Neural Radiance Fields for the Real World: A Survey

WENHUI XIAO, REMI CHIERCHIA, and RODRIGO SANTA CRUZ, Queensland University of Technology, Australia and CSIRO Data61, Australia
XUESONG LI, CSIRO Agriculture & Food, Australia and Australian National University, Australia
DAVID AHMEDT-ARISTIZABAL, CSIRO Data61, Australia
OLIVIER SALVADO, CLINTON FOOKES, and LEO LEBRAT, Queensland University of Technology, Australia

arXiv:2501.13104v2 [cs.CV] 9 Dec 2025

Neural Radiance Fields (NeRFs) have remodeled 3D scene representation since release. NeRFs can effectively reconstruct complex 3D scenes from 2D images, advancing different fields and applications such as scene understanding, 3D content generation, and robotics. Despite significant research progress, a thorough review of recent innovations, applications, and challenges is lacking. This survey compiles key theoretical advancements and alternative scene representations and investigates emerging challenges. It further explores applications on reconstruction, highlights NeRFs' impact on computer vision and robotics, and reviews essential datasets and toolkits. By identifying gaps in the literature, this survey discusses open challenges and offers directions for future research.

CCS Concepts: • **Computing methodologies** → **Computer vision representations**; **Computer graphics**; **Computer vision**.

Additional Key Words and Phrases: neural radiance fields, neural rendering, novel view synthesis, 3D reconstruction

ACM Reference Format:
Wenhui Xiao, Remi Chierchia, Rodrigo Santa Cruz, Xuesong Li, David Ahmedt-Aristizabal, Olivier Salvado, Clinton Fookes, and Leo Lebrat. 2025. Neural Radiance Fields for the Real World: A Survey. 1, 1 (December 2025), 46 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The perception and interpretation of three-dimensional (3D) space fundamentally shapes human interaction with the physical world. Enabling computer vision systems to understand and process 3D information in the real world is crucial for numerous applications including robotics, telehealth, immersive simulations, and many others. A key ingredient of a 3D-capable system is the selection of appropriate 3D primitives for representing objects and scenes. Unlike the *pixel* as a unified representation of 2D images, there are diverse options for 3D representations, e.g. voxels, meshes, and point clouds. An effective 3D representation must maintain precision and scalability to

Authors' addresses: Wenhui Xiao, wenhui.xiao@hdr.qut.edu.au; Remi Chierchia, remi.chierchia@hdr.qut.edu.au; Rodrigo Santa Cruz, r2.santacruz@qut.edu.au, Queensland University of Technology, Brisbane, Australia and CSIRO Data61, Brisbane, Australia; Xuesong Li, xuesong.li@csiro.au, CSIRO Agriculture & Food, Canberra, Australia and Australian National University, Canberra, Australia; David Ahmedt-Aristizabal, david.ahmedtaristizabal@data61.csiro.au, CSIRO Data61, Canberra, Australia; Olivier Salvado, olivier.salvado@qut.edu.au; Clinton Fookes, c.fookes@qut.edu.au; Leo Lebrat, leo.lebrat@qut.edu.au, Queensland University of Technology, Brisbane, Australia.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Association for Computing Machinery.
XXXX-XXXX/2025/12-ART $15.00
https://doi.org/XXXXXXX.XXXXXXX

efficiently represent the 3D world while being learnable to integrate with deep learning frameworks and advanced 3D vision systems.

Traditional 3D representations exhibit significant limitations in satisfying those computational requirements. While voxels and point clouds are compatible with deep-learning-based frameworks, they suffer from discretization artifacts—voxels face scalability constraints, while point clouds lack connectivity—and yield limited geometry accuracy. Meshes excel at representing complex geometry; however, their irregular topology and element shapes impede integration with deep-learning frameworks.

Recent research has introduced implicit neural representations (INRs) [154, 169] to overcome these limitations. INRs implement continuous implicit functions by neural networks, enabling accurate and memory-efficient scene representation. However, the scarcity of high-quality 3D ground-truth data poses a significant challenge for 3D supervision. To this end, 3D neural rendering has become a central paradigm connecting 2D and 3D. It is grounded in volumetric rendering, which models the physical imaging process where pixel colors are obtained through integration of radiance fields along each ray. The idea of modeling light transport as a radiance function dates back to the 1960s [19] and was formally introduced to computer graphics in the late 1980s [96]. In the 2000s, subsequent works began to represent 3D scenes directly using color voxel grids [107, 211]. In computer vision, differentiable volume rendering began to gain traction for high-quality novel view synthesis. Early methods incorporated depth cues or stereo-image priors using representations such as soft 3D volumes [174], Multi-Plane Images (MPI) [157], and voxel-grid decoders [142], but suffered from the scalability limitations. Meanwhile, Multilayer Perceptrons (MLPs) and the PointNet architecture [179] have shown a strong ability to learn 3D scene representations from geometric data [154, 169]. Subsequent works [163, 221] addressed the reliance of ground-truth geometry by using image-based rendering losses, but did not account for view-dependent effects.

Neural Radiance Fields (NeRF) [159], as known today, arises at the crossroads of these approaches, combining differentiable volume rendering with view-dependent radiance fields to enable photorealistic 3D scene synthesis. It introduces an elegant framework to learn a 3D scene as a continuous volumetric function, mapping a 3D position and a view direction to RGB color and density. Through volume rendering, NeRF enables the reconstruction of a 3D scene from only 2D images and associated camera poses. The NeRF formulation achieves high-quality novel view rendering of a real-world scene while preserving accurate 3D geometry, in contrast to MPI approaches that often introduce specular artifacts. This breakthrough is enabled by several design innovations, including ray batching, differentiable volume rendering, and positional encoding.

Since its introduction in 2020, NeRF has been drawing significant research attention in both computer graphics and 3D vision communities. Fundamental works have enhanced NeRF models to be more efficient and accurate, addressing challenges posed by complex real-world scenarios, while numerous applications have emerged to advance various 3D vision tasks. With thousands of up-to-date NeRF-related research, a comprehensive survey on this topic is both urgent and critical.

1.1 Our Contributions

Several early surveys have reviewed the broader landscape of neural rendering [235] or 2D/3D neural fields [279], which cover NeRF as one representative method but consider only a handful of early advancements; conversely, [291] focuses on progress between 2022 and 2024 and covers only part of the spectrum of NeRF's developments. However, the breadth and rapid progress of NeRF necessitate an up-to-date, dedicated, in-depth survey. Although [13] and [185] summarize a variety of NeRF's improvements, they lack a coherent taxonomy and pay little attention to real-world challenges and applications that are vital for translating NeRF research into practical deployment. [321] and [61] review NeRF's foundational advancements and discuss several applications. Nonetheless, they

Table 1. Comparison of our review to other related survey papers. ✓: main focus; Δ: mention.

			[61]	[235]	[279]	[321]	[185]	[291]	[13]	Ours
Fundamental progress	Framework		✓	✓	✓	✓	✓	✓	✓	✓
	NeRF-agnostic enhancements									✓
	Alternative scene representations		✓				Δ			✓
Real-world challenges	Degraded views		✓	✓				✓		✓
	Sparse training views		✓	✓		✓	✓	✓		✓
	Inaccurate camera poses		✓	✓		✓		✓		✓
	Complex light effects			✓	✓	✓		✓		✓
	NeRF-in-the-wild		Δ	Δ						✓
	Complex scenes	Dynamic		✓	✓					✓
		Indoor						✓		✓
		Unbounded	✓	Δ						✓
	Uncertainty quantification									✓
	Generalizability to unseen scenes		Δ	✓			Δ			✓
Applications - Reconstruction tasks	3D surface		✓	✓	✓	✓				✓
	Large-scale scenes		✓	✓		✓	Δ			✓
	Medical images					✓				✓
	Digital human		✓	✓	✓	✓	✓	✓		✓
Applications - Beyond reconstruction	Robotics		Δ		✓	Δ				✓
	3D generation and editing		✓	✓	✓	✓	✓	✓		✓
	Recognition		✓			✓		✓		✓
Datasets			✓				✓	✓	✓	✓
Tools										✓

do not analyze the factors affecting NeRF's performance in real-world settings, and omit critical challenges such as uncertainty quantification, dynamic scenes, and indoor environments. Compared with existing surveys, our work provides the first unified taxonomy that connects NeRF's technical developments to real-world deployment challenges and evaluation protocols, as outlined in Table 1. Given the explosion of NeRF-related research after 2023, we aim to provide a timely survey of NeRFs, enabling readers to pinpoint key research directions, match solutions to specific challenges, explore NeRF's diverse applications, and gain insights for future work.

To this end, this survey uniquely integrates both foundational and emerging research in a consistent taxonomy. We distill core progress in improving NeRF's rendering quality and efficiency, and clarify how alternative scene representations have evolved from the NeRF framework. Beyond methodological advances, we identify factors affecting NeRF's robustness and generalizability in real-world scenarios, offering insights into its strengths and limitations. We further contextualize NeRF within broader applications, from reconstruction tasks to computer vision and robotics pipelines, highlighting its growing role as a crucial representation across diverse 3D tasks. To support ongoing research, we curate an extensive collection of publicly available tools, datasets, and benchmarking evaluation protocols. Finally, we identify key research gaps and open challenges, providing perspectives to guide future exploration in NeRF-related studies.

1.2 Scope and Organization

The scope of this survey is summarized in Fig. 1. Our survey covers NeRF's major developments in training and rendering, robustness under real-world conditions, applicability across diverse applications, and resources (tools and datasets) benefiting both academia and industry. Building on its central innovation, NeRF has given rise to a variety of alternative scene representations, such as Gaussian Splatting [98] for real-time rendering. However, this survey does not cover the rapidly evolving line of work on Gaussian Splatting, as its methodology has diverged significantly from NeRFs and merits a dedicated survey of its own [55].

Our organization is as follows. Section 2 provides an overview of NeRF and an in-depth review of fundamental improvement strategies for rendering quality and efficiency, followed by a brief introduction to alternative scene representations in Section 2.6. Then, Section 3 offers a thorough

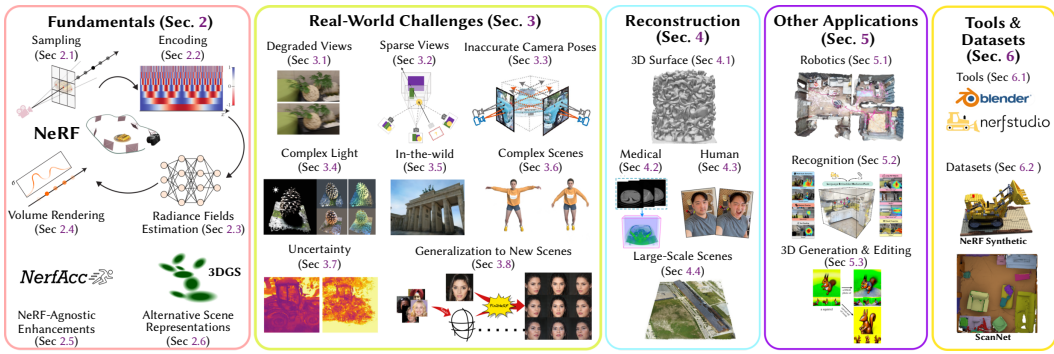


Fig. 1. Overview of our paper structure, outlining key sections covering NeRF fundamentals and improvement strategies, real-world challenges with corresponding solutions, diverse application domains, and practical resources. Images adapted from [5, 14, 23, 34, 42, 56, 66, 90, 98, 100, 150, 159, 171, 176, 178, 231, 242, 254, 313, 322].

analysis of key real-world challenges and their corresponding solutions. We explore various NeRF application tasks including reconstruction (Section 4), robotics (Section 5.1), recognition (Section 5.2), and 3D generation and editing (Section 5.3). Moreover, Section 6 compiles frequently used tools and datasets. Finally, we conclude by discussing open challenges and key research directions in Section 7.

2 FUNDAMENTALS OF NERF

NeRF [159] defines a 3D scene as a continuous volumetric function. Given a 3D position $\mathbf{p} \in \mathbb{R}^3$ in the world system and a view direction $\mathbf{d} \in \mathbb{R}^2$, it outputs the volume density $\sigma \in \mathbb{R}$ and the emitted radiance $\mathbf{c} \in \mathbb{R}^3$. Here, the volume density represents the probability of a ray terminating at the current position. $\mathcal{F}_\theta$ denotes such a learnable function, formally defined by

$$\mathcal{F}_\theta : (\mathbf{p}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma), \tag{1}$$

where σ depends only on $\mathbf{p}$, whereas $\mathbf{c}$ depends on both $\mathbf{d}$ and $\mathbf{p}$, to allow for modeling non-Lambertian effects. The learning process goes through a differentiable rendering procedure: (a) *ray casting and sampling*, (b) *samples encoding*, (c) *radiance field estimation*, (d) *volume rendering*, and (e) *optimization*.

Ray casting and sampling. Given the camera's intrinsic and extrinsic parameters, a ray is associated with each pixel. The union of all rays and pixels is used for the supervision of NeRFs, and for each training step, a subset of rays is *randomly* selected. To avoid unnecessary queries on empty or occluded regions while marching rays, points are sampled using *hierarchical volume sampling*. This process involves stratified sampling for a *coarse* model $\mathcal{F}_c$ predicting a rough density distribution, followed by inverse transform sampling from this distribution for a *fine* model $\mathcal{F}_f$.

Samples encoding. Along a ray, positional encoding is applied to each sampled point, which is associated with a 5D coordinate (three for position and two for view direction). Each element u of the 5D coordinate is normalized to $[-1, 1]$. Positional encoding of u is a L-levels frequency embedding, implemented by a γ_L function mapping u to a higher dimension space $[-1, 1]^{2L}$. Formally:

$$\gamma_L(u) = \left(\sin(2^0\pi u), \cos(2^0\pi u), \ldots, \sin(2^{L-1}\pi u), \cos(2^{L-1}\pi u)\right). \tag{2}$$

Radiance field estimation. The estimation of the radiance field is traditionally achieved through two MLPs for density and color. For conciseness, the rest of this section will use the germinal MLPs

as proposed in [159]. This backbone can be replaced by more efficient architectures or sparser formulations (see Section 2.3 and Section 2.6).

Volume rendering. The volume rendering [96] involves a continuous integration process through time-parameterized rays. NeRF performs numerical approximation of these integrals via a quadrature rule. Given a set of discretized time $\{t_i\}_{i=0}^{N}$ along a ray $\mathbf{r}$, the predicted color $\mathbf{c}_i$, and the density σ_i; the target pixel's integrated color between t_0 and t_N is approximated by $\hat{\mathbf{C}}(\mathbf{r})$ and given by:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=0}^{N-1} \mathbf{w}_i \mathbf{c}_i, \tag{3}$$

$$\text{where} \quad \mathbf{w}_i = T_i(1 - e^{-\sigma_i \delta_i}), \quad \delta_i = t_{i+1} - t_i, \quad \text{and} \quad T_i = e^{-\sum_{j=0}^{i-1} \sigma_j \delta_j},$$

where δ_i is the distance between contiguous samples t_i and t_{i+1}, and T_i is the accumulated transmittance measuring the likelihood of a ray traveling from t_0 to t_i without being intercepted.

Optimization. Using the pixel's color $\mathbf{C}(\mathbf{r})$ associated with each ray $\mathbf{r}$, the NeRF model is trained in a supervised fashion by minimizing the total squared error against the reconstructed color $\hat{\mathbf{C}}(\mathbf{r})$. For each batch of randomly sampled rays $\mathcal{R}$, the parameters of both the *coarse* and *fine* models are optimized simultaneously:

$$\mathcal{L} = \sum_{\mathbf{r} \in \mathcal{R}} \left(\left\| \hat{\mathbf{C}}_c(\mathbf{r}) - \mathbf{C}(\mathbf{r}) \right\|_2^2 + \left\| \hat{\mathbf{C}}_f(\mathbf{r}) - \mathbf{C}(\mathbf{r}) \right\|_2^2 \right), \tag{4}$$

where the subscripts c and f stand for the *coarse* and *fine* models.

As discussed in [159], 3D INRs mark a major advancement in 3D vision and novel view synthesis. However, NeRF, the most popular method in this field, has a lengthy runtime even on modern GPU workstations. Improvements on rendering quality and efficiency to the original NeRF pipeline include sampling (Section 2.1), encoding (Section 2.2), radiance fields estimation (Section 2.3), and volume rendering (Section 2.4). Section 2.5 summarizes backbone-agnostic improvements of rendering efficiency and quality, whereas Section 2.6 introduces alternative scene representations.

2.1 Sampling

The Monte Carlo sampling approximation described in Eq. (3) is numerically intensive without prior knowledge of the reconstructed geometry. From an unknown density distribution, sampling must be exhaustive to properly approximate the pixel's color $\hat{\mathbf{C}}$, and improvement of the approximation error is linear with the number of points sampled, making it computationally unaffordable.

Higher sampling efficiency can be achieved by wisely selecting integration segments in the non-empty space of the scene. Towards this end, *hierarchical volume sampling* scheme in NeRF employs a *coarse* model to estimate a rough density distribution, guiding the sampling of the *fine* model and improving numerical approximation efficiency. However, in Eq. (4) the *coarse* model's supervision requires rendering but is not used in the final rendered image.

Two lines of research aim to address limitations in this point sampling scheme: proposal-based and occupancy-based methods. Proposal-based methods e.g. Mip-NeRF 360 [6], replace the *coarse* model with a small *proposal* model, producing only density instead of both density and color. However, with this new formulation, the optimization of the *proposal* has to be revisited. Mip-NeRF 360 [6] comes up with an online distillation approach, where the *proposal* model is learned using the interlevel loss. This loss aligns the predicted weights histogram between the *proposal* model and the final NeRF model. The reduction in the size of the *proposal* model allows for a larger *fine* model (15×), contributing to higher synthesis performance with only a modest increase in computational time (2×). Different from proposal-based methods, occupancy-based methods manage to efficiently rule out sampled points with low density. For example, PlenOctrees [299] introduces the use of

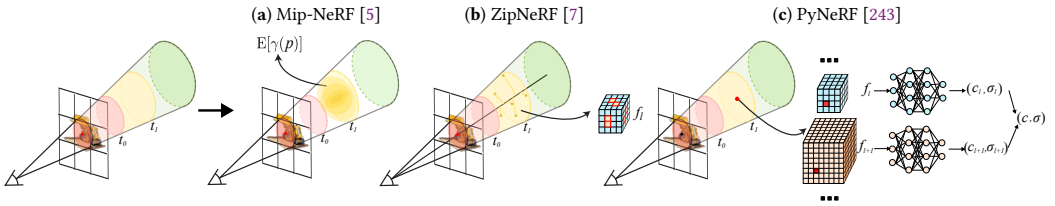


Fig. 2. Different multi-scale encoding designs.

an auxiliary NeRF-SH network trained with a sparsity prior, which is then converted to a sparse Plenoctree data structure to avoid computations on empty spaces of the scene. Hu et al. [81] analyze the weight and density distribution of NeRF's sampled points and introduce valid sampling to the coarse stage and pivotal sampling to the fine stage. Their proposed sampling strategies allow sampling points only in parts of the scene that contribute the most to the final color, allowing an 80% reduction of training time.

