual Image Patch Similarity), and computational efficiency in terms of rendering speed (Frames Per Second, FPS) and training time.

Table 1 specifically summarizes the performance of various methods aimed at accelerating NeRF training and inference, evaluated on the widely used NeRF-Synthetic (Blender) dataset. It highlights the significant progress made in reducing computational cost while maintaining or even improving rendering quality compared to the original NeRF.

Note: The data presented in Table 1 is adapted from the survey by He et al. [10].

4. Conclusions

This paper has presented a comprehensive survey of the field of Novel View Synthesis (NVS), charting its evolution from early geometry-based methods through the advancements brought by deep learning, generative models, and most recently, diffusion models. My primary focus has been an in-depth analysis of Neural Radiance Fields (NeRF), a technique that has revolutionized the field by enabling unprecedented levels of photorealism and geometric consistency through implicit volumetric scene representation and differentiable rendering.

I examined the core principles underpinning NeRF and systematically reviewed the significant progress made by the research community in addressing its initial limitations. Key advancements were discussed in handling challenging real-world scenarios, including:

- Extending NeRF to represent large-scale, unbounded environments.
- Adapting NeRF to model dynamic elements and appearance variations within scenes.
- Improving reconstruction quality from sparse or few input views through priors and regularization.
- Drastically accelerating NeRF's training and inference speed via hybrid representations and optimized data structures, paving the way for real-time applications.

The quantitative comparisons, such as those presented in Section 3, highlight the remarkable performance gains achieved by these NeRF variants across various benchmarks.

Despite this rapid progress, several challenges and exciting avenues for future research remain. Achieving true real-time performance on diverse hardware, including mobile devices, while maintaining high fidelity remains an ongoing pursuit. Handling highly complex dynamic scenes with intricate motion, topology changes, and extreme lighting variations requires more robust solutions. Further improving generalization from extremely sparse or casually captured data is crucial for democratizing high-quality NVS. Moreover, enhancing the editability and controllability of NeRF representations—allowing intuitive manipulation of scene content, appearance, and lighting—is a key direction for practical content creation workflows. The deeper integration of powerful generative priors from diffusion models or GANs, advancements in theoretical understanding, and ap-

plications bridging NVS with robotics and simulation also represent promising frontiers.

In conclusion, NeRF has fundamentally reshaped the landscape of novel view synthesis and implicit 3D scene representation. The continued exploration of its capabilities and the ongoing efforts to overcome its limitations promise a vibrant future for this field, with profound implications for computer vision, computer graphics, virtual/augmented reality, and beyond.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5855–5864, 2021. 3

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5479, 2022. 2, 4

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-NeRF: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 732–742, 2023. 4

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingdong Yu, and Hao Su. TensoRF: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, pages 333–350. Springer, 2022. 4, 5, 6

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Jingyi Zhang, and Andreas Geiger. MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14124–14133, 2021. 2, 4

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 8780–8794, 2021. 2

[7] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deep stereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5515–5524, 2016. 1

[8] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5501–5510, 2022. 2, 5, 6

[9] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. Multi-view stereo for community photo collections. In *2007 IEEE 11th International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007. 1

[10] Gaoxiang He, Bin Zhu, Bo Xie, and Yi Chen. Progress in novel view synthesis using neural radiance fields. *Laser & Optoelectronics Progress*, 61(12):1200005, 2024. 5, 6

[11] Tao Hu, Shu Liu, Yilun Chen, Tiancheng Chen, and Qi Shen. EfficientNeRF: Efficient neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12892–12901, 2022. 6

[12] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting NeRF on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5885–5894, 2021. 2, 4

[13] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. In *ACM SIGGRAPH Computer Graphics*, volume 18, pages 165–174. ACM, 1984. 3

[14] Animesh Karnewar, Tobias Ritschel, Oliver Wang, David Novotny, Andrea Vedaldi, and Aman Makadia. ReLU fields: The little non-linearity that could. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 6

[15] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3D mesh renderer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3907–3916, 2018. 1

[16] Jáchym Kulhánek, Erik Tretschk, Zexiang Wang, Christian Henkel, Yuning Jiang, Ali Gokaslan, and Vincent Sitzmann. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12516–12525, 2023. 2

[17] Ruilong Li, Hang Gao, Matthew Tancik, and Angjoo Kanazawa. NeRFAcc: Efficient sampling accelerates NeRFs, 2023. 6

[18] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. NSVF: Neural sparse voxel fields, 2020. 6

[19] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF-W: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7210–7219, 2021. 2, 4

[20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 2, 3, 6

[21] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022. 2, 4, 5, 6

[22] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. RegNeRF: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5480–5490, 2022. 4

[23] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature

fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11453–11464, 2021. 1

[24] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[25] Xuanchi Ren, Hao Ling, Ailing Zeng, Shizun Liu, and Hong Li. Look outside the room: Synthesizing a consistent long-term 3D scene video from a single image. In *European Conference on Computer Vision (ECCV)*, pages 473–490. Springer, 2022. 1

[26] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3D priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13033–13043, 2021. 1

[27] Self-reference. Related Work Summary Points, 2024. Placeholder for points summarized from broader related work.

[28] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3D. In *ACM SIGGRAPH 2006 papers*, pages 835–846. ACM, 2006. 1

[29] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5459–5469, 2022. 5, 6

[30] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8248–8258, 2022. 2, 4

[31] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-NeRF: Scalable construction of large-scale NeRFs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12859–12869, 2022. 2

[32] Daniel Watson, William Chan, Ricardo Polis, Alexander Tokmakov, Dima Metaxas, and Abhishek Kar. Novel view synthesis with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[33] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. PixelSynth: Generating a 3D-consistent experience from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4324–4333, 2020. 1

[34] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016. 2

[35] Hao Yan, Chenglei Liu, Chuan Ma, Jiaming Wang, Hesen Wang, Chengyuan Tu, Andrea Tagliasacchi, and Xuejin Liu. PlenVDB: Memory efficient VDB-based radiance fields for fast training and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 88–97, 2023. 6

[36] Jiawei Yang, Marco Pavone, and Yilun Wang. FreeNeRF: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8264–8273, 2023. 2, 4

[37] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5752–5761, 2021. 2, 4, 6

[38] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 1, 4

[39] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NeRF++: Analyzing and improving neural radiance fields, 2020. 4