Recently, some researchers have emphasized the significance of pixel and view selection in NeRF training, and have improved NeRF training efficiency via *ray-* or *view-level sampling methods*. For instance, Zhang et al. [312] enhance training convergence by focusing on pixels with "*dramatic color change*" and regions with high rendering error. Xiao et al. [278] investigate the role of view sampling in NeRFs and benchmark different view selection methods to assess their impact. Their findings highlight that with better-selected samples, a NeRF obtains higher rendering quality with less computational time or supervision.

2.2 Encoding

The standard encoding of Eq. (2) treats each frequency signal equally, leading to issues when rendering at different resolutions. High-frequency positional encoding features can become aliased, resulting in artifacts known as "jaggies". Additionally, deep networks directly process the radiance field from the encoded input coordinates, which necessitates large and computationally expensive MLPs to represent the scene accurately. Later research on advanced encoding approaches has evolved towards anti-aliasing ability and high efficiency. We categorize existing encoding approaches into *integrated positional encoding*, *feature encoding*, *anti-aliased feature encoding*, and *compact feature encoding*.

Integrated positional encoding. The mipmapping technique [268] is a widely used anti-aliasing approach in the computer graphics rendering pipeline. To remove high-frequency aliased details in low-resolution signals, one creates mipmaps by downsampling and pre-filtering the texture map to different scales and picking an appropriate scale for rendering. Inspired by this idea, Mip-NeRF [5] extends NeRF to represent a scene with continuous scales by tracing a cone instead of a ray for pixel rendering, effectively modeling pre-filtered radiance fields. Mip-NeRF encodes a conical frustum for the MLP input via the *Integrated Positional Encoding (IPE)*. A conical frustum is approximated for efficient integral computation using a multivariate Gaussian with mean μ and covariance Σ, as presented in Fig. 2(a). An IPE feature $\gamma(\bullet)$ of a volume covered by the frustum can be computed through the expected high-frequency function of the mean and the diagonal of the covariance matrix, which can be mathematically expressed as:

$$\gamma(\mu, \Sigma) = \mathbb{E}_{\mathbf{p}\sim\mathcal{N}(\mu_\gamma,\Sigma_\gamma)}\left[\gamma(\mathbf{p})\right]. \tag{5}$$

The cone casting scheme and IPE allow Mip-NeRF to encode scale information into features and pre-filter high-frequency aliasing. Ultimately, Mip-NeRF can achieve anti-aliasing rendering with multi-scale input views.

Feature encoding. The idea behind *Feature Encoding (FE)* is to make input features learnable, thereby enabling the training and rendering to be accelerated without using a large MLP for inference. This approach uses auxiliary grid-based data structures, such as voxel octrees [226], hash grids [161], or planes [21], to store input features. These features are iteratively updated alongside the model's parameters through gradient backpropagation during training, adapting to the scene characteristics. Instant NGP [161] is a representative work in this area. It arranges trainable features into multi-level hash grids, corresponding to different resolutions of voxel grids. Instant NGP retrieves features from vertices of the voxel $\mathcal{V}_l$ where a 3D point $\mathbf{p}$ is located at each level l. These multiresolution hash encodings are aggregated using trilinear interpolation to form the level-wise feature f_l, which is formally expressed as:

$$f_l = trilerp(n_l \cdot \mathbf{p}; \mathcal{V}_l), \tag{6}$$

where n_l is the grid's linear size. A compact MLP for rendering takes encodings concatenated from each level as its input. The design of multi-resolution hash grids and the use of small MLPs successfully achieves fast training and high-quality rendering with a compact memory size.

Anti-aliased feature encoding. A recent trend in encoding is *Anti-Aliased Feature Encoding (AAFE)*, which combines the advantages of both IPE and FE for efficient, anti-aliased rendering. IPE struggles with efficiency because of its reliance on a heavy MLP for decoding scale information, while grid-based FE suffers from aliased artifacts. The challenge of achieving efficient anti-aliased rendering lies in making grid-based data structures scale-aware. To tackle this, ZipNeRF [7] adapts the complex multi-sample conical frustum used in Mip-NeRF to a multi-resolution hash-grid-based encoding. As shown in Fig. 2(b), ZipNeRF *multisamples* a conical frustum with a set of isotropic Gaussians to encode scale information, querying features through trilinear interpolation across grid levels. This scale-aware interpolation reduces artifacts by *downweighting* high-frequency signals. Although ZipNeRF is faster than Mip-NeRF [5], it is still slower than FE-based architectures like Instant NGP [161] in training time.

Other approaches, such as Mip-VoG [80] and Tri-MipRF [82], use grid-based structures across scales to predict color and density with scale-aware feature mipmaps. PyNeRF [243] trains a pyramid of NeRF models at different grid resolutions, accessing the closest resolutions corresponding to the sample. It then interpolates their scale-aware outputs to produce anti-aliased results compatible with any grid-based NeRF, as depicted in Fig. 2(c).

Compact feature encoding. Considering the extensive memory cost of grid-based feature encodings, a key trend in advancing feature encoding is to provide more compact and accurate representations. The successful application of quantization and frequency transform in codecs like JPEG2000 [148] has inspired some works [200, 217] that incorporate these techniques to increase feature sparsity. Binary Radiance Fields [217] applies binarization operations to the encoding parameters stored in the grid. Therefore, these parameters can be stored as either +1 or −1 instead of expensive float numbers in terms of 16-bit or 32-bit type. Similarly, Rho et al. [200] propose to use Discrete Wavelet Transform (DWT) and a binarization mask to convert high-dimensional features into masked wavelet coefficient grids. The feature used for color and density estimation is predicted through inverse DWT. Run-Length Encoding and Huffman Encoding can be applied to further compress wavelet coefficients and mask bitmaps, yielding a more compact feature representation.

2.3 Radiance Fields Estimation

The advancement of radiance fields estimation has developed towards higher rendering quality, faster training and rendering speed, and more compact models. In this subsection, we will discuss four mainstream techniques to improve efficiency and compactness.

2.3.1 Pre-caching. Pre-caching (or *baking*) accelerates rendering by pre-computing and tabulating geometry or appearance data into grid-based data structures like voxel grids. The inherent challenge is storing view-dependent color information with low memory footprints.

Some methods decompose view-dependent appearance into a combination of basis functions. For example, PlenOctrees [299] uses *Spherical Harmonics (SH)* functions and stores learned density and SH coefficients in the leaves of an octree-based grid. Similarly, FastNeRF [63] factorizes appearance as the inner product of the view-dependent weights and a position-dependent deep radiance map, which can be cached independently.

Other methods store pre-computed partial information of radiance fields and features, using a small MLP to predict view-dependent appearance. SNeRG [76] builds a 3D texture atlas composed of the predicted diffuse color, volume density, and specular features, which are stored in voxel grids. This 3D texture atlas can be further compressed with advanced codecs like JPEG. During rendering, SNeRG utilizes a lookup table to query accumulated color and feature vectors to then use a small MLP, which evaluates the view-dependent component. Similarly, MERF [196] stores data in a low-resolution 3D voxel grid and three high-resolution 2D planes, which are parameterized as multi-resolution hash encoding. To further improve rendering efficiency and model compactness, MERF adopts a binary 3D occupancy grid to sparsely store non-zero elements and applies quantization to each entry. DVGO [226] employs two voxel grids—a coarse one for geometry and a fine one for detailed reconstruction, combined with a small MLP for refinement.

2.3.2 Tensor rank decomposition/Factorization. Tensor rank decomposition is a technique to factorize complex feature vectors with multiple low-rank tensor components, enhancing training efficiency and reducing memory footprint. For example, the traditional CANDECOMP/PARAFAC (CP) decomposition [16] represents a tensor decomposition as a sum of multiple rank-one vectors outer products.

Chen et al. [21] apply CP decomposition to model a scene as a tensorial radiance field (TensoRF) with a set of rank-one tensor components, outperforming the vanilla NeRF model. This approach enables the use of compact MLPs or SH to compute the density and directional color from local features. To capture more complex scene elements, they further introduce the vector-matrix (VM) decomposition factorizing a tensor as the sum of outer products of vectors and matrices. The matrices describe the geometry and appearance characteristics of a scene for each plane of a 3D coordinate system. K-Planes [57] extends the tri-plane factorization to multi-plane factorization, allowing for the representation of dynamic scenes. Not limited to simple orthogonal transformation basis, Dictionary Fields [22] enhances representation capability by using a basis field with periodic transformations and a localized coefficient field, allowing for greater diversity in transformation bases.

2.3.3 Divide-and-conquer. To improve the computational efficiency of a NeRF model, it is a commonly used strategy to decompose a complex NeRF model into multiple smaller specialized units and then merge their rendering outputs. Leveraging this strategy, KiloNeRF [195] divides a large NeRF model into thousands of independent small MLPs, each of which learns a specific subpart of the scene. During the volume rendering, this grid of tiny MLPs is queried along the ray casts. The training of these small NeRFs is achieved through distillation. A large-capacity NeRF model is first trained as the *teacher* model. The distillation is then performed by matching the predicted color and density between the *teacher* model and the locally defined smaller student NeRF model. A similar approach, DeRF [193], proposes to conduct the Voronoi spatial decomposition of the scene. Then the Painter's algorithm is applied to efficiently draw the rendering of each Voronoi cell and compose the final rendering. It can effectively improve the rendering speed under the same quality. The divide-and-conquer mechanism is also borrowed when the complexity of a scene is too large to

be represented by one single model, which is common in large-scale scene reconstruction scenarios discussed in Supp. A.1, allowing to represent a large scene through an ensemble of small models.

2.3.4 Alternative Architectures. Some researchers have revisited the architecture selection and employed the advancements of recent large models or generative models to build a NeRF-like model and enhance rendering performance, e.g. GNT [245] and GANeRF [202]. Different from these works, Fridovich-Keil et al. [58] claim that scene learning can be accomplished without heavy MLPs. They propose Plenoxels, an MLP-less solution that efficiently learns the radiance field through an interpolation process of nearby voxels for opacity and color estimation. In this method, a sparse voxel grid stores the density and SH coefficients for each voxel. A trilinear interpolation on pre-cached information from nearby voxels is applied to model the continuous plenoptic function.

2.4 Volume Rendering

Classical volume rendering [96] assumes a 3D space filled with particles that can absorb, emit, and scatter light, integrating radiance along time-parameterized rays. NeRF simplifies this framework by only considering absorption and emission, using the volume rendering equation in Eq. (3) as a cornerstone ingredient in its self-supervised training. This allows for efficient optimization of a 3D INR from 2D images, but it still suffers in terms of scene representation accuracy and efficiency.

Using the quadrature rule, Eq. (3) assumes piecewise constant density and color within the interval $[t_i, t_{i+1})$. This assumption can lead to instability in sampling and ray supervision, causing fuzzy surfaces [244]. PL-NeRF [244] provides a more stable volume rendering approximation by replacing a piecewise constant integration rule with a piecewise linear one, where the weights $\mathbf{w}_i$ and transmittance T_i are given by

$$\mathbf{w}_i = T_i \left(1 - e^{-\frac{(\sigma_{i+1}+\sigma_i)\delta_i}{2}}\right), \quad \delta_i = t_{i+1} - t_i, \quad \text{and} \quad T_i = e^{-\sum_{j=0}^{i-1} \frac{(\sigma_{j+1}+\sigma_j)\delta_j}{2}}. \tag{7}$$

Increasing the order of the approximation scheme leads to increased sampled points within each interval, ultimately decreasing rendering efficiency. Instead of improving point sampling strategies as discussed in Section 2.1, Lindell et al. [133] alleviate this approximation issue by directly learning the primitive function of density and color to achieve fast and accurate ray integration. More specifically, an efficient AutoInt [133] framework is proposed to replace the numerical approximation of the integral. The derivative of the coordinate-based network models the original signal during training and reuses the parameters to build an integral network for automatic integration.

2.5 NeRF-Agnostic Enhancements

Despite fundamental progress, some works serve as NeRF-agnostic plugins to enhance either the efficiency or rendering quality. For instance, NerfAcc [118] provides flexible APIs fusing operations with CUDA kernels to incorporate efficient sampling methods for NeRFs. On the other hand, NeRFLix [316] enhances the rendering quality by correcting NeRF-style degradations like splatted Gaussian noise, ray jittering, and blurs in a post-processing manner.

2.6 Alternative Scene Representations

NeRF has inspired multiple alternative scene representations that extended beyond the original framework, aiming for higher cross-scene generalization and efficiency. This subsection reviews these approaches, highlighting their distinctions and benefits compared to the vanilla NeRF framework, with a focus on image-based rendering, explicit geometric representations, and differentiable rasterization.

In vanilla NeRF, the color of a pixel in a novel view is estimated by integrating the radiance field along its corresponding ray. Alternatively, inspired by image-based rendering, the color of a pixel can be obtained through warping, resampling, and blending related pixels from other views. Building on these ideas, methods such as IBRNet [256] and PixelNeRF [300] predict color and density from image features of source views, enabling novel scene rendering without retraining. Subsequent methods, like MVSNeRF [23] and NeX [225], further refine multi-view feature aggregation using plane-swept cost volumes and epipolar relationships. These approaches outperform traditional NeRF in generating new views of unseen scenes, and with scene-specific fine-tuning, they can achieve results comparable to State-of-the-Art (SOTA) single-scene NeRF.

Another issue with NeRF is the inefficiency in computing the volumetric integral, as evaluating an MLP at numerous sampled points within often sparse 3D space results in long training and rendering times. To alleviate this problem, techniques such as hierarchical sampling [159], "baking" [33, 195], and tensor factorization [15, 21] can be employed. Alternatively, some approaches offer a different solution by using explicit geometric structures to support radiance inference. For instance, Point-NeRF [285] employs a point-cloud representation along with multi-view features to guide the sampling and computation of radiance and density during ray marching. Similar approaches, like NeuMesh [287] and NSVF [138], leverage triangular meshes and sparse voxel grids, respectively. These methods provide a less costly inference compared to NeRF [159] and explicit 3D representations of the geometry, which are key for applications beyond novel view synthesis.

Besides ray marching, neural rendering can also leverage differentiable rasterization methods [109, 141], which project 3D primitives directly onto the 2D image plane for rendering. A notable method in this domain is 3D Gaussian splatting [98], where scenes are represented as collections of 3D Gaussians, and novel views are rendered by projecting each Gaussian onto the image plane and compositing them via alpha-blending. Other visual primitives, such as marching tetrahedons [75], textured polygons [33], and view lifted point cloud [104], can also be used. These methods often take advantage of optimized and hardware-accelerated rasterization procedures, enabling real-time throughput.

3 ADVANCEMENT FACING REAL-WORLD CHALLENGES

NeRF-based approaches excel in view synthesis under ideal conditions, such as densely distributed and high-quality training views, accurate camera poses, simple lighting, static objects, and controlled backgrounds. However, real-world settings often challenge these assumptions, leading to artifacts in synthesized views. Key challenges include variations in views, camera poses, scene complexity, lighting, uncertainty, and generalizability. This section illustrates the progress made in addressing these real-world challenges.

3.1 Degraded Views

Traditional NeRF models are optimized by minimizing the total squared error between predicted and observed pixel values, relying on an accurate correspondence between a ray and its associated pixel value provided by a high-quality training view. However, in real-world settings, training views often suffer from different corruptions such as 1) noise and Low Dynamic Range (LDR) in low-light scenes, 2) motion blur and defocus, 3) low resolution, and 4) haze. These corruptions can result in inconsistent pixel values, leading to poor reconstructions with artifacts when used for training. This subsection explores the challenges posed to the rendering performance of NeRFs from the aforementioned degradation types and discusses related advancements.

Low-light scenes: noise and low dynamic range. Images (LDR on default) often fail to represent the High Dynamic Range (HDR) information of a low-light scene, where there is a high contrast

between the brightest and darkest region, lacking high-frequency details. RawNeRF [156] addresses these challenges by optimizing NeRF directly on RAW data in linear HDR color space, applying variable exposure adjustment and tonemapping for better HDR signal learning. Though RawNeRF can produce HDR rendering with minimal noise, its requirement of RAW image data has the limitation of large memory footprints. Other HDR enhancement research manages to incorporate the physical imaging process upon the HDR NeRF framework to simulate the LDR outputs and match the LDR input views. The simulated imaging process transforms the HDR output to the LDR one through a tone mapper, which considers exposure information in the image metadata [85] or white balance [94].

Blurriness: motion blur and defocus. Image blurriness from camera motion or defocus cannot be modeled with a static ray, requiring the pixel value to be represented as a weighted sum of sharp rays originating from different spatial positions. Deblur-NeRF [147] addresses this by modeling the blurred pixel values as the convolution of the sharp image and a sparse blur kernel. This blur kernel is derived from a canonical sparse kernel and learned view embeddings. Another approach consists of simulating the physical image formation process throughout the exposure time. Specifically, the rendered blurry image is the weighted sum of a sequence of predicted sharp images captured along the camera's trajectory during this period. The camera trajectory can be approximated using the Lie-Algebra of $\mathbf{SE}(3)$ [147], a cubic B-Spline curve [147], a Bezier curve [112], or parameterized by a screw axis [111]. DP-NeRF [111] further introduces the adaptive combination of sharp rays regarding the relationship between depth and defocus blur. Additionally, event streams preserve HDR details in low-illumination and fast-motion environments and recently have been used for enhancing deblurring and low-light scenes [180, 206].

Low resolution. Learning a high-resolution (HR) 3D representation of the scene from a set of low-resolution (LR) images constitutes a super-resolution (SR) task. This can be achieved in either *pre-processing*, *in-processing*, or *post-processing* ways. HRNVS [298] works as a *pre-processing* method that super-samples the available training views, providing HR samples for training a HR NeRF model. The *in-processing* NeRF SR method [249] simulates the super-sampling scheme by casting multiple rays to each sub-pixel within each pixel. The learning of each sub-pixel ray is supervised by a sum of weighted errors between each predicted sub-pixel and the corresponding ground-truth integer pixel. A refinement step is further applied to enhance high-frequency details from an HR reference frame. In *post-processing* approaches [84, 130], a SR module (either single-image-based or reference-based) is used to super-sample the rendered LR views. Patch-based refinement enables the SR module to exploit the information embedded in related regions across nearby rays or HR reference views.

Haze. Haze in training views introduces the ambiguity between the reconstructed target and light-scattering atmospheric particles, complicating the ray density estimation. To solve this ill-posed problem, dehazing methods [28, 190] incorporate a physics-based scattering model into the volume rendering equation. These methods restore a clear scene from a hazy environment by disentangling the actual appearance and geometry of the target from the ambient particles.

3.2 Sparse Training Views

NeRF is trained from a collection of posed images. Like multi-view stereo, training of NeRFs intensely relies on fusing information from different viewpoints to successfully extract meaningful 3D information. The lack of diversity in view directions provides limited information on the correct geometry, turning the inverse problem of training a NeRF ill-posed. This subsection reviews the latest advancements in learning a 3D scene from sparse views or even a single view.

A popular line of works focuses on developing additional geometric constraints for regularizing under-constrained geometry learning. For instance, [201, 252] leverages depth priors to constrain NeRF's training. Considering external geometric priors are not always available, some research imposes priors based on information theory [101], frequency bands [288], and density estimation [212]. Hand-crafted geometric constraints are usually based on specific scenes or architectural hypotheses that cannot be generalized. Some methods generate training samples in unobserved viewpoints and take advantage of learned priors from pre-trained deep models. For example, DietNeRF [87] generates pseudo labels for both observed and unobserved views, constraining the learned geometry with consistent semantics in a semi-supervised manner. Rather than solely utilizing hand-crafted or learned priors, DiffusioNeRF [275] learns a joint probability distribution of geometry. The regularization-based methods help reduce visual artifacts in novel view generation. However, they suffer from poor generalization properties, when the queried view is far from the training set.

Concurrent works treat a NeRF as a decoder, conditioned on either codebooks or image features. Codebooks-conditioned methods [89, 194] retain latent codes for different scene components e.g. shape and appearance or foreground and background, which condition a NeRF with a shared latent space in the same category. Despite category-level view generalizability, these methods are limited to synthetic datasets and struggle with complex backgrounds. Image feature-based methods [23, 78, 300] condition NeRFs on available 2D image features by projecting 3D points into 2D planes and aggregating pixel-aligned image features, such as PixelNeRF [300] and MVSNeRF [23] presented in Section 2.6. ReconFusion [271] further marries diffusion priors with the PixelNeRF-based models for out-painting unobserved geometry and texture. These image feature-based methods produce good novel view synthesis results from a small number of views or even a single view. However, they are prone to yielding blurry rendering or artifacts in complex real-world settings.

A recent trend is to revisit the novel view synthesis as a generative task, aiming to extrapolate the appearance and shape of a scene with limited observations. To equip NeRFs with generative ability, this line of work incorporates the NeRF scene representation into generative models, such as generative adversarial network (GAN) and diffusion models. Using the GAN framework, early approaches build a NeRF model conditioned on the scene components [162] or image features [18]. Though impressive, these methods may yield inconsistent rendering and lack fine-grained details. With the great success of diffusion models in 2D image generation, other works manage to leverage diffusion priors to predict unseen details in single-view/sparse-view NeRF reconstruction. They reach this by generating feature encodings with diffusion models [24, 216], distilling a 3D scene from the view-conditioned diffusion models [140, 319], or employing a NeRF as the diffusion decoder [286]. Despite their compelling results, these methods are still facing the common limitations associated with diffusion models including expensive optimization time, limited rendering resolution, and the inability to handle unbounded scenes.

3.3 Inaccurate Camera Poses

The training of NeRFs requires accurate camera poses to precisely learn the 3D scene. Recent works have explored how to estimate or correct pose information during training. iNeRF [296] is the first to perform pose estimation from a trained NeRF by optimizing pose parameters to minimize the photogrammetric loss in Eq. (4). This approach can be used to predict the pose of additional images to extend NeRF's training set. NeRF−− [263] only requires a set of unposed images and allows for computing camera parameters while training a NeRF. The proposed optimization is robust for translation and rotation noises up to 20% and 20° respectively, limiting its use to forward-facing scenes. SiNeRF [276] improves upon NeRF−− [263] by replacing NeRF's activation function with

SIREN [220] and proposing a mixed region sampling based on SIFT detectors for ray selection instead of random sampling.

Other methods introduce geometric constraints to learn camera parameters more effectively. BARF [129] extends the 2D image alignment problem into 3D by minimizing the photometric loss between pairs of corresponding images through planar alignment. To make their method robust to poor initialization, a coarse-to-fine approach is used by gradually adding higher frequencies in the positional encoding throughout the optimization. GARF [38] addresses the spectral bias of sine and cosine functions used for positional encoding in Eq. (2) (biased towards low-frequency functions). Aimed at this, it uses Gaussian activation functions for a better representation of the first-order derivatives of the encoded signal. L2G-NeRF [30] closely follows BARF [129] through a new local-to-global approach. By combining a flexible pixel-wise alignment and a frame-wise constrained parametric alignment, this method provides improved camera pose computation.

Geometric constraints also provide support for the estimation of camera intrinsics addressing challenges posed by camera distortions and large displacements. SCNeRF [90] is capable of recovering camera distortions' parameters. It introduces a geometric loss called projected ray distance. This reprojection error is computed for 3D correspondences between two views. NoPe-NeRF [10] focuses on large camera displacements. It demonstrates that correcting monocular depth priors with scale and shift parameters leads to better renderings. It utilizes a monocular depth network to obtain a depth map, from which they extract a point cloud used for a point cloud loss and a surface-based photometric loss. It reports improvements over both BARF and SCNeRF. Finally, CamP [170] employs a camera-preconditioning method to facilitate the optimization process, leading to improved convergence compared to SCNeRF.

Previous methods assume that dense input views are available; with fewer views retrieving accurate camera extrinsics becomes more challenging. In this context, SPARF [241] introduces a depth consistency loss, which generates depth maps from training views, to condition the geometry of novel renderings in a pseudo-depth supervision fashion. Additionally, it incorporates a visibility mask estimated from rays' transmittance to account for occlusions. SC-NeuS [83] focuses on reconstructing accurate geometry by implementing an Signed Distance Function (SDF) approach, allowing for efficient mapping of 2D features to 3D space. Subsequently, a view-consistent projection loss is used to compute the reprojection error and correct camera poses. Furthermore, it defines a view-consistent patch-warping loss, comparing patches between two views using the homography matrix with a normalized cross-correlation score.

3.4 Complex Light Effects

NeRFs model expected color as a function of position and view direction but often struggle to capture the appearance of objects under complex real-world lighting. Visual effects such as reflections, occlusions, and shadows also depend on surface characteristics like material properties, smoothness, and environmental lighting. To model these effects, some works have adopted a physics-based approach using the rendering equation [95]. The rendering equation describes the total radiance $L_o(\mathbf{p}, w_o)$ observed at a point $\mathbf{p}$ from a direction w_o as the sum of surface-emitted radiance $L_e(\mathbf{p}, w_o)$ and reflected radiance $L_r(\mathbf{p}, w_o)$. Mathematically,

$$L_o(\mathbf{p}, w_o) = L_e(\mathbf{p}, w_o) + L_r(\mathbf{p}, w_o) = L_e(\mathbf{p}, w_o) + \int_\Omega B(\mathbf{p}, w_i, w_o) L_i(\mathbf{p}, w_i)(n \cdot w_i)\, dw_i, \tag{8}$$

where the reflected radiance is the integral over all incoming directions Ω, with incoming radiance $L_i(\mathbf{p}, w_i)$ modulated by the Bidirectional Reflectance Distribution Function (BRDF) B and the cosine of the angle between the incident direction w_i and the surface normal n. Further modeling can also

account for the light scattering that occurs at the subsurface level. However, this is often neglected in most computer vision applications.

IDR [294] and PhySG [309] are the first approaches to explore the disentanglement of volumetric representation and appearance in the NeRF context. While IDR focuses on an implicit approach to modeling the scene BRDF, PhySG explicitly estimates light components such as albedo, specular BRDF, and environment map. Both methods jointly learn a surface representation and exploit well-conditioned normals for appearance estimation. However, both assume several approximations and neglect phenomena such as interreflections and occlusions. NeRV [223] estimates a shadow map, surface roughness, and interreflections by approximating environment lighting visibility with a dedicated MLP. However, it assumes known illumination, which limits its applicability. NeRD [11], on the other hand, models illumination using Spherical Gaussians, and implicitly encodes shadows together with the diffuse albedo. Ref-NeRF [246] introduces a reflection direction parametrization to simplify the appearance modeling. It estimates the color as a linear combination of diffuse and specular components. The diffuse component is estimated by the NeRF model, while the specular component results from a directional MLP that encodes surface roughness, surface normal, and view direction. Ref-NeuS [65] builds on Ref-NeRF by estimating an SDF. Additionally, it uses the SDF to access a visibility parameter to refine the reflection score, which mitigates the contribution of a reflective region to the total color loss. S3-NeRF [290] exploits shades and shadows to disentangle geometry and appearance.

An important application in this domain is scene relighting, where new renderings can be generated from different lighting conditions for interactive graphics or animation purposes. For instance, NeRF-OSR [205] focuses on outdoor scene relighting, closely following NeRD [11]; NeRFactor [313] models free-viewpoint relighting assuming a single unknown illumination; and PS-NeRF [289] instead models multiple unknown directional light images.

Other applications explore real-world phenomena. NeRF-W [151] tackles novel view rendering for outdoor, unstructured collections of images with different appearances and illumination. NeRFReN [72] demonstrates that NeRF density has two peaks—one corresponding to the medium depth and the other to the scene reflected or behind glass—and uses this observation to model mirror-like reflections. Similarly, works such as [168, 237, 320] focus on rendering scenes through glass, eliminating the undesired reflection and refraction effects. Hazy or underwater scenes pose challenges from scattering effects; related methods are detailed in Sec 3.1 and Sec 4.4.2. Moreover, ORCa [236] models the environment's appearance from specular objects on unknown surfaces.

3.5 NeRF-in-the-Wild

NeRFs were originally formulated under the assumption that scenes remain geometrically, materially, and photometrically static, such that two observations from an identical camera pose yield identical measurements. Real-world environments rarely satisfy these conditions: illumination changes over time, imaging devices vary in resolution and spectral response, objects move, occlusions appear and disappear, and shadows shift. These factors undermine the baseline assumptions of the NeRF formulation, resulting in rendering artifacts that limit its applicability beyond controlled laboratory conditions.

The research area extending NeRFs to unconstrained settings is often referred to as "NeRF-in-the-Wild". Within this domain, several approaches tackle different aspects of real-world variability. Photo-tourism [3] exemplifies the challenge of reconstructing landmarks from unstructured internet photo collections, where even images captured from the same viewpoint on the same day can differ significantly in illumination, shadows, and moving objects. To address these issues, NeRF-W [151] introduces per-image appearance embeddings to model illumination and camera-dependent variations, alongside transient embeddings to capture and suppress dynamic elements such as

moving objects and temporary occlusions. This decomposition improves rendering consistency and mitigates artifacts such as ghosting and color shifts.

These dynamic elements, commonly referred to as *distractors*, have been further explored in subsequent works. RobustNeRF [207] treats them as outliers within the optimization process, using robust loss functions to suppress their impact. NeRF *On-the-Go* [199] predicts per-pixel uncertainty and assumes distractors correspond to high-uncertainty regions, which are progressively filtered out during training. IE-NeRF [259] takes a complementary approach by coupling NeRF with an inpainting module that generates transient masks to explicitly remove occlusions, thereby refining static scene reconstruction. Together, these strategies improve novel-view synthesis by emphasizing stable scene content.

Beyond robustness to distractors, some *"in-the-wild"* works push NeRF models toward new capabilities. Ha-NeRF [29] hallucinates appearance from single images to synthesize novel views under unseen lighting conditions, eliminating NeRF-W's per-image optimization requirement. Neural 3D Reconstruction in the Wild [227] extends NeRF to neural surface reconstruction, enabling geometry recovery resilient to distractors in large-scale photo collections. Complementarily, NeRS [307] introduces a surface-based neural representation that yields closed, reflectance-aware surfaces from sparse, real-world multi-view data, expanding NeRF's utility into surface reconstruction with materials and lighting decomposition. Together, these methods underscore NeRF-in-the-wild's evolution beyond rendering stability to broader applications in appearance modeling and 3D reconstruction.

Scaling from landmarks to entire urban environments, as in photo-tourism, requires handling massive, heterogeneous image collections captured under inconsistent conditions and frequent occlusions. For example, Urban Radiance Fields [198] integrate multimodal data for city-scale reconstruction; related large-scale methods are detailed in Sec 4.4.3.

3.6 Complex Scene Configurations

Original NeRF formulation assumes the reconstructed scene to be static, object-centric or front-facing, and spatially bounded. However, real scenes may violate some of these requirements. This subsection explores three typical relaxations of these assumptions: dynamic scenes, indoor scenes, and unbounded scenes.

3.6.1 Dynamic scenes. Modeling of dynamic scenes with NeRF is a complex task that involves designing a neural representation over a four-dimensional (4D) space of 3D space and time. A naive idea consists of extending the input with a new dimension for the time component. However, this often fails to precisely model dynamic scenes and can be very costly to train. Several approaches have been developed to efficiently parameterize neural rendering over time.

The first group of methods learns a canonical representation of a scene along with a deformation field for each time step. Using a time-indexed deformation, novel coordinates for a scene can be computed and substituted directly into the volumetric rendering equation to learn a canonical NeRF representation [178, 239]. Nerfies [171] parameterizes the deformation field on $\mathbf{SE}(3)$. To enforce a robust optimization, penalties for non-rigid transformation and displacement of points in the static background are introduced. The deformations are estimated using a coarse-to-fine regularization scheme. Park et al. [172] propose to express NeRF's canonical space in a higher dimension, to allow for non-regular deformations such as dynamic topological change. Wang et al. [248] introduce optical flow constraints for better temporal regularization of video with rapid motion.

Another approach is to learn a coupled representation of appearance and scene flows, capturing the local deformation across neighboring frames. In NSFF [125], the training imposes temporal photometric consistency and cycle consistency between the forward and backward scene flow. NSFF further considers geometric consistency and single-view depth for added regularization.

To reduce the need for extensive supervision and enable continuous-time training from a single camera, Du et al. [51] re-parameterize the flow field with neural ordinary differential equations [25]. Recently, this formulation has shown SOTA results for complex camera and scene motion using a volumetric image-based rendering framework [126].

Compositional approaches separately model fixed and moving scene contents. For instance, Ost et al. [167] learn a scene graph representation to encode independent object transformations and radiance for the static scene. However, these frameworks are challenged by complex object deformations and lighting variations. Recent methods further use self-supervised schemes to decompose a dynamic scene into static and dynamic data [273], or into static, deforming, and newly appearing objects [222].

Instead of explicitly learning the deformation information, one can parameterize a 4D dynamic scene representation with compact time-conditioned feature vectors. N3DV [120] introduces a NeRF model conditioned on a set of temporal latent codes, allowing for time interpolation. However, this approach suffers from lengthy training time and is challenged by long video sequences.

Recent methods build an efficient and sparse 4D scene representation by decomposing a dynamic scene into different components, storing them in plane-based [57, 222] or voxel-based structures [222, 251]. For example, K-Planes [57] decomposes a dynamic scene into six sparse feature planes: three for space representation, and three for spatial-temporal variations.

3.6.2 Indoor scenes. In contrast to traditional NeRF pipelines, where objects of interest are reconstructed using concentric views, indoor scenes require outward-facing capture, a process often referred to as "inside-out" in many NeRF reconstruction codebases. A typical characteristic of these views is limited overlap between perspectives, with a high prevalence of self-occlusion. Additionally, they often comprise numerous low-texture regions such as walls and floors, which are well-known obstacles in traditional multiple-view geometry reconstruction. Without adequate additional supervisory inputs, the novel view synthesis becomes a challenging extrapolation task requiring in-painting of the missing regions. To tackle these challenges, learning-based priors can be leveraged to provide additional supervision for NeRFs' optimization. Some approaches [201, 264] exploit depth priors to guide the sampling around the actual geometric surface. Roessle et al. [201] also use depth for additional supervision to handle the shape-radiance ambiguity during optimization. Semantic information from pre-trained models like CLIP [186] can also provide additional guidance to further constrain NeRFs' optimization. It can be utilized as a sampling guide [183], or a regularizer enforcing cross-view semantics consistency [106].

Notably, the aforementioned methods are designed for perspective views. Another common type of indoor image is the 360° panorama. The scene representation of multi-sphere images (MSI) [4] has been widely used in 360° view synthesis. Recent approaches [32, 106] have adopted a similar design to deal with 360° images. These methods project each 3D position from Cartesian space to spherical space, ensuring the volume rendering is conducted along the actual ray trajectory.

3.6.3 Unbounded scene. In unbounded scenes, like outdoor 360° scenes and street views, objects can be captured from an arbitrary direction and distance. These scenes are often boundless, posing significant challenges in terms of sampling complexity and model capability when attempting to represent an unbounded learnable space.

The original NeRF addresses these challenges in a front-facing setting by applying the perspective projection that contracts distant samples into normalized device coordinates (NDC) space along the z-axis, as shown in Fig. 3 (left). Nevertheless, this contraction method is not suitable for unbounded 360° scenes where rays aim in all directions. To parameterize challenging unbounded 360 scenes, an early technical report [310] decomposes the scene into two volumes: an inner sphere volume representing the foreground and an MSI-based outer volume representing the

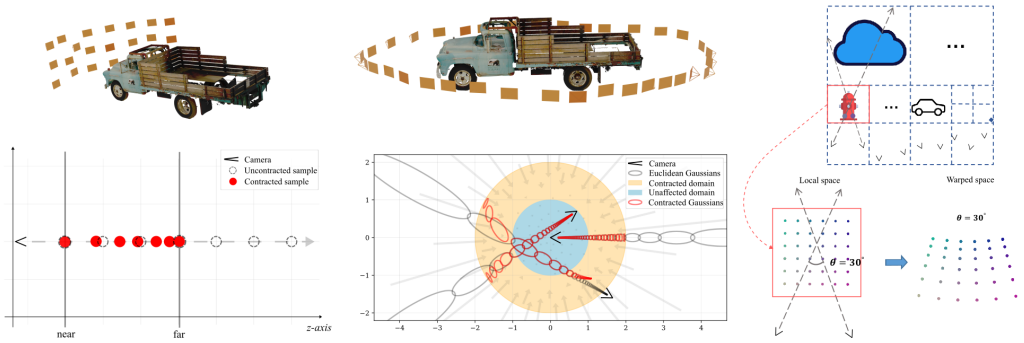


Fig. 3. Different contraction functions are applied for different scene settings: (**left**) a front-facing scene [159], (**middle**) a 360° scene [6], and (**right**) a scene of free camera trajectories [255]. The **top** row presents the camera trajectories in different settings and the **bottom** row demonstrates corresponding contraction methods. Figures are reproduced from Mip-NeRF 360 [6] and F^2-NeRF [255].

background. This design is also adopted in Plenoxels [58]. Another line of methods [6, 196, 284] aim to parameterize the scene with a novel space contractor function. For instance, Mip-NeRF 360 [6] contracts sampled Gaussians in distant space to a radius-2 ball and correspondingly transforms ray intervals proportional to their disparity, as demonstrated in Fig. 3 (middle). In practice, cameras can follow an arbitrary trajectory. To achieve fast scene rendering with free trajectories, F^2-NeRF [255] subdivides the scene into regions and introduces a perspective warping to each region, as presented in Fig. 3 (right). This warping ensures that the distances between samples in the warped space align with those observed in visible cameras.

Another challenge in unbounded scene reconstruction is the scale of a scene. Directly training a NeRF model on the entire large-scale scene data is impractical due to excessive computational and memory overheads. Section 4.4 further explores the challenges and advanced research in applying NeRFs to large-scale scene reconstruction.

3.7 Uncertainty Quantification

Quantifying uncertainty in a NeRF model is crucial for NeRFs' practical applications under complicated and ambiguous real-world scenes, such as those with limited training data. The uncertainty in deep learning is typically categorized into two types [97] — *aleatoric uncertainty* and *epistemic uncertainty*. For NeRFs, *aleatoric uncertainty* refers to the inherent noise (e.g. camera noise) or randomness (e.g. transient objects or illumination changes) in the training data. On the other hand, *epistemic uncertainty* considers the uncertainty within the entire system due to model misspecification or insufficient training data, such as limited available training views.

NeRF-W [151] paves the way for uncertainty quantification of NeRF models by reparameterizing the NeRF model with an additional output of uncertainty. It quantifies the aleatoric uncertainty and mitigates the influence of transient scene elements. A similar method is used [192] to quantify the uncertainty for the next best view planning problem. Following the Bayesian approach, S-NeRF [214] and its follow-up CF-NeRF [213] assess the epistemic uncertainty based on the variational inference by sampling and estimating an approximate posterior distribution across multiple potential radiance fields and optimizing the network. Nonetheless, the aforementioned uncertainty quantification methods need to reformulate the NeRF architecture and modify the training pipeline. It is difficult to introduce these reformulations to more contemporary models like Instant NGP, which comprise complex sampling mechanisms and feature encoding designs. Instead, Bayes' Rays [66] does not require modifying the NeRF training and estimates the posterior distribution through a Laplace

approximation based on model perturbations. Despite compatibility with any NeRF architecture, it can only assess epistemic uncertainty. In contrast, Density-aware NeRF Ensembles [229] quantifies both the aleatoric and epistemic uncertainty of the NeRF as the sum of the variance of predicted color and termination probability across multiple observations. It does not require modifying NeRF training. However, the ensemble-based approach requires multiple substantial computations for one NeRF model, which has efficiency limitations considering broad real-world applications.

3.8 Generalizability to Unseen Scenes

The vanilla NeRF encodes a 3D scene into the weights of coordinate-based MLPs, requiring per-scene optimization. As such, it still relies on a costly optimization from scratch each time it encounters new scenes and geometries. A generalizable NeRF should indeed be capable of novel view synthesis on an unseen scene with minimal retraining or without retraining. Aiming at this, pioneering research overlaps with partial research within conditional NeRFs discussed in Section 3.2. Codebooks-conditioned methods e.g. LOLNeRF [194] share category-level 3D prior knowledge with scene latent codes such that it can predict out-of-distribution scenes after fine-tuning the scene latent codes during inference. Adopting an image encoder to predict the latent code and view direction for GAN-based NeRF, Pix2NeRF [14] can generalize to unseen scenes without requiring inference time fine-tuning. However, these methods are dependent on prescribed instances within the same category. Image feature-based methods like PixelNeRF [300], MVSNeRF [23], and GRF [240] extract general image features with a 2D convolutional neural network (CNN). They present the potential to generalize to unseen scenes across categories while producing blurry renderings with artifacts.

More recently, a new trend is to leverage recent advances in large vision models (LVMs) for enhancing cross-scene generalization performance. Some approaches obtain a generic 3D prior using vision transformers. For instance, VisionNeRF [131] uses a 2D CNN for local image information and a pre-trained vision transformer [49] for global information. LRM [78] employs a pre-trained vision model to extract image features, which are then projected to tri-planes [21] using a transformer decoder via cross-attention mechanism. Furthermore, GNT [245] restructures the entire architecture with two transformer-based modules—a view transformer, for aggregating multi-view image features, and a ray transformer, for color predictions, instead of volume rendering. These methods demonstrate impressive performance in single-view reconstruction and cross-scene generalization. Nevertheless, they remain challenged when relying solely on semantic information to recover fine-grained details in unobserved regions.

4 RECONSTRUCTION TASKS

Use cases include 3D surface reconstruction (Section 4.1), large-scale scene reconstruction (Section 4.4), medical images reconstruction (Section 4.2), and digital human reconstruction (Section 4.3).

4.1 3D Surface Reconstruction

NeRFs can generate novel views, but extracting high-quality surfaces from the density-based scene representation using methods such as marching cubes results in poor outcomes due to insufficient constraints on surface [254]. To overcome this problem, several methods embed surface representation into NeRFs formulations, either using implicit or explicit representations.

4.1.1 Implicit representations. Implicit surface representations, such as occupancy fields and distance fields, can model many different geometries. Although UNISURF [164] manages to obtain accurate surfaces through occupancy fields, Signed Distance Function (SDF) have become the preferred alternative for surface representations in NeRF frameworks.

An SDF is an implicit representation that describes a geometry through a signed distance function. Such a function can be learned by a NeRF and is denoted $s_\theta : \mathbb{R}^3 \to \mathbb{R}$. Using this representation, the geometry of the scene can be easily extracted by computing the 0-level set of s_θ.[1] Leveraging this representation, approaches such as NeuS [254] and VolSDF [293] propose different formulations to derive the density σ from the SDF.
NeuS [254] defines the density function to be maximal at the zero-crossings of the SDF s_θ as,

$$\sigma(\mathbf{p}) = \max\left(\frac{-H'_\tau(s_\theta(\mathbf{p}))}{H_\tau(s_\theta(\mathbf{p}))}, 0\right), \text{where } H_\tau(u) = \frac{1}{1+e^{-\tau u}}, \tag{9}$$

where τ is a hyper-parameter that controls the sharpness of the logistic sigmoid function H.
Similarly, VolSDF [293] defines,

$$\sigma(\mathbf{p}) = \alpha\Psi_\beta(-s_\theta(\mathbf{p})), \text{where } \Psi_\beta(u) = \begin{cases} \frac{1}{2}\exp\left(\frac{u}{\beta}\right) & \text{if } u \leq 0 \\ 1 - \frac{1}{2}\exp\left(-\frac{u}{\beta}\right) & \text{if } u > 0 \end{cases}, \tag{10}$$

where α and β are learnable parameters.

Other methods explore variations of the SDF to customize surface representation: 3PSDF [27] introduces the NULL sign to represent non-closed surfaces by excluding surface regions from the optimization pipeline. NeuralRecon [228], on the other hand, utilizes Truncated Signed Distance Functions (TSDF) to represent local surfaces rather than optimizing the whole geometry at once. These parametrizations preserve the advantages of traditional NeRFs while learning smooth surfaces.

Due to their watertight nature, SDF formulations struggle to represent complex geometries like closed surfaces with holes or nested surfaces. Unsigned Distance Functions (UDFs) can represent such surfaces by eliminating the sign convention [36]. However, UDFs suffer from two limitations: they cannot account for occlusions, and their non-negative behavior causes unstable gradient computation on the 0-level set due to cusps, which are non-differentiable [143]. To overcome these challenges, NeuralUDF [143] introduces a differentiable visibility indicator function approximating the surface with a soft probabilistic distribution rather than relying on a hard threshold.

4.1.2 Explicit representations. Explicit representations describe surfaces using parametric geometries (e.g. meshes or point clouds). NVDIFFREC [75] employs a coarse-to-fine approach to estimate an SDF on a deformable tetrahedral grid. Via differentiable rendering for the scene's appearance, they extend previous work [215] to fit 2D supervision of NeRF approaches. VMesh [71] uses a hybrid volume-mesh representation merging geometry and volume primitives. Following discretization, they store texture, geometry, and volumetric information in efficient explicit assets. Wang et al. [262] utilize spatially varying kernel sizes τ (in Eq. (9)) for shell extraction of different frequency details (e.g. smooth surfaces or volumetric content such as clouds). Via dilation and erosion of the explicit shell, they reduce the number of samples for smooth surfaces while maintaining a sufficient amount for high-level details of volumetric content. Although these explicit surface representations are often less flexible than implicit surfaces, they offer faster training and rendering and enable the direct implementation of geometric constraints.

4.1.3 Surface Losses. Accessing a surface representation during training allows adding surface regularization terms to the neural rendering loss. With regard to implicit methods, Gropp et al. [68] show that it is possible to obtain a smooth and accurate representation by employing the Eikonal term, which encourages the representation to have a unit norm gradient. To further

[1]We recall that the ℓ-level set of a function f, is given by $L_\ell(f) = \{\mathbf{p} \in \mathbb{R}^3 | f(\mathbf{p}) = \ell\}$.

encourage smoothness, Neuralangelo [124] proposes to regularize the mean curvature of their SDF representation.

Explicit methods, on the other hand, often customize regularization terms to their representation. NVDIFFREC [75] penalizes the SDF sign change at the edges of the tetrahedral grid, reducing random faces and floaters. VMesh [71] optimizes the mesh by comparing its silhouette and depth map with the opacity and depth value rendered from the initial SDF stage. Wang et al. [262] encourage smoothness of a shell representation by regularizing its spatially varying kernel size.

4.1.4 Additional Surface Supervision. To extend the control over surface learning, various methods utilize information computed by the structure-from-motion (SfM) step or rely on off-the-shelf methods for additional priors. Geo-Neus [59] and RegSDF [308] directly supervise their SDF using the 3D sparse point cloud generated from SfM algorithms (e.g. COLMAP [209]). Following the assumption that those points lie on the surface, they enforce the SDF values to be zero. Darmon et al. [44] use additional multi-view constraints and structural similarity to better recover high-frequency details. When only a few views are available, surface methods significantly deteriorate. To overcome this issue, s-VolSDF [270] leverages multi-view stereo (MVS) priors to guide implicit surfaces from sparse views. In turn, the reconstructed depth information refines MVS depth sampling, providing more accurate estimations. MonoSDF [303], instead, uses monocular geometric priors predicted by a pre-trained network to directly optimize depth and normal estimations. Although these approaches yield more accurate and smoother surfaces, they come at the cost of additional supervisory data or pre-trained models.

4.2 Medical Images Reconstruction

Cameras are used in several medical applications such as skin inspection for wounds, cancer detection, and endoscopic examination. Similarly to other applications, NeRF's ability to generalize to unseen views enables [177] to expand the limited dataset in endoscopic imaging.

EndoNeRF [260] models human anatomy using NeRFs from surgical video. The authors show how NeRFs could improve Simultaneous Localization and Mapping (SLAM)-based approaches previously used for 3D reconstruction in estimating non-rigid tissue deformation. They draw inspiration from D-NeRF [178] which maps the deformation space to a trained canonical space. EndoNeRF exploits stereo views used in an operating room to estimate coarse depth maps for additional supervision.

Additional constraints using SDF are investigated in EndoSurf [304], extending EndoNeRF [260] and showing improved performance in surface estimation. These approaches leverage SDF formulations allowing for disentangling geometry and using additional geometric cues to regularize the reconstructed surface.

Batlle et al. [9] develop LightNeuS for endoscopic surface reconstruction using monocular videos. They identify two properties of endoscopic videos. First, the surface can be considered tubular and therefore easily modeled by SDFs; second, the light characteristics of endoluminal cavities can be approximated by an inverse-squared law. This allows for a rewriting of the rendering equation based on depth. Reconstruction improves, but this method lacks the ability to handle soft deformation.

NeRF can also be used for imaging modalities that are intrinsically 3D such as Magnetic Resonance Imaging (MRI), and Computed Tomography (CT). CuNeRF [34] generates high-resolution MRI and CT images from low-resolution inputs using NeRF formulations. Fang et al. [54] employ NeRF-style implicit functions to enhance resolution and accuracy in 3D tooth reconstruction from Cone Beam CT scans. In another example, [267] shows the generalizability of SDF representations for generating models of living cell shapes (microscopy imaging). They show how additional input

parameters such as time and a latent code can condition the representation to predict the evolution of the cell shapes.

Assessing the effectiveness of emerging technologies like neural rendering in medical applications is essential for medical research. In the context of chronic wound documentation, Syn3DWound [110] assesses a NeRF-based model for 3D wound reconstruction from synthetic images, comparing its performance to well-established photogrammetry solutions.

4.3 Digital Human Reconstruction

3D human reconstruction involves challenges ranging from motion to non-rigid deformation, often found in tasks of inverse kinematics and skeletal animation. Unlike for the reconstruction of arbitrary scenes, the human body can be represented with a mesh template such as the Skinned Multi-Person Linear model (SMPL) [145], which can accurately model pose-dependent shape variations. However, a central disadvantage of this model is its limited capacity to represent fine details such as clothing and hairstyle [91].

In this direction, several works integrated NeRF with the scope of rendering photorealistic images. Unlike standard NeRF settings, 3D human reconstruction typically involves monocular video of dynamic subjects. One approach to handling this is to encode temporal changes alongside the standard three dimensions in a unified 4D space. Neural Body [173] encodes structured latent codes anchored to an SMPL model into a single neural radiance field, providing implicit geometric guidance. Alternatively, other methods optimize a 3D canonical space from pose-estimation priors [265], extending Dynamic Scenes (see Section 3.6.1) to target anatomical structures, such as hands [73] and heads [77], and non-rigid deformations regularizing deformations across time [171, 172].

To enforce multi-view geometric correspondence of finer details, [62, 318] utilize human masks to optimize an implicit SDF field, preserving high-frequency details at image boundaries. [281] further addresses visibility challenges by leveraging the visibility-aware occupancy field of the learned geometry, enabling more precise surface reconstruction even in scenarios with significant occlusion. Building on this idea, [262] employs a hybrid volumetric-surface method by fitting explicit shells near learned surfaces, drastically accelerating rendering while maintaining geometric accuracy. This approach proves particularly effective for modeling complex parts of the human head, where the head is represented by an SDF and the hair structure is captured by a volumetric field.

While previous methods primarily focus on reconstructing geometry, generation-based techniques aim to synthesize novel appearances. For instance, [165] combines an SDF with a 2D style generator to synthesize novel photorealistic appearances. In contrast, [92] leverages explicit reconstruction priors derived from a 2D human generator to learn accurate and detailed 3D human geometry. For a more detailed overview of these techniques, in particular for methods focusing on human generation, we refer the reader to [91, 257].

4.4 Large-scale Scene Reconstruction

Large-scale scene reconstruction presents unique challenges for NeRF, due to the large data volumes that must be stored and processed, the variability in sensing conditions arising from extended capture times in complex environments such as city-scale landscapes, and modifications required in the rendering function to handle varied scenarios like underwater scenes.

4.4.1 Remote Sensing Scene Reconstruction. Reconstructing scenes using satellite or aircraft sensing data is challenging due to time, lighting, weather variations, limited viewing angles, and moving objects. Stereo-matching approaches [45, 175] are commonly used in satellite-based photogrammetry but face difficulties with lighting inconsistencies from shadow movements and sky conditions.

NeRF, with its implicit neural scene representation, offers a novel solution. S-NeRF [48] models direct sunlight through a local visibility field and learns indirect illumination from diffuse light via a non-local color field. Performance is improved for 3D reconstruction and novel view synthesis, but transient objects remain an issue. Sat-NeRF [149] employs a task-uncertainty learning approach [151] to model transient objects and uses rational polynomial coefficient camera models [67] to assist point sampling, yielding better results.

4.4.2 Underwater Reconstruction. The original NeRF [159] and many of its variants assume that multi-view images are captured in a non-hazy medium, focusing solely on radiance from foreground and background objects. This assumption fails in real-world scenarios like foggy or underwater environments, where scattering effects are present. For example, SeaThru-NeRF[114] directly integrates the scattering medium into its rendering equation through a wavelength-independent attenuation model, using two wideband coefficients to represent the medium.

4.4.3 City-Scale Reconstruction. Unlike the original NeRF, which performs well in small, controlled settings, city-scale scenes present three key challenges: various detail levels caused by wide viewing positions, real-world complexities (e.g. moving objects, weather changes, and dynamic lighting), and long rendering times with the large image sets.

City-NeRF [277] employs a progressive approach, increasing model detail as views get closer, efficiently managing multi-scale complexities in urban environments. BlockNeRF [230] improves scalability by dividing large datasets into manageable blocks, enabling parallel processing and reducing computational demands. This approach is particularly effective for rendering extensive city models. Switch-NeRF [155] introduces a learnable gating network to dynamically assign 3D points to sub-networks, optimizing scene decomposition and radiance field reconstruction without hand-crafted segmentation. Mega-NeRF [242] further optimizes large-scale scene reconstruction and rendering through a scalable architecture, enabling real-time, detailed urban fly-throughs for virtual reality and geographic exploration. UE4-NeRF [69] partitions large scenes into sub-NeRFs, represented by polygonal meshes, and integrates NeRF with Unreal Engine 4 (UE4) to achieve real-time rendering at 4K resolution with high frame rates. Finally, Urban neural fields [198] adapts NeRF for dynamic urban elements, enabling realistic reconstruction of scenes with moving vehicles and changing lighting.

5 COMPUTER VISION TASKS BEYOND RECONSTRUCTION

NeRF has emerged as a foundational 3D scene representation and powered a recent surge in applications beyond 3D reconstruction. This section highlights the significance of NeRFs in two common applications: robotics-related tasks (Section 5.1) and traditional computer vision tasks (Section 5.2). We further provide a concise review of 3D generation and editing in Section 5.3.

5.1 Applications on Robotics

This section discusses NeRF's capabilities in several key areas of robotics-related applications. We first discuss NeRF's capabilities for SLAM (see Section 5.1.1) — a fundamental and long-established component of robotic perception. Beyond that, NeRF offers a high-fidelity 3D scene representation that provides richer geometry and supports more accurate 3D scene understanding than 2D images, thereby improving performance in trajectory planning (see Section 5.1.2) and complex control tasks (see Section 5.1.3), e.g., object grasping.

5.1.1 SLAM. SLAM simultaneously maps an unknown environment while tracking the robot's location, inheriting challenges associated with camera pose estimation, 3D reconstruction, and object tracking. Traditional SLAM methods rely on feature extraction or depth estimation from camera

images or LiDAR scans. NeRF enhances SLAM by providing a continuous, differentiable scene representation from 2D images. NeRF-SLAM[204] combines NeRFs with dense monocular SLAM to improve 3D mapping and real-time scene reconstruction from monocular images, enhancing geometric and photometric accuracy through accurate pose and depth estimations. NICE-SLAM[322] addresses scalability and detail preservation in large-scale SLAM by using hierarchical, grid-based neural encoding with multi-level local information and pre-trained geometric priors, allowing for efficient local updates. Loopy-SLAM [136] incorporates loop closure into neural SLAM to mitigate error accumulation during tracking. It generates point-based sub-maps for efficient corrections without storing entire input frames and ensures global consistency through pose graph optimization and online loop closure detection. iSDF [166] enables incremental updates to SDF, allowing SLAM systems to adapt more rapidly to environmental change for real-time, high-resolution mapping. Dense RGB SLAM [115] employs a hierarchical feature volume to improve accuracy in large scenes and incorporates a multi-scale patch-based warping loss for a robust dense visual SLAM system by constraining camera motion and scene map. The fusion of NeRF and SLAM addresses key SLAM challenges, including dynamic environments and construction of an accurate map, facilitating more robust tracking and enhancement of downstream tasks, such as navigation.

5.1.2 Path Planning. Path planning is essential for autonomous robotic navigation, ensuring collision-free trajectories. NeRF improves path planning by offering high-fidelity 3D representations from 2D images, enabling detailed geometric understanding and smoother trajectory planning in dynamic environments. NeRF-Nav [2] utilizes NeRFs to create detailed geometric maps, optimizing trajectories by minimizing collision risks through density fields. It also incorporates a vision-based state estimator for real-time dynamic replanning and robust navigation. RNR-Map [108] integrates visual and semantic cues for efficient mapping and localization, demonstrating highly accurate and robust navigation even in changing environments. CATNIPS [26] transforms NeRFs into Poisson Point Processes to quantify collision probabilities. It integrates a voxel-based probabilistic unsafe robot region to ensure safe path planning. NeRF2Real [12] transfers vision-based bipedal locomotion policies from simulation to reality, leveraging NeRF to generate photorealistic 3D scenes combined with physics simulations to train reinforcement learning policies. MultiOb [152] introduces two neural implicit representations for agent navigation: one for predicting object positions and the other for encoding occupancy and exploration, significantly enhancing navigation capabilities in complex environments. Integrating NeRFs into path planning enhances autonomous navigation by improving 3D mapping, collision avoidance, trajectory optimization, and semantic understanding. Future efforts should optimize computational efficiency for real-time use and integrate NeRFs with sensors like LiDAR and radar for greater effectiveness.

5.1.3 Robotic Control. High-level tasks such as path planning and perception rely on controlling robots in complex environments with challenging objects. NeRF's high-fidelity 3D representations enhance robotic systems by providing detailed data for improved control. For example, NeRF-CBF [238] uses NeRFs for single-step visual foresight, ensuring control barrier functions to enhance safety and reliability through future observation predictions. Real-time simulations show its superior performance in preventing unsafe actions. Controlling3D [123] connects visual perception with motor control, generating detailed 3D environment representations that improve the accuracy and robustness of visuomotor control in dynamic and unstructured environments. NeRF-supervision [295] introduces a self-supervised pipeline to generate accurate dense correspondences, enhancing visual descriptors for challenging objects, such as thin and reflective surfaces. Evo-NeRF [99] improves robotic grasping efficiency for transparent objects by combining rapid NeRF training with sequential adaptation, allowing quick updates to the scene representation as objects are removed. GraspNeRF [43] employs a multi-view RGB-based 6-DoF grasp detection network

with generalizable NeRF for real-time material-agnostic object grasping, including transparent and specular objects in cluttered environments. To improve object manipulation and navigation, SPARTN[315] uses an offline data augmentation technique with NeRF to inject corrective noise into visual demonstrations, enabling real-time, RGB-only 6-DoF grasping policies that improve grasp success rates. NGDF[266] models grasping as the level set of a continuous implicit function, enabling joint optimization of grasping and motion planning through gradient-based trajectory optimization. NeRF's powerful 3D scene representation offers new opportunities for robots to handle challenging objects and navigate complex environments. However, achieving real-time performance remains a major challenge for effective interaction with the real world.

5.2 Recognition Tasks

Visual recognition is a fundamental problem in computer vision, involving the identification of semantically meaningful entities within visual data. This encompasses a wide spectrum of tasks, including but not limited to semantic segmentation, object detection, and object tracking. While current methods are effective on 2D data such as images and video frames [35], they heavily rely on pre-trained architectures fine-tuned on large annotated 2D datasets. Extending these methods to 3D data (e.g. point cloud and triangular mesh) is challenging due to the scarcity of 3D annotations, as annotating in 3D is labor-intensive. Methods that use only 2D annotations struggle to integrate 2D knowledge into accurate 3D representations, often resulting in errors.

In this context, NeRF emerges as a promising solution to seamlessly integrate multiple 2D observations into a unified, geometrically and semantically consistent 3D model, reducing the need for extensive 3D annotations. SemanticNeRF [314] performs 3D semantic segmentation by introducing a NeRF model that jointly learns scene appearance, semantics, and geometry from annotated 2D images. This work extends the original NeRF model by including a view-invariant function that maps world coordinates to semantic labels, enabling the rendering of precise 2D segmentation through the conventional NeRF's rendering equations. It can perform 2D-to-3D label propagation, even with sparse, noisy, and low-resolution 2D annotations. Similarly, works like [60, 218] achieve 3D panoptic segmentation by using coarse 3D annotations, such as 3D bounding boxes, and designing a label propagation scheme to handle ambiguities and noise.

Targeting the challenging task of multimodal open-set recognition, LERF [100] integrates CLIP embeddings [186] into NeRF to enable natural-language-based search and localization of objects, visual attributes, and geometric details within 3D scenes. SA3D [17] "lifts" 2D vision foundation model like Segment Anything (SAM) [102] to 3D, providing an interactive 3D segmentation through simple user prompts. NeRFs can also leverage scene physics and rendering effects as self-supervision signals to learn visual recognition models without relying on 2D or 3D annotations. For instance, D^2-NeRF [273] segments moving objects by decoupling the static and dynamic scene elements.

While the previous methods are scene-specific, other approaches aim for a general model to perform visual recognition across multiple NeRF-encoded scenes. Trained NeRFs can act as a versatile visual data format, capable of generating 2D images from various viewpoints or creating sequences that simulate video. Building on this, some methods perform visual recognition within the NeRF's implicit scene representation. For instance, NeRF-RPN [79] detects 3D objects in a NeRF scene by training a standard object detector on sampled NeRF primitives (*e.g.*, density and color). NeSF [247] achieves semantic segmentation across scenes by training a neural network to map sampled NeRF representations to generic features, which can then be segmented by deep-learning segmentation models.

A crucial aspect of applying NeRFs to visual recognition tasks lies in the interplay between the geometric and semantic features of a scene. Often, one aims to exploit geometric features to semantically distinguish objects within a scene. However, research has also demonstrated

that integrating semantic reasoning into NeRF models can contribute to a more comprehensive understanding of scene geometry and appearance, thereby improving the model's efficacy for novel view synthesis. For instance, DietNeRF [87] and SinNeRF [282] utilize semantic reasoning to address data deficiencies such as limited view diversity and inaccurate camera poses, boosting generalization across unseen viewpoints and environments.

5.3 3D Generation and Editing

Recently, generative AI techniques such as Stable Diffusion [203] and DALL-E [191] have gathered impressive results in text-to-image generation and image editing. Extending those generative capabilities from 2D to 3D content opens new opportunities for entertainment like video games and creative industries like immersive experiences, where rich 3D assets are essential but costly to craft manually. Nevertheless, lifting generation to 3D is not as simple as it seems. It requires 1) proper scene representation, and 2) semantic alignment between the input text descriptions and generated 3D content. NeRF [159] has been a popular choice for scene representations as its adopted volume rendering allows for consistent generation of 3D contents from 2D images.

For semantics alignment, early works like DreamFields [86] aim to minimize the CLIP distance [187] between rendered views and input text prompts. However, the use of pre-trained 2D image-text models compromises realism and accuracy in the generated 3D content. To address this, DreamFusion [176] extends the generative capabilities of 2D diffusion models [208] to 3D. Specifically, it replaces the CLIP loss with a loss derived from the distillation of a 2D diffusion model, introducing the Score Distillation Sampling (SDS) framework. This enhancement improves the quality and realism of the generated 3D content. Expanding on this foundational idea, various architectural improvements have been proposed to enhance quality and computational efficiency. Some approaches initialize NeRF with a 3D shape prior [283], or a 3D point cloud [301]. Magic3D [128] introduces a coarse-to-fine reconstruction strategy using Instant NGP [161] alongside mesh extraction and differentiable mesh rendering.

Images can also inform 3D generation tasks by explicitly providing geometry and appearance information of 3D contents. The image-to-3D task focuses on inferring the 3D structures and appearance of objects from single or sparse images. Some research [153, 232] manages to engineer textual prompts for the input image and utilize the DreamFusion-like scheme to learn the 3D shape. However, these methods are time-consuming and often suffer from the 3D inconsistency problem as the adopted 2D diffusion models fail to provide view-consistent supervision. Differently, Zero-1-to-3 [140] designs a view-conditioned diffusion model from a single image for 3D-consistent novel view synthesis and then lifts these synthesized views for 3D shape learning. Based on this scheme, One-2-3-45 [139] further enhances the 3D generation performance and efficiency with a few-shot SDF-based model (SparseNeuS [144]), leading to a 3D mesh generation from any single image in 45 seconds.

The generative properties of these models can be further tailored for controllable 3D content generation. Related works focus on incorporating multiple appearance or geometrical constraints in 3D generation. For instance, Dreambooth3D [188] constrains text-to-3D generation using a few input images. Similarly, Control3D [31] directs geometry through an input sketch, while Set-the-Scene [40] allows 3D scene generation with object proxies determining the placement of different elements. LaTeRF [160] extracts 3D assets from a pre-trained NeRF model, guided by a few pixel annotations and a natural language description of the target object. The ability to finetune 3D-generated content is another valuable feature of these methods. Magic3D [128] proposes prompt-based editing through fine-tuning, whereas DreamBooth3D [188] allows 3D object generation with simple prompt adjustments such as re-contextualization, accessorization, or stylization.

While major works revolve around single-object or dream-style generation, research interest grows towards expanding this generative capability to larger and more realistic environments. For example, Text2NeRF [306] explores the use of progressive in-painting and depth supervision to enhance the realism of 3D scene generation. The generative potential of these approaches has also sparked wide interest in digital human creation. Its applications range from 3D human generation [280] and 3D face model generation [18], to talking-head synthesis [70] and face morphing [250].

In addition to generation, NeRF-based 3D AIGC techniques show promise in scene editing. Given a prompt, one can modify the local properties of an object, such as appearance [105] and geometry [88], or manipulate objects within a scene [323]. Other works focus on altering a scene's global properties while keeping its original structure, performing relighting (see Section 3.4) or stylization [137]. Recently, Instruct-NeRF2NeRF [74] has shown compelling editing ability in both local (*e.g.* appearance) and global (*e.g.* style) modification using text prompts.

An important avenue of future research on 3D generation techniques includes extending related advancements to dynamic content, as explored by MAV3D [219], and further reducing the generation time through amortized optimization as demonstrated by ATT3D [146]. We refer our readers to recent surveys [121, 305] for a detailed discussion on recent advancements of 3D generation techniques their respective applications.

6 TOOLS AND DATASETS

This section compiles a list of tools to facilitate the development and analysis of NeRFs. We also provide commonly used datasets and evaluation protocols for benchmarking in Section 6.2.

6.1 Tools

This subsection introduces tools related to data generation, data preprocessing, NeRF training and optimization, and visualization/rendering.

6.1.1 Dataset generation. Synthetic datasets are usually created with fully controllable variables, making them a crucial first step in developing effective NeRF algorithms. Tools commonly used to create synthetic datasets for NeRF include, but are not limited to, Blender [56] and Unity [234]. Blender, an open-source 3D computer graphics software, is widely used for synthetic dataset generation, such as D-NeRF [178], by simulating various environments and lighting conditions. BlenderNeRF [184] integrates with Blender to streamline the process of generating high-quality and diverse datasets for novel-view synthesis tasks. Unity, a game engine, can also create synthetic data through custom scripts and scenarios. Unity NeRF[2] is developed specifically for generating NeRF datasets within the Unity environment.

6.1.2 Data preprocessing. Most NeRF approaches require camera poses as priors, calculating camera poses is an important step in preprocessing multi-view images. Several tools are available for this purpose, including COLMAP [209], pixSfM [134], and OpenDroneMap [64]. COLMAP is widely used for 3D reconstruction and preprocessing image data, providing accurate camera poses and point clouds. PixSfM improves the accuracy of COLMAP by refining 3D points and camera poses with direct alignment of 2D image features extracted from deep-learning methods. OpenDroneMap is useful for preprocessing large-scale datasets, allowing rapid 3D reconstruction with the possibility of including image GPS coordinates, camera yaw, pitch, and roll to accelerate SfM.

[2]https://github.com/kwea123/nerf_Unity. Accessed: 2024-11-05.

6.1.3 Training and optimization. To standardize the training and optimization of NeRF models and facilitate user-friendly interaction with NeRF models, NeRFStudio [231] provides an easy-to-use API for modular NeRF development. It simplifies model architecture design, training, and hyperparameter tuning, facilitating efficient development and optimization of NeRF models. Building on NeRFStudio, SDFStudio [302] focuses on 3D reconstruction and provides a unified and modular framework for neural implicit surface reconstruction. Focusing on efficient sampling in NeRF, NerfAcc [118] offers a PyTorch NeRF acceleration toolbox designed for both training and inference. It is a universal and plug-and-play tool compatible with most NeRF models.

6.2 Datasets and Evaluation Protocols

This section provides an overview of widely used datasets in NeRF-based research, as summarized in Table 2. It first reviews benchmarking datasets and evaluation metrics for two fundamental tasks — *novel view synthesis* and *3D surface reconstruction*, along with representative datasets and their corresponding leading methods provided in Table 3. It then discusses datasets that facilitate the broad adoption of NeRFs in real-world applications.

Novel View Synthesis. A typical dataset is the synthetic dataset from the original NeRF paper [159], NeRF Synthetic. It consists of eight 360° object-centric scenes covering different levels of geometric complexities (e.g., reflection, occlusion, and thin details) and offers accurate camera poses and a predefined train-test split. Another popular choice is LLFF [158], a real-world dataset containing 24 forward-facing scenes. Common categories include 8 representative scenes: Room, Fern, Leaves, Fortress, Orchids, Flower, T-Rex, and Horns. The standard protocol holds out one view out of every eight for testing, while the remaining views are used for training.

Several real-world datasets highlight challenges from unbounded indoor and outdoor environments. For example, MipNeRF 360 [6] provides 360° scene captures with rich high-frequency details, and ETH3D [210] offers high-resolution images for large scenes but with sparse spatial coverage. TanksAndTemples [103] captures large indoor and outdoor scenes that inevitably contain minor moving objects and variations in exposure. Phototourism [93] is used to evaluate performance under complex in-the-wild conditions, covering variations in lighting and the presence of transient objects over time.

Beyond these, additional datasets are designed to target specific real-world challenges, providing tailored benchmarks for evaluating an approach's effectiveness. Shiny Blender [246] and the Shiny dataset [269] specifically focus on the challenges posed by reflective surfaces. Deblur-NeRF dataset [147] highlight the low-quality training views with motion blur and camera focus blur. In the context of dynamic changes, D-NeRF [178] and DyNeRF [119] datasets feature dynamic objects with slow motions, Nerfies dataset [171] capture casual free-viewpoint selfies, while HyperNeRF dataset [172] focuses on the challenges associated with topological changes, e.g., mouth opening.

The novel view synthesis performance is commonly evaluated through three full-reference quality metrics:

- Peak Signal-to-Noise Ratio (PSNR), which is a distortion metric comparing the mean squared error between the rendered and ground-truth pixel in the *log* space,

[3]ShapeNetCore dataset does not provide rendered images. A common way is to follow 3D-R2N2 [39] to render images and split the dataset.
[4]Objaverse provides scripts for rendering images with Blender.
[5]Replica originally contains the 3D model of scenes. One can render datasets using their provided engines.
[6]GSO dataset does not provide images. One can follow Zero-1-to-3 [140] to generate rendering.
[7]Most NeRF-related methods use these two variants of pre-processed datasets—Plenoxels [58] (with backgrounds) and NSVF [138] (without backgrounds).
[8]Objaverse-XL dataset provides scripts for rendering images with Blender.

Table 2. Commonly used datasets for NeRFs-related methods. In the column of **Content**, J means objects, I means indoor scenes, and O means outdoor scenes. In the column of **Type**, S means static scenes and D means dynamic scenes.

Scene	Name	Content	Type	Scene Num./Size	Max. Resolution (Width, Height)					Depth	Semantics	3D	Website
					[0, 1k)	[1k, 2k)	[2k, 3k)	[3k, 4k)	[4k, +∞)				
Synthetic	NeRF Synthetic [159]	J	S	8 scenes 100 training 200 testing	✓	✗	✗	✗	✗	✓	✗	✗	
	ShapeNetCore[3] [20]	J	S	51300 3D models 55 object categories						✗	✗	✓	
	Shiny Blender [246]	J	S	6 scenes 100 training 200 testing	✓	✗	✗	✗	✗	✓	✗	✗	
	D-NeRF [178]	J	D	8 scenes 100 training 200 testing	✓	✗	✗	✗	✗	✓	✗	✗	
	Objaverse[4] [47]	J	S & D	818K objects 21K classes						✗	✗	✓	
	BlendedMVS [292]	J & O	S	113 scenes 20 to 1K images	✓	✗	✓	✗	✗	✓	✗	✓	
	MatrixCity [122]	O	D	2 scenes 67k aerial images 452k street-level images	✗	✓	✗	✗	✗	✓	✗	✗	
	Deblur-NeRF [147]	O & I	S	10 scenes 34 images	✓	✗	✗	✗	✗	✗	✗	✗	
	Replica[5] [224]	I	S	18 scenes						✓	✓	✓	
Real	Deblur-NeRF [147]	J	S	21 scenes 27 to 52 images	✓	✗	✓	✗	✗	✗	✗	✗	
	Shiny dataset [269]	J	S	8 scenes 32 to 307 images	✗	✓	✗	✗	✗	✗	✗	✓	
	DyNeRF [119]	J	D	6 scenes 10 seconds video (30 fps)	✗	✗	✗	✓	✗	✗	✗	✗	
	HyperNeRF [172]	J	D	17 scenes 82 to 415 images	✓	✓	✗	✗	✗	✗	✗	✗	
	Nerfies [171]	J	D	9 scenes 40 to 356 images	✓	✓	✗	✗	✗	✗	✗	✗	
	DTU [1]	J	S	80 scenes 49 to 64 images	✗	✓	✗	✗	✗	✗	✗	✓	
	CO3D [197]	J	S	19K objects 50 categories 1.5M images	✓	✗	✗	✗	✗	✓	✗	✓	
	GSO[6] [50]	J	S	1030 objects 17 categories						✗	✗	✓	
	OmniObject3D [272]	J	S	6000 objects 190 categories	✓	✓	✗	✗	✗	✓	✗	✓	
	ScanNet [42]	I	S	1503 scenes	✗	✓	✗	✗	✗	✓	✓	✓	
	ScanNet++ [297]	I	S	460 scenes 280000 DSLR images 3.7M RGB-D frames	✗	✓	✗	✗	✓	✓	✓	✓	
	RealEstate10K [317]	I	S	1500 scenes around 750K frames	✗	✓	✗	✗	✗	✗	✗	✗	
	ARKitScenes [8]	I	S	1661 scenes 5047 scans	✓	✓	✗	✗	✗	✓	✓	✓	
	EyefulTower [284]	I	S	13 scenes 765 to 8030 images	✗	✓	✓	✓	✓	✗	✗	✓	
	Dl3dv-10k [135]	I & O	S	10510 videos 51.2M frames	✓	✗	✓	✓	✗	✗	✗	✗	
	ETH3D [210]	I & O	S	25 scenes 7 to 110 images	✗	✗	✗	✗	✓	✓	✗	✓	
	LLFF-Real [158]	I & O	S	24 scenes 20 to 30 images	✗	✗	✗	✗	✓	✗	✗	✗	
	TanksAndTemples[7] [103]	I & O	S	14 scenes	✗	✓	✓	✓	✓	✗	✗	✓	
	MipNeRF-360 [6]	I & O	S	9 scenes 100 to 330 images	✓	✓	✓	✗	✓	✗	✗	✗	
	Phototourism [93]	O	S	25 scenes 75 to 3765 images	✗	✓	✗	✗	✗	✓	✗	✓	
	Mill 19 [242]	O	D	2 scenes 1678 and 1940 images	✗	✗	✗	✗	✓	✗	✗	✗	
	Kitti-360 [127]	O	D	11 scenes 320K images 80K laser scans	✗	✓	✗	✗	✗	✗	✓	✓	
	Waymo Block-NeRF [230]	O	D	12000 images	✗	✓	✗	✗	✗	✗	✗	✗	
Real & Synthetic	UrbanScene3D [132]	O	S	10 synthetic 6 real 128K images	✗	✗	✗	✗	✓	✓	✓	✓	
	Objaverse-XL[8] [46]	J	S & D	10.2M objects						✗	✗	✓	

- Structural Similarity Index Measure (SSIM) [261], which is a perceptual metric comparing the similarity of structural details, brightness, and contrast between the rendered and ground-truth image, and

Table 3. Representative datasets and leading methods for the novel view synthesis and 3D surface reconstruction tasks.

Task Category	Scene Content	Representative Datasets	Representative Leading Methods
Novel View Synthesis	Static	NeRF Synthetic [159], LLFF-Real [158]	Instant NGP [161], Mip-NeRF [5], Plenoxels [94]
	Unbounded	Mip-NeRF 360 [6], Tanks And Temples [103]	Zip-NeRF [7], Mip-NeRF 360 [6], SMERF [52]
	Dynamic	D-NeRF [178], DyNeRF [119], HyperNeRF [172]	K-Planes [57], TiNeuVox [53], HyperNeRF [172]
3D Surface Reconstruction	Small-scale objects	DTU [1], BlendedMVS [292]	MonoSDF [303], Neuralangelo [124]
	Large-scale objects	Tanks And Temples [103]	Neuralangelo [124]

- Learned Perceptual Image Patch Similarity (LPIPS) [311], which is a perceptual metric comparing the difference of derived features from a deep-learning-based model.

3D Surface Reconstruction. Three datasets are commonly used for evaluating the 3D surface reconstruction improvement: two object-centric datasets — DTU [1] and BlendedMVS [292] — and TanksAndTemples [103] for scenes with a larger scale. BlendedMVS originally contains 113 synthetic scenes covering architectures, sculptures, and small objects; within NeRF-based 3D surface reconstruction techniques, a subset of low-resolution scenes is commonly used, including Dog, Stone, Clock, Sculpture, Bear, Durian, Jade, and Man. DTU has real-world object scans of different categories and contains various appearances and geometries, among which 15 scans are commonly used. It offers RGB images associated with well-calibrated camera poses, 3D point clouds from a structured-light scanner, and observability masks for benchmarking.

During evaluation, the surface (in the form of meshes) is extracted from implicit scene representations, and then a set of points is sampled from the predicted surface. Typical evaluation metrics used to compare the accuracy of reconstructed surface include:

- Chamfer distance, quantifying geometric dissimilarity by measuring the average nearest-point distance between predicted and ground-truth sampled point clouds,
- F1 score, measuring the percentage of correctly reconstructed points within a threshold distance of the ground truth, and
- normal consistency, comparing surface normals at corresponding sampled points to assess the fidelity of local geometric details.

Towards Real-world Application. Large-scale datasets such as Objaverse [47], Objaverse-XL [46], and CO3D [197] provide extensive collections of multi-view 3D assets, boosting research for 3D generation and NeRFs' generalizability to unseen views and scenes. Indoor datasets like ScanNet++ [297] and ARKitScenes [8] offer both 3D ground truth and semantic annotations, supporting research in indoor scene reconstruction, 3D semantic segmentation, as well as robotics. Large-scale urban scene datasets, including Kitti-360 [127], Mill 19 [242], and Waymo Block-NeRF dataset [230], captured either from UAVs or moving vehicles, cover a large area of city- or site-scale environments and serve as valuable benchmarks for large-scale reconstruction and autonomous driving applications. In addition, realistic datasets with ultra-high-resolution imagery (4K/8K), such as Eyeful Tower [284], have further advanced the development of AR/VR applications.

Tele-health remains one of the most essential application areas for NeRF-based techniques. The recent release of diverse medical-related datasets has greatly benefited progress in the reconstruction of medical-related data. For instance, the MedNeRF dataset [41] contributes 20 CT chest scans and 5 CT knee scans; the EndoNeRF dataset [260] focuses on endoscopic surgical scenes with tissue deformations; and the SALVE dataset [37] offers three scans of 3D wound data associated with ultra-high-resolution 4K images, supporting development in wound care.

7 DISCUSSION AND OPEN CHALLENGES

This survey reviews key NeRF-related advancements towards higher accuracy and efficiency for representing 3D scenes, and their applicability to real-world problems. We covered strategies for enhancing rendering quality and efficiency, analyzed solutions to real-world challenges, and explored applications including reconstructions, 3D generation and editing, recognition, and robotics. Finally, we compiled a list of frequently used datasets and useful tools for NeRF-related research. This section takes a step back and explores open challenges for advancing NeRFs in the future.

Current NeRF enhancements yield impressive results with images up to 1K resolution. However, when handling higher resolutions such as 4K, the rendering time is multiplied by an order of magnitude compared to 1K images, limiting widespread adoption [117]. The rise of inexpensive mobile devices, capable of acquiring high-resolution images but limited in computing resources, highlights the need to optimize NeRFs towards higher efficiency and compactness with minimal GPU requirements. One approach is to incorporate explicit representations such as 3D Gaussian Splatting [98] as they can achieve real-time renderings. This could involve creating a hybrid representation or transcoding a NeRF into alternative explicit representations. Additionally, data compression techniques like quantization might be helpful to improve the model's compactness [76].

As NeRFs' performance continues to improve, it is crucial to revisit the choice of testing sets [278] and evaluation metrics [182]. Current evaluations heavily rely on reference 2D images, making them sensitive to testing camera distributions, testing view quality, and scene occlusions.

In-the-wild scenes pose multiple challenges, including heterogeneous image quality, limited observations, diverse dynamics (e.g., transient objects and complex lighting/weather variations), and large-scale scenes. Developing methods that can efficiently and accurately model such scenes while robustly handling distractors remains an active area of research [207].

Recent research [274, 316] applies recent blind image restoration techniques to improve the quality of NeRF trained on degraded inputs, and we expect more effective and efficient restoration systems to emerge. Several few-shot NeRFs utilize learned priors from pre-trained deep models (e.g., monocular depth priors) to guide training under limited views; however, ensuring metric accuracy and maintaining multi-view consistency in priors remains an open challenge [252]. In addition, limited views make SfM-based camera pose estimation underconstrained and prone to inaccuracy. Two promising directions include: i) developing online refinement algorithms for feature correspondences and camera parameters [170, 241], and ii) leveraging feed-forward 3D reconstruction methods, e.g., DUSt3R [258] and MASt3R [113].

Accurately and efficiently modeling realistic non-rigid deformations remains a long-standing and important challenge in dynamic NeRFs, where hybrid representations may help [57]. In-the-wild dynamic scenes involve not only local deformations (e.g., articulation/motion and topological changes) but also global variations (e.g., illumination/weather changes). Capturing both scales efficiently remains challenging [189]. Physics-aware methods prove effective in modeling complex outdoor scene lighting [205] and soft-body deformations [181]. Integrating richer physics simulation priors offers a promising path toward physically grounded dynamic NeRF models.

Despite significant advances in large-scale scene modeling, challenges of memory scalability and computational efficiency remain open. Achieving streamable performance in large-scale environments remains an active research direction for interactive VR/AR [52], e.g., enabling exploration of multi-room spaces in real-estate applications.

In addition, improving uncertainty quantification and generalizability of NeRFs is essential for practical real-world use. Open questions in uncertainty quantification include: 1) What constitutes well-quantified uncertainty for NeRFs? 2) Can we efficiently quantify both epistemic and aleatoric

uncertainty? For generalizability, regardless of achieving impressive renderings of unseen scenes, challenges remain with lengthy training time and insufficient details in unobserved areas.

The field of 3D vision has evidenced the broad adoption of NeRFs, transforming 3D surface reconstruction across various data types e.g. medical imaging[34]. Real-world applications, such as robotics, require handling diverse signals. Recent research demonstrates the potential of using multi-modal inputs—such as semantic labels [314], event streams [180], multispectral data [116], and audio [70]—to inform and enhance 3D scene learning. Although in its early stages, this research direction presents a promising yet underexplored area: how to efficiently leverage priors from diverse signals for 3D scene learning. Another exciting challenge lies in learning from modalities other than traditional visual data, *e.g.* developing NeRF-like frameworks from LiDAR sensors [233].

Beyond reconstruction tasks, NeRF increasingly impacts traditional computer vision tasks, such as 3D semantic segmentation. Learning 3D scenes naturally supports self-supervised learning. It is exciting to see how NeRFs can seamlessly integrate with and improve various large-scale 3D downstream tasks [100]. Moreover, advancements in large models may further boost the end-to-end learning of NeRFs, potentially bypassing the traditional camera pose estimation and then reconstruction steps [253]. This could lead to direct camera pose estimation and 3D geometry learning with large models like Transformers.

NeRF has also refashioned 3D generation and editing tasks, though existing NeRF-based 3D generation techniques are limited to dream-style object generation, often lacking fine-grained realistic details and requiring long optimization times [176]. Developing high-fidelity, real-time, faithful, and controllable 3D generation techniques would benefit AR/VR and creative industries. Conversely, generative models can be used to improve reconstruction. There is a trend in using generative priors learned from large-scale image assets to enrich details on unobserved areas and facilitate scene learning, although decreasing their computational overheads presents an ongoing challenge [140]. Moreover, it is crucial to consider the ethical and copyright implications of these generative methods.

ACKNOWLEDGMENTS

We thank Rongkai Ma and Ethan Goan for their valuable feedback on our manuscript. This work was partially supported by the CSIRO AI4M Research Program.

REFERENCES

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. 2016. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision* 120 (2016), 153–168.

[2] Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. 2022. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters* 7, 2 (2022), 4606–4613.

[3] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. 2011. Building rome in a day. *Commun. ACM* 54, 10 (2011), 105–112.

[4] Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. 2020. MatryODShka: Real-time 6DoF video view synthesis using multi-sphere images. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 441–459.

[5] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5835–5844.

[6] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5460–5469.

[7] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2023. Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 19697–19705.

[8] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. 2021. ARKitScenes - A Diverse Real-World Dataset For 3D Indoor Scene Understanding Using Mobile RGB-D Data. In *Advances in Neural Information Processing Systems*.

[9] Víctor M. Batlle, José M. M. Montiel, Pascal Fua, and Juan D. Tardós. 2023. LightNeuS: Neural Surface Reconstruction in Endoscopy Using Illumination Decline. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 502–512.

[10] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. 2023. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4160–4169.

[11] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. 2021. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12684–12694.

[12] Arunkumar Byravan, Jan Humplik, Leonard Hasenclever, Arthur Brussee, Francesco Nori, Tuomas Haarnoja, Ben Moran, Steven Bohez, Fereshteh Sadeghi, Bojan Vujatovic, et al. 2023. Nerf2real: Sim2real transfer of vision-guided bipedal motion skills using neural radiance fields. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9362–9369.

[13] Jintong Cai and Huimin Lu. 2024. NeRF-based Multi-View Synthesis Techniques: A Survey. In *2024 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 208–213.

[14] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. 2022. Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3981–3990.

[15] Ang Cao and Justin Johnson. 2023. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 130–141.

[16] J Douglas Carroll and Jih-Jie Chang. 1970. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika* 35, 3 (1970), 283–319.

[17] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, et al. 2023. Segment Anything in 3D with NeRFs. *Advances in Neural Information Processing Systems* 36, 25971–25990.

[18] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16123–16133.

[19] Subrahmanyan Chandrasekhar. 1960. *Radiative Transfer*. Courier Corporation.

[20] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

[21] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. TensoRF: Tensorial Radiance Fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 333–350.

[22] Anpei Chen, Zexiang Xu, Xinyue Wei, Siyu Tang, Hao Su, and Andreas Geiger. 2023. Dictionary fields: Learning a neural basis decomposition. *ACM Transactions on Graphics* 42, 4 (2023), 1–12.

[23] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14124–14133.

[24] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. 2023. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2416–2425.

[25] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. 2018. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*.

[26] Timothy Chen, Preston Culbertson, and Mac Schwager. 2024. Catnips: Collision avoidance through neural implicit probabilistic scenes. *IEEE Transactions on Robotics* 40 (2024), 2712–2728.

[27] Weikai Chen, Cheng Lin, Weiyang Li, and Bo Yang. 2022. 3psdf: Three-pole signed distance function for learning surfaces with arbitrary topologies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18522–18531.

[28] Wei-Ting Chen, Wang Yifan, Sy-Yen Kuo, and Gordon Wetzstein. 2024. DehazeNeRF: Multi-image Haze Removal and 3D Shape Reconstruction using Neural Radiance Fields. In *Proceedings of the International Conference on 3D Vision (3DV)*. 247–256.

[29] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. 2022. Hallucinated neural radiance fields in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12943–12952.

[30] Yue Chen, Xingyu Chen, Xuan Wang, Qi Zhang, Yu Guo, Ying Shan, and Fei Wang. 2023. Local-to-global registration for bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8264–8273.
[31] Yang Chen, Yingwei Pan, Yehao Li, Ting Yao, and Tao Mei. 2023. Control3d: Towards controllable text-to-3d generation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1148–1156.
[32] Zheng Chen, Yan-Pei Cao, Yuan-Chen Guo, Chen Wang, Ying Shan, and Song-Hai Zhang. 2023. PanoGRF: generalizable spherical radiance fields for wide-baseline panoramas. In *Advances in Neural Information Processing Systems*.
[33] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. 2023. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16569–16578.
[34] Zixuan Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. 2023. CuNeRF: Cube-Based Neural Radiance Field for Zero-Shot Medical Image Arbitrary-Scale Super Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 21185–21195.
[35] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention Mask Transformer for Universal Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1290–1299.
[36] Julian Chibane, Gerard Pons-Moll, et al. 2020. Neural unsigned distance fields for implicit function learning. In *Advances in Neural Information Processing Systems*, Vol. 33.
[37] Remi Chierchia, Leo Lebrat, David Ahmedt-Aristizabal, Olivier Salvado, Clinton Fookes, and Rodrigo Santa Cruz. 2025. SALVE: A 3D Reconstruction Benchmark of Wounds from Consumer-Grade Videos. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 4205–4214.
[38] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. 2022. Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 264–280.
[39] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 628–644.
[40] Dana Cohen-Bar, Elad Richardson, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. 2023. Set-the-scene: Global-local training for generating controllable nerf scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2920–2929.
[41] Abril Corona-Figueroa, Jonathan Frawley, Sam Bond-Taylor, Sarath Bethapudi, Hubert PH Shum, and Chris G Willcocks. 2022. Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology society (EMBC)*. IEEE, 3843–3848.
[42] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[43] Qiyu Dai, Yan Zhu, Yiran Geng, Ciyu Ruan, Jiazhao Zhang, and He Wang. 2023. Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1757–1763.
[44] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. 2022. Improving Neural Implicit Surfaces Geometry With Patch Warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6260–6269.
[45] Carlo De Franchis, Enric Meinhardt-Llopis, Julien Michel, J-M Morel, and Gabriele Facciolo. 2014. An automatic and modular stereo pipeline for pushbroom images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 2 (2014), 49–56.
[46] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. 2023. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems* 36 (2023), 35799–35813.
[47] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13142–13153.
[48] Dawa Derksen and Dario Izzo. 2021. Shadow neural radiance fields for multi-view satellite photogrammetry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1152–1161.
[49] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International*

Conference on Learning Representations (ICLR).

[50] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. 2022. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2553–2560.

[51] Yilun Du, Yinan Zhang, Hong Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. 2021. Neural Radiance Flow for 4D View Synthesis and Video Processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14304–14314.

[52] Daniel Duckworth, Peter Hedman, Christian Reiser, Peter Zhizhin, Jean-François Thibert, Mario Lučić, Richard Szeliski, and Jonathan T Barron. 2024. SMERF: Streamable memory efficient radiance fields for real-time large-scene exploration. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–13.

[53] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. 2022. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.

[54] Yu Fang, Zhiming Cui, Lei Ma, Lanzhuju Mei, Bojun Zhang, Yue Zhao, Zhihao Jiang, Yiqiang Zhan, Yongsheng Pan, Min Zhu, and Dinggang Shen. 2022. Curvature-Enhanced Implicit Function Network for High-quality Tooth Model Generation from CBCT Images. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 225–234.

[55] Ben Fei, Jingyi Xu, Rui Zhang, Qingyuan Zhou, Weidong Yang, and Ying He. 2024. 3d gaussian splatting as new era: A survey. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[56] Blender Foundation. 1994. Home of the Blender project - Free and Open 3D Creation Software. https://www.blender.org/. Accessed: 2024-11-05.

[57] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. 2023. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12479–12488.

[58] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance Fields without Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5491–5500.

[59] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. 2022. Geo-Neus: Geometry-Consistent Neural Implicit Surfaces Learning for Multi-view Reconstruction. In *Advances in Neural Information Processing Systems*.

[60] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. 2022. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *Proceedings of the International Conference on 3D Vision (3DV)*. 1–11.

[61] Kyle Gao, Yina Gao, Hongjie He, Dening Lu, Linlin Xu, and Jonathan Li. 2022. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379v5* (2022).

[62] Qingzhe Gao, Yiming Wang, Libin Liu, Lingjie Liu, Christian Theobalt, and Baoquan Chen. 2023. Neural novel actor: Learning a generalized animatable neural representation for human actors. *IEEE Transactions on Visualization and Computer Graphics* 30, 8 (2023), 5719–5732.

[63] Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. 2021. FastNeRF: High-Fidelity Neural Rendering at 200FPS. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14326–14335.

[64] Augustine-Moses Gaavwase Gbagir, Kylli Ek, and Alfred Colpaert. 2023. OpenDroneMap: multi-platform performance analysis. *Geographies* 3, 3 (2023), 446–458.

[65] Wenhang Ge, Tao Hu, Haoyu Zhao, Shu Liu, and Ying-Cong Chen. 2023. Ref-neus: Ambiguity-reduced neural implicit surface learning for multi-view reconstruction with reflection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4251–4260.

[66] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. 2024. Bayes' Rays: Uncertainty Quantification for Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20061–20070.

[67] Jacek Grodecki. 2001. IKONOS stereo feature extraction-RPC approach. In *ASPRS annual conference St. Louis*.

[68] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. 2020. Implicit geometric regularization for learning shapes. In *Proceedings of the 37th International Conference on Machine Learning*. 3789–3799.

[69] Jiaming Gu, Minchao Jiang, Hongsheng Li, Xiaoyuan Lu, Guangming Zhu, Syed Afaq Ali Shah, Liang Zhang, and Mohammed Bennamoun. 2024. Ue4-nerf: Neural radiance field for real-time rendering of large-scale scene. In *Advances in Neural Information Processing Systems*.

[70] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5784–5794.

[71] Yuan-Chen Guo, Yan-Pei Cao, Chen Wang, Yu He, Ying Shan, and Song-Hai Zhang. 2023. VMesh: Hybrid volume-mesh representation for efficient view synthesis. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.

[72] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. 2022. Nerfren: Neural radiance fields with reflections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18409–18418.

[73] Zhiyang Guo, Wengang Zhou, Min Wang, Li Li, and Houqiang Li. 2023. HandNeRF: Neural Radiance Fields for Animatable Interacting Hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21078–21087.

[74] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 19740–19750.

[75] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. 2022. Shape, light, and material decomposition from images using monte carlo rendering and denoising. In *Advances in Neural Information Processing Systems*, Vol. 35. 22856–22869.

[76] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. 2021. Baking Neural Radiance Fields for Real-Time View Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5855–5864.

[77] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. HeadNeRF: A Real-Time NeRF-Based Parametric Head Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20374–20384.

[78] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2024. LRM: Large Reconstruction Model for Single Image to 3D. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[79] Benran Hu, Junkai Huang, Yichen Liu, Yu-Wing Tai, and Chi-Keung Tang. 2023. NeRF-RPN: A general framework for object detection in NeRFs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[80] Dongting Hu, Zhenkai Zhang, Tingbo Hou, Tongliang Liu, Huan Fu, and Mingming Gong. 2023. Multiscale Representation for Real-Time Anti-Aliasing Neural Rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 17726–17737.

[81] Tao Hu, Shu Liu, Yilun Chen, Tiancheng Shen, and Jiaya Jia. 2022. EfficientNeRF: Efficient Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12892–12901.

[82] Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, and Yuewen Ma. 2023. Tri-MipRF: Tri-Mip Representation for Efficient Anti-Aliasing Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 19717–19726.

[83] Shi-Sheng Huang, Zixin Zou, Yichi Zhang, Yan-Pei Cao, and Ying Shan. 2024. Sc-neus: Consistent neural surface reconstruction from sparse and noisy views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 2357–2365.

[84] Xudong Huang, Wei Li, Jie Hu, Hanting Chen, and Yunhe Wang. 2023. RefSR-NeRF: Towards High Fidelity and Super Resolution View Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8244–8253.

[85] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. 2021. HDR-NeRF: High Dynamic Range Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18377–18387.

[86] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 867–876.

[87] Ajay Jain, Matthew Tancik, and Pieter Abbeel. 2021. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5885–5894.

[88] Clément Jambon, Bernhard Kerbl, Georgios Kopanas, Stavros Diolatzis, Thomas Leimkühler, and George Drettakis. 2023. Nerfshop: Interactive editing of neural radiance fields. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 6, 1 (2023).

[89] Wonbong Jang and Lourdes Agapito. 2021. CodeNeRF: Disentangled Neural Radiance Fields for Object Categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12929–12938.

[90] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. 2021. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5846–5854.

[91] Zhen Jia, Zhang Zhang, Liang Wang, and Tieniu Tan. 2024. Human image generation: A comprehensive survey. *Comput. Surveys* 56, 11 (2024), 1–39.

[92] Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, and Lan Xu. 2023. HumanGen: Generating Human Radiance Fields With Explicit Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12543–12554.

[93] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. 2021. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision* 129, 2 (2021), 517–547.

[94] Kim Jun-Seong, Kim Yu-Ji, Moon Ye-Bin, and Tae-Hyun Oh. 2022. Hdr-plenoxels: Self-calibrating high dynamic range radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 384–401.

[95] James T Kajiya. 1986. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*. 143–150.

[96] James T Kajiya and Brian P Von Herzen. 1984. Ray tracing volume densities. *ACM SIGGRAPH computer graphics* 18, 3 (1984), 165–174.

[97] Alex Kendall and Yarin Gal. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?. In *Advances in Neural Information Processing Systems*. 5575–5585.

[98] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (2023), 139–1.

[99] Justin Kerr, Letian Fu, Huang Huang, Yahav Avigal, Matthew Tancik, Jeffrey Ichnowski, Angjoo Kanazawa, and Ken Goldberg. 2022. Evo-nerf: Evolving nerf for sequential robot grasping of transparent objects. In *Proceedings of The 6th Conference on Robot Learning*.

[100] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. 2023. LERF: Language Embedded Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[101] Mijeong Kim, Seonguk Seo, and Bohyung Han. 2022. InfoNeRF: Ray Entropy Minimization for Few-Shot Neural Volume Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12902–12911.

[102] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[103] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. 2017. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics* 36, 4 (2017).

[104] Georgios Kopanas, Julien Philip, Thomas Leimkühler, and George Drettakis. 2021. Point-Based Neural Rendering with Per-View Optimization. *Computer Graphics Forum* 40, 4 (2021), 29–43.

[105] Zhengfei Kuang, Fujun Luan, Sai Bi, Zhixin Shu, Gordon Wetzstein, and Kalyan Sunkavalli. 2023. Palettenerf: Palette-based appearance editing of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20691–20700.

[106] Shreyas Kulkarni, Peng Yin, and Sebastian Scherer. 2023. 360fusionnerf: Panoramic neural radiance fields with joint guidance. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7202–7209.

[107] Kiriakos N Kutulakos and Steven M Seitz. 2000. A theory of shape by space carving. *International journal of computer vision* 38, 3 (2000), 199–218.

[108] Obin Kwon, Jeongho Park, and Songhwai Oh. 2023. Renderable neural radiance map for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9099–9108.

[109] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. 2020. Modular Primitives for High-Performance Differentiable Rendering. *ACM Transactions on Graphics* 39, 6 (2020).

[110] Léo Lebrat, Rodrigo Santa Cruz, Remi Chierchia, Yulia Arzhaeva, Mohammad Ali Armin, Joshua Goldsmith, Jeremy Oorloff, Prithvi Reddy, Chuong Nguyen, Lars Petersson, et al. 2024. Syn3DWound: A Synthetic Dataset for 3D Wound Bed Analysis. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1–5.

[111] Dogyoon Lee, Minhyeok Lee, Chajin Shin, and Sangyoun Lee. 2023. DP-NeRF: Deblurred Neural Radiance Field with Physical Scene Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12386–12396.

[112] Dongwoo Lee, Jeongtaek Oh, Jaesung Rim, Sunghyun Cho, and Kyoung Mu Lee. 2023. ExBluRF: Efficient Radiance Fields for Extreme Motion Blurred Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 17593–17602.

[113] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. 2024. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*. Springer, 71–91.

[114] Deborah Levy, Amit Peleg, Naama Pearl, Dan Rosenbaum, Derya Akkaynak, Simon Korman, and Tali Treibitz. 2023. Seathru-nerf: Neural radiance fields in scattering media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 56–65.

[115] Heng Li, Xiaodong Gu, Weihao Yuan, Zilong Dong, Ping Tan, et al. 2023. Dense RGB SLAM with Neural Implicit Maps. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[116] Jiabao Li, Yuqi Li, Ciliang Sun, Chong Wang, and Jinhui Xiang. 2024. Spec-nerf: Multi-spectral neural radiance fields. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2485–2489.

[117] Quewei Li, Feichao Li, Jie Guo, and Yanwen Guo. 2023. UHDNeRF: Ultra-High-Definition Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 23097–23108.

[118] Ruilong Li, Hang Gao, Matthew Tancik, and Angjoo Kanazawa. 2023. NerfAcc: Efficient Sampling Accelerates NeRFs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 18491–18500.

[119] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. 2022. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5521–5531.

[120] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, and Zhaoyang Lv. 2022. Neural 3D Video Synthesis from Multi-view Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5511–5521.

[121] Xiaoyu Li, Qi Zhang, Di Kang, Weihao Cheng, Yiming Gao, Jingbo Zhang, Zhihao Liang, Jing Liao, Yan-Pei Cao, and Ying Shan. 2024. Advances in 3d generation: A survey. *arXiv preprint arXiv:2401.17807* (2024).

[122] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. 2023. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3205–3215.

[123] Yunzhu Li, Shuang Li, Vincent Sitzmann, Pulkit Agrawal, and Antonio Torralba. 2022. 3d neural scene representations for visuomotor control. In *Conference on Robot Learning*. PMLR, 112–123.

[124] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H. Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. 2023. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8456–8465.

[125] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. 2021. Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6494–6504.

[126] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. 2023. DynIBaR: Neural Dynamic Image-Based Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4273–4284.

[127] Yiyi Liao, Jun Xie, and Andreas Geiger. 2022. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3292–3310.

[128] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 300–309.

[129] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. 2021. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5741–5751.

[130] Chien-Yu Lin, Qichen Fu, Thomas Merth, Karren Yang, and Anurag Ranjan. 2024. Fastsr-nerf: Improving nerf efficiency on consumer devices with a simple super-resolution pipeline. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 6036–6045.

[131] Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. 2023. Vision transformer for nerf-based view synthesis from a single input image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 806–815.

[132] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. 2022. Capturing, reconstructing, and simulating: the urbanscene3d dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 93–109.

[133] David B. Lindell, Julien N.P. Martel, and Gordon Wetzstein. 2021. Autoint: Automatic integration for fast neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14551–14560.

[134] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. 2021. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5987–5997.

[135] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. 2024. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22160–22169.

[136] Lorenzo Liso, Erik Sandström, Vladimir Yugay, Luc Van Gool, and Martin R Oswald. 2024. Loopy-slam: Dense neural slam with loop closures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20363–20373.

[137] Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. 2023. Stylerf: Zero-shot 3d style transfer of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8338–8348.

[138] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat Seng Chua, and Christian Theobalt. 2020. Neural Sparse Voxel Fields. In *Advances in Neural Information Processing Systems*.

[139] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. 2024. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *Advances in Neural Information Processing Systems*, Vol. 36.

[140] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9298–9309.

[141] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. 2019. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7708–7717.

[142] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural volumes: learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–14.

[143] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Yuan Liu, Peng Wang, Christian Theobalt, Taku Komura, and Wenping Wang. 2023. Neuraludf: Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20834–20843.

[144] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. 2022. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 210–227.

[145] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2023. *SMPL: A Skinned Multi-Person Linear Model* (1 ed.). Association for Computing Machinery, New York, NY, USA.

[146] Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. 2023. Att3d: Amortized text-to-3d object synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 17946–17956.

[147] Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V. Sander. 2022. Deblur-NeRF: Neural Radiance Fields from Blurry Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12851–12860.

[148] Michael W Marcellin, Michael J Gormish, Ali Bilgin, and Martin P Boliek. 2000. An overview of JPEG-2000. In *Proceedings DCC 2000. Data compression conference*. IEEE, 523–541.

[149] Roger Marí, Gabriele Facciolo, and Thibaud Ehret. 2022. Sat-nerf: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using rpc cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1311–1321.

[150] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7210–7219.

[151] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7210–7219.

[152] Pierre Marza, Laetitia Matignon, Olivier Simonin, and Christian Wolf. 2023. Multi-Object Navigation with dynamically learned neural implicit representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 11004–11015.

[153] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. 2023. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8446–8455.

[154] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4460–4470.

[155] Zhenxing Mi and Dan Xu. 2023. Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[156] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. 2021. NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16169–16178.

[157] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)* 38, 4 (2019), 1–14.

[158] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. 2019. Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines. *ACM Transactions on Graphics* (2019).

[159] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 405–421.

[160] Ashkan Mirzaei, Yash Kant, Jonathan Kelly, and Igor Gilitschenski. 2022. Laterf: Label and text driven object radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 20–36.

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

[162] Michael Niemeyer and Andreas Geiger. 2020. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11448–11459.

[163] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. 2020. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3504–3515.

[164] Michael Oechsle, Songyou Peng, and Andreas Geiger. 2021. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5589–5599.

[165] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. 2022. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13503–13513.

[166] Joseph Ortiz, Alexander Clegg, Jing Dong, Edgar Sucar, David Novotny, Michael Zollhoefer, and Mustafa Mukadam. 2022. iSDF: Real-time neural signed distance fields for robot perception. In *Robotics: Science and Systems*.

[167] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. 2021. Neural Scene Graphs for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2855–2864.

[168] Jen-I Pan, Jheng-Wei Su, Kai-Wen Hsiao, Ting-Yu Yen, and Hung-Kuo Chu. 2022. Sampling Neural Radiance Fields for Refractive Objects. In *SIGGRAPH Asia 2022 Technical Communications*. Article 5, 4 pages.

[169] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 165–174.

[170] Keunhong Park, Philipp Henzler, Ben Mildenhall, Jonathan T Barron, and Ricardo Martin-Brualla. 2023. Camp: Camera preconditioning for neural radiance fields. *ACM Transactions on Graphics* 42, 6 (2023), 1–11.

[171] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5845–5854.

[172] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021. HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields. *ACM Transactions on Graphics* 40, 6 (2021).

[173] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural Body: Implicit Neural Representations With Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9054–9063.

[174] Eric Penner and Li Zhang. 2017. Soft 3d reconstruction for view synthesis. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 1–11.

[175] Roland Perko, Hannes Raggam, and Peter M Roth. 2019. Mapping with pléiades—end-to-end workflow. *Remote Sensing* 11, 17 (2019), 2052.

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

[177] Dimitrios Psychogyios, Francisco Vasconcelos, and Danail Stoyanov. 2023. Realistic Endoscopic Illumination Modeling for NeRF-Based Data Generation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 535–544.

[178] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10318–10327.

[179] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 652–660.

[180] Yunshan Qi, Lin Zhu, Yu Zhang, and Jia Li. 2023. E2NeRF: Event Enhanced Neural Radiance Fields from Blurry Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13254–13264.

[181] Yi-Ling Qiao, Alexander Gao, and Ming Lin. 2022. Neuphysics: Editable neural geometry and physics from monocular videos. *Advances in Neural Information Processing Systems* 35 (2022), 12841–12854.

[182] Qiang Qu, Hanxue Liang, Xiaoming Chen, Yuk Ying Chung, and Yiran Shen. 2024. NeRF-NQA: No-Reference Quality Assessment for Scenes Generated by NeRF and Neural View Synthesis Methods. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[183] Yansong Qu, Yuze Wang, and Yue Qi. 2023. SG-NeRF: Semantic-guided Point-based Neural Radiance Fields. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. 570–575.

[184] Maxime Raafat. 2024. *BlenderNeRF*. https://doi.org/10.5281/zenodo.13250725

[185] AKM Rabby and Chengcui Zhang. 2023. BeyondPixels: A comprehensive review of the evolution of neural radiance fields. *arXiv preprint arXiv:2306.03000* (2023).

[186] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*. 8748–8763.

[187] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 8748–8763.

[188] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. 2023. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2349–2359.

[189] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Anton Van Den Hengel. 2024. Blirf: Bandlimited radiance fields for dynamic scene modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4641–4649.

[190] Andrea Ramazzina, Mario Bijelic, Stefanie Walz, Alessandro Sanvito, Dominik Scheuble, and Felix Heide. 2023. ScatterNeRF: Seeing Through Fog with Physically-Based Inverse Neural Rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 17911–17922.

[191] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125* (2022).

[192] Yunlong Ran, Jing Zeng, Shibo He, Jiming Chen, Lincheng Li, Yingfeng Chen, Gimhee Lee, and Qi Ye. 2023. NeurAR: Neural Uncertainty for Autonomous 3D Reconstruction With Implicit Neural Representations. *IEEE Robotics and Automation Letters* 8, 2 (2 2023), 1125–1132.

[193] Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, and Andrea Tagliasacchi. 2020. DeRF: Decomposed Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14148–14156.

[194] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. 2022. LOLNeRF: Learn from One Look. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1548–1557.

[195] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. 2021. KiloNeRF: Speeding up Neural Radiance Fields with Thousands of Tiny MLPs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14315–14325.

[196] Christian Reiser, Rick Szeliski, Dor Verbin, Pratul Srinivasan, Ben Mildenhall, Andreas Geiger, Jon Barron, and Peter Hedman. 2023. MERF: Memory-Efficient Radiance Fields for Real-time View Synthesis in Unbounded Scenes. *ACM Transactions on Graphics* 42, 4 (2 2023).

[197] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. 2021. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 10901–10911.

[198] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. 2022. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12932–12942.

[199] Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. 2024. Nerf on-the-go: Exploiting uncertainty for distractor-free nerfs in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8931–8940.

[200] Daniel Rho, Byeonghyeon Lee, Seungtae Nam, Joo Chan Lee, Jong Hwan Ko, and Eunbyung Park. 2022. Masked Wavelet Representation for Compact Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR). 20680–20690.

[201] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Niebner. 2022. Dense Depth Priors for Neural Radiance Fields from Sparse Input Views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12882–12891.

[202] Barbara Roessle, Norman Müller, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Niessner. 2023. GANeRF: Leveraging Discriminators to Optimize Neural Radiance Fields. *ACM Transactions on Graphics* 42, 6 (12 2023).

[203] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[204] Antoni Rosinol, John J Leonard, and Luca Carlone. 2023. Nerf-slam: Real-time dense monocular slam with neural radiance fields. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3437–3444.

[205] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. 2022. NeRF for Outdoor Scene Relighting. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 615–631.

[206] Viktor Rudnev, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. 2022. EventNeRF: Neural Radiance Fields from a Single Colour Event Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4992–5002.

[207] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. 2023. Robustnerf: Ignoring distractors with robust losses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20626–20636.

[208] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in neural information processing systems*, Vol. 35. 36479–36494.

[209] Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4104–4113.

[210] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3260–3269.

[211] Steven M Seitz and Charles R Dyer. 1999. Photorealistic scene reconstruction by voxel coloring. *International journal of computer vision* 35, 2 (1999), 151–173.

[212] Seunghyeon Seo, Donghoon Han, Yeonjin Chang, and Nojun Kwak. 2023. Mixnerf: Modeling a ray with mixture density for novel view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20659–20668.

[213] Jianxiong Shen, Antonio Agudo, Francesc Moreno-Noguer, and Adria Ruiz. 2022. Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 540–557.

[214] Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno-Noguer. 2021. Stochastic Neural Radiance Fields: Quantifying Uncertainty in Implicit 3D Representations. In *Proceedings of the International Conference on 3D Vision (3DV)*. 972–981.

[215] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems*.

[216] Ruoxi Shi, Xinyue Wei, Cheng Wang, and Hao Su. 2024. ZeroRF: Fast Sparse View 360deg Reconstruction with Zero Pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21114–21124.

[217] Seungjoo Shin and Jaesik Park. 2024. Binary radiance fields. In *Advances in Neural Information Processing Systems*.

[218] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. 2023. Panoptic Lifting for 3D Scene Understanding With Neural Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9043–9052.

[219] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. 2023. Text-To-4D Dynamic Scene Generation. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 31915–31929.

[220] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. 2020. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, Vol. 33. 7462–7473.

[221] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. 2019. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in neural information processing systems* 32 (2019).

[222] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. 2023. NeRFPlayer: A Streamable Dynamic Scene Representation with Decomposed Neural Radiance Fields. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–11.

[223] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. 2021. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7495–7504.

[224] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797* (2019).

[225] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. 2022. Light field neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8269–8279.

[226] Cheng Sun, Min Sun, and Hwann Tzong Chen. 2022. 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)*. 5449–5459.

[227] Jiaming Sun, Xi Chen, Qianqian Wang, Zhengqi Li, Hadar Averbuch-Elor, Xiaowei Zhou, and Noah Snavely. 2022. Neural 3d reconstruction in the wild. In *ACM SIGGRAPH 2022 conference proceedings*. 1–9.

[228] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. 2021. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15598–15607.

[229] Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. 2023. Density-aware NeRF Ensembles: Quantifying Predictive Uncertainty in Neural Radiance Fields. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. 9370–9376.

[230] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben P. Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. 2022. Block-NeRF: Scalable Large Scene Neural View Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8238–8248.

[231] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. 2023. Nerfstudio: A Modular Framework for Neural Radiance Field Development. In *ACM SIGGRAPH 2023 Conference Proceedings*.

[232] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 2023. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22819–22829.

[233] Tang Tao, Longfei Gao, Guangrun Wang, Yixing Lao, Peng Chen, Hengshuang Zhao, Dayang Hao, Xiaodan Liang, Mathieu Salzmann, and Kaicheng Yu. 2024. Lidar-nerf: Novel lidar view synthesis via neural radiance fields. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 390–398.

[234] Unity Technologies. 2005. Unity Real-Time Development Platform | 3D, 2D, VR & AR Engine. https://unity.com/. Accessed: 2024-11-05.

[235] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. 2022. Advances in neural rendering. *Computer Graphics Forum* 41, 2 (2022), 703–735.

[236] Kushagra Tiwary, Akshat Dave, Nikhil Behari, Tzofi Klinghoffer, Ashok Veeraraghavan, and Ramesh Raskar. 2023. Orca: Glossy objects as radiance-field cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20773–20782.

[237] Jinguang Tong, Sundaram Muthu, Fahira Afzal Maken, Chuong Nguyen, and Hongdong Li. 2023. Seeing through the glass: Neural 3d reconstruction of object inside a transparent container. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12555–12564.

[238] Mukun Tong, Charles Dawson, and Chuchu Fan. 2023. Enforcing safety for vision-based controllers via control barrier functions and neural radiance fields. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 10511–10517.

[239] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhofer, Christoph Lassner, and Christian Theobalt. 2021. Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Dynamic Scene From Monocular Video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12939–12950.

[240] Alex Trevithick and Bo Yang. 2021. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 15182–15192.

[241] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. 2023. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR). 4190–4200.

[242] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. 2022. 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)*. 12912–12921.

[243] Haithem Turki, Michael Zollhöfer, Christian Richardt, and Deva Ramanan. 2024. Pynerf: Pyramidal neural radiance fields. In *Advances in Neural Information Processing Systems*, Vol. 36.

[244] Mikaela Angelina Uy, George Kiyohiro Nakayama, Guandao Yang, Rahul Krishna Thomas, Leonidas Guibas, and Ke Li. 2023. NeRF Revisited: Fixing Quadrature Instability in Volume Rendering. In *Advances in Neural Information Processing Systems*, Vol. 36.

[245] Mukund Varma, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. 2022. Is attention all that neRF needs?. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[246] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. 2022. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5481–5490.

[247] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi S. M. Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. 2022. NeSF: Neural Semantic Fields for Generalizable Semantic Segmentation of 3D Scenes. *Transactions on Machine Learning Research* (2022).

[248] Chaoyang Wang, Lachlan Ewen MacDonald, Laszlo A. Jeni, and Simon Lucey. 2023. Flow supervision for Deformable NeRF. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21128–21137.

[249] Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. 2022. Nerf-sr: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6445–6454.

[250] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. 2022. Morf: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.

[251] Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, Yafei Song, and Huaping Liu. 2023. Mixed Neural Voxels for Fast Multi-view Video Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 19649–19659.

[252] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. 2023. SparseNeRF: Distilling Depth Ranking for Few-shot Novel View Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9031–9042.

[253] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5294–5306.

[254] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems*.

[255] Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, and Wenping Wang. 2023. F2-nerf: Fast neural radiance field training with free camera trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4150–4159.

[256] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. 2021. IBRNet: Learning Multi-View Image-Based Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4690–4699.

[257] Ruihe Wang, Yukang Cao, Kai Han, and Kwan-Yee K Wong. 2024. A Survey on 3D Human Avatar Modeling–From Reconstruction to Generation. *arXiv preprint arXiv:2406.04253* (2024).

[258] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. 2024. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20697–20709.

[259] Shuaixian Wang, Haoran Xu, Yaokun Li, Jiwei Chen, and Guang Tan. 2025. IE-NeRF: Exploring transient mask inpainting to enhance neural radiance fields in the wild. *Neurocomputing* 618 (2025), 129112.

[260] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. 2022. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, 431–441.

[261] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

[262] Zian Wang, Tianchang Shen, Merlin Nimier-David, Nicholas Sharp, Jun Gao, Alexander Keller, Sanja Fidler, Thomas Müller, and Zan Gojcic. 2023. Adaptive Shells for Efficient Neural Radiance Field Rendering. *ACM Transactions on Graphics* 42, 6, Article 260 (2023).

[263] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. 2022. NeRF–: Neural Radiance Fields Without Known Camera Parameters. *arXiv preprint arXiv:2102.07064* (2022).

[264] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. 2021. NerfingMVS: Guided Optimization of Neural Radiance Fields for Indoor Multi-view Stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5590–5599.

[265] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. 2022. HumanNeRF: Free-Viewpoint Rendering of Moving People From Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16210–16220.

[266] Thomas Weng, David Held, Franziska Meier, and Mustafa Mukadam. 2023. Neural grasp distance fields for robot manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1814–1821.

[267] David Wiesner, Julian Suk, Sven Dummer, David Svoboda, and Jelmer M. Wolterink. 2022. Implicit Neural Representations for Generative Modeling of Living Cell Shapes. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 58–67.

[268] Lance Williams. 1983. Pyramidal parametrics. In *Proceedings of the 10th annual conference on Computer graphics and interactive techniques*. 1–11.

[269] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. 2021. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8534–8543.

[270] Haoyu Wu, Alexandros Graikos, and Dimitris Samaras. 2023. S-VolSDF: Sparse Multi-View Stereo Regularization of Neural Implicit Surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3556–3568.

[271] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. 2024. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21551–21561.

[272] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. 2023. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 803–814.

[273] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. 2022. Dˆ 2NeRF: Self-Supervised Decoupling of Dynamic and Static Objects from a Monocular Video. In *Advances in Neural Information Processing Systems*, Vol. 35. 32653–32666.

[274] Zhongkai Wu, Ziyu Wan, Jing Zhang, Jing Liao, and Dong Xu. 2024. RaFE: Generative Radiance Fields Restoration. In *European Conference on Computer Vision*. Springer, 163–179.

[275] Jamie Wynn and Daniyar Turmukhambetov. 2023. Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[276] Yitong Xia, Hao Tang, Radu Timofte, and Luc Van Gool. 2022. SiNeRF: Sinusoidal Neural Radiance Fields for Joint Pose Estimation and Scene Reconstruction. In *British Machine Vision Conference*.

[277] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. 2022. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 106–122.

[278] Wenhui Xiao, Rodrigo Santa Cruz, David Ahmedt-Aristizabal, Olivier Salvado, Clinton Fookes, and Leo Lebrat. 2024. NeRF Director: Revisiting View Selection in Neural Volume Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20742–20751.

[279] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. 2022. Neural fields in visual computing and beyond. *Computer Graphics Forum* 41, 2 (2022), 641–676.

[280] Zhangyang Xiong, Di Kang, Derong Jin, Weikai Chen, Linchao Bao, Shuguang Cui, and Xiaoguang Han. 2023. Get3dhuman: Lifting stylegan-human into a 3d generative model using pixel-aligned reconstruction priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9287–9297.

[281] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. 2022. Icon: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 13286–13296.

[282] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. 2022. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 736–753.

[283] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. 2023. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20908–20918.

[284] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, Dahua Lin, Michael Zollhöfer, and Christian Richardt. 2023. VR-NeRF: High-Fidelity Virtualized Walkable Spaces. In *SIGGRAPH Asia 2023 Conference Papers*. 23.

[285] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. 2022. Point-NeRF: Point-based Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5428–5438.

[286] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. 2024. DMV3D: Denoising Multi-view Diffusion Using 3D Large Reconstruction Model. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[287] Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. 2022. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 597–614.

[288] Jiawei Yang, Marco Pavone, and Yue Wang. 2023. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[289] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong. 2022. Ps-nerf: Neural inverse rendering for multi-view photometric stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 266–284.

[290] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong. 2022. S^3-NeRF: Neural reflectance field from shading and shadow under a single viewpoint. In *Advances in Neural Information Processing Systems*, Vol. 35. 1568–1582.

[291] Mingyuan Yao, Yukang Huo, Yang Ran, Qingbin Tian, Ruifeng Wang, and Haihua Wang. 2024. Neural Radiance Field-based Visual Rendering: A Comprehensive Review. *arXiv preprint arXiv:2404.00714* (2024).

[292] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. 2020. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1790–1799.

[293] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems*, Vol. 34. 4805–4815.

[294] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems*, Vol. 33. 2492–2502.

[295] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Tsung-Yi Lin, Alberto Rodriguez, and Phillip Isola. 2022. Nerf-supervision: Learning dense object descriptors from neural radiance fields. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 6496–6503.

[296] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. 2021. iNeRF: Inverting Neural Radiance Fields for Pose Estimation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 1323–1330.

[297] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. 2023. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12–22.

[298] Youngho Yoon and Kuk-Jin Yoon. 2023. Cross-Guided Optimization of Radiance Fields with Multi-View Image Super-Resolution for High-Resolution Novel View Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12428–12438.

[299] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021. PlenOctrees for Real-time Rendering of Neural Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5732–5741.

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

[301] Chaohui Yu, Qiang Zhou, Jingliang Li, Zhe Zhang, Zhibin Wang, and Fan Wang. 2023. Points-to-3d: Bridging the gap between sparse points and shape-controllable text-to-3d generation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6841–6850.

[302] Zehao Yu, Anpei Chen, Bozidar Antic, Songyou Peng, Apratim Bhattacharyya, Michael Niemeyer, Siyu Tang, Torsten Sattler, and Andreas Geiger. 2022. SDFStudio: A Unified Framework for Surface Reconstruction. https://github.com/autonomousvision/sdfstudio

[303] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. 2024. MonoSDF: exploring monocular geometric cues for neural implicit surface reconstruction. In *Advances in Neural Information Processing Systems*.

[304] Ruyi Zha, Xuelian Cheng, Hongdong Li, Mehrtash Harandi, and Zongyuan Ge. 2023. EndoSurf: Neural Surface Reconstruction of Deformable Tissues with Stereo Endoscope Videos. In *International Conference on Medical Image*

Computing and Computer Assisted Intervention (MICCAI). 13–23.
[305] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric Xing. 2023. Multimodal image synthesis and editing: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
[306] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. 2024. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics* (2024).
[307] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. 2021. Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. *Advances in Neural Information Processing Systems* 34 (2021), 29835–29847.
[308] Jingyang Zhang, Yao Yao, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. 2022. Critical Regularizations for Neural Surface Reconstruction in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6270–6279.
[309] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. 2021. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5453–5462.
[310] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492* (2020).
[311] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 586–595.
[312] Wenyuan Zhang, Ruofan Xing, Yunfan Zeng, Yu Shen Liu, Kanle Shi, and Zhizhong Han. 2023. Fast Learning Radiance Fields by Shooting Much Fewer Rays. *IEEE Transactions on Image Processing* 32 (2023), 2703–2718.
[313] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. 2021. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics* 40, 6 (2021), 1–18.
[314] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. 2021. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 15838–15847.
[315] Allan Zhou, Moo Jin Kim, Lirui Wang, Pete Florence, and Chelsea Finn. 2023. Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 17907–17917.
[316] Kun Zhou, Wenbo Li, Yi Wang, Tao Hu, Nianjuan Jiang, Xiaoguang Han, and Jiangbo Lu. 2023. NeRFLiX: High-Quality Neural View Synthesis by Learning a Degradation-Driven Inter-viewpoint MiXer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[317] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. 2018. Stereo Magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics* 37 (2018).
[318] Xiaowei Zhou, Sida Peng, Zhen Xu, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, and Hujun Bao. 2024. Animatable implicit neural representations for creating realistic avatars from videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 6 (2024), 4147–4159.
[319] Zhizhuo Zhou and Shubham Tulsiani. 2023. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12588–12597.
[320] Chengxuan Zhu, Renjie Wan, and Boxin Shi. 2022. Neural transmitted radiance fields. In *Advances in Neural Information Processing Systems*, Vol. 35. 38994–39006.
[321] Fang Zhu, Shuai Guo, Li Song, Ke Xu, Jiayu Hu, et al. 2023. Deep review and analysis of recent nerfs. *APSIPA Transactions on Signal and Information Processing* 12, 1 (2023).
[322] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. 2022. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12786–12796.
[323] Jingyu Zhuang, Chen Wang, Liang Lin, Lingjie Liu, and Guanbin Li. 2023. Dreameditor: Text-driven 3d scene editing with neural fields. In *SIGGRAPH Asia 2023 Conference Papers*. 1–10.

Received 08 January 2025; revised 12 November 2025