International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| Volume 3, Issue 4, 115-127, 2022
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X.IJERET-V3I4P113

Original Article

3D Reconstruction from Monocular Videos Using Neural Radiance Fields (NeRF)

Sajud Hamza Elinjulliparambil
Pace University.

Abstract - *Monocular video-based 3D reconstruction has emerged as a fundamental yet challenging problem in computer vision, due to depth ambiguity, scale uncertainty, and limited viewpoint coverage. Traditional geometry-related approaches, which include Structure-from-Motion (SfM), Multi-View Stereo (MVS) and SLAM, are partial solutions and usually result in incomplete or noisy reconstruction. Neural Radiance Fields (NeRF) broke the previous paradigm of 3D generation by modelling the scene as a continuous volumetric generator, which takes 3D coordinates and viewing directions as inputs and neural colour and density as outputs to generate photorealistic novel-view images. This review follows the history of NeRF and its initial extensions to monocular video, such as sparse-view adaptations (PixelNeRF, DietNeRF, RegNeRF), dynamic and deformable scene modeling (D-NeRF, NSFF, NeRF-T), and optimization strategies, such as pose estimation, regularization, and efficiency. We address evaluation policies, datasets, and applications in the areas of AR/VR, robotics, cultural heritage, and digital content creation. Lastly, we provide a critical reflection on the limitations of NeRF and are able to identify future perspectives, such as improved priors in monocular input, faster inference, generalizable architectures, and lightweight models. The paper is a detailed overview of the methods that form the basis of neural-radiance-field-based monocular-video reconstruction and preconditions for further progress in that direction.*

Keywords - *Neural Radiance Fields, 3D Reconstruction, Neural Radiance Fields, Monocular Video, Implicit Neural Representations, Neural Radiance Fields, Pose Optimization.*

1. Introduction

The concept of constructing three-dimensional (3D) structure and appearance using monocular video has been a long time challenge in computer vision [1]. The fact that it is possible to deduce a complete scene geometry with minimal input being a sequence of images captured with a single moving camera is an opening to cheap, versatile, and ubiquitous generation of 3D content. Talent is essential in many areas, such as augmented and virtual reality (AR/VR), robotics, preservation of digital heritage, and visual effects. Monocular reconstruction in AR/VR allows an experience of the real world by immersing with a believable scene without the need to acquire costly multi-camera setups[2]. In robotics, it enables autonomous systems to perceive and maneuver through complicated settings with low-weight sensors [3]. In other applications, such as entertainment and visualization markets, the monocular video capture is a useful alternative to the conventional multi-view scanning systems.

Although this is important, 3D reconstruction using monocular input is not an easy task. The explicit depth information is not available in a single camera which makes the depth ambiguous, the scale indeterminate, and the uncertainty under low parallax motion [4]. Estimation of accurate geometry is even harder in cases where the camera path lacks a sufficient amount of baseline variation. Other challenging conditions that are commonly experienced with monocular capture include occlusions, dynamic objects, low light, and surfaces without any texture and which may decrease the quality of reconstruction[5].The most notable systems based on classical geometrical methods, including Structure-from-Motion (SfM), Multi-View Stereo (MVS), and many of the Simultaneous Localization and Mapping (SLAM) systems, are important building blocks to reconstruction[6]. SfM approximates camera positions and sparse 3D points whereas MVS tries to densify these point clouds into surfaces, which are more complete. The purpose of SLAM techniques is to coordinate the movements of the camera and to map the environment in real-time.[7] Nevertheless, each of them is overly dependent on solid feature recognition, sound photometric stability, and preferable camera movement. Practically, they tend to generate incomplete or noisy geometry in situations of low-textured scenes, reflective materials, repeating patterns or moderate changes of viewpoint typical of monocular video recording. In addition, conventional pipelines do not couple geometry estimation with appearance modeling which means that the reconstructions are not photorealistic, view-dependent, or have a smooth transition between lighting conditions[8][9].

Neural Radiance Fields (NeRF) has become an important conceptual advance on the way scenes are represented and reconstructed. Rather than representing the scene as a collection of discrete point clouds or meshes, NeRF represents a continuous volumetric representation of the scene, learned as a 3D volumetric predictor that takes 3D coordinates and viewing directions as input and outputs color and density [10]. This implicit representation, optimized-end-to-end on input image allows unbelievably high-quality novel view synthesis with naturalistic lighting and structural fine detail. The differentiable volumetric rendering formulation that NeRF uses is a combination of predicted sample values along camera rays, which the network is able to learn representations that have both geometry and appearance in a single formulation [11].

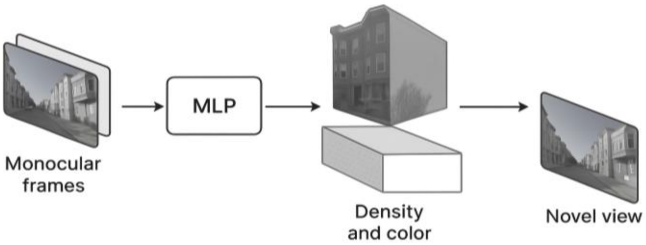

Fig 1: Conceptual Illustration of the Nerf Pipeline

Monocular frames serve as input, from which camera rays are cast into the scene. An MLP predicts density and color for each sampled point along a ray[12]. A volumetric rendering procedure then integrates these predictions to reconstruct the 3D scene and synthesize novel views.

Subsequent extensions were aimed at correcting some of the drawbacks of NeRF, such as high computational cost, dependence on dense multi-view data, and the problem of sparse or monocular input. Enhanced sampling strategies, regularization, pose optimization and representational efficiency expanded the range of applications of NeRF to monocular video. The present review is devoted to these pioneering contributions, and it is possible to trace the development of the ideas and approaches that led to the appearance of the first radiance-field-based reconstruction systems.

2. Background and Foundational Concepts

For this conceptual basis one needs to learn about radiance-field-based monocular 3D reconstruction. It starts with a discussion of classical methods of geometry that were created before neural implicit models and proceeds to mention early neural rendering formulations which directly inspired the development of NeRF-like methods. Lastly, it proposes the principles of volumetric rendering that are the mathematical foundation of radiance-field models.

2.1. Classical 3D Reconstruction from Monocular Videos

2.1.1. Structure-from-Motion (SfM)

SfM tries to obtain camera paths and 3D geometry using a set of 2D images by identifying and correlating keypoints (e.g., SIFT, ORB) across frames [13]. Epipolar constraints are used to estimate camera motions and triangulation generates sparse 3D points. Both bundling and positing solutions are refined in the process of bundle adjustment to come up with solutions that are globally consistent. Whereas SfM is mathematically beautiful and well-posed in textured scenes, it is susceptible to variations in lighting, blur, repetitive, textures and low-parallax motion typically found in monocular video because of its feature matching [14]. In addition, SfM usually makes sparse reconstructions that need further densification.

2.1.2. Multi-View Stereo (MVS)

MVS is an extension of SfM and results in dense depth or surface reconstructions of a scene based on multiple calibrated views of a scene. Classical pipelines consist of plane-sweep pipelines, patch-based pipelines, and volumetric MVS [15]. The techniques rely on proper calibration of the cameras and interview consistency in photometry. MVS is capable of producing dense geometry of high quality, but it is not able to operate effectively where the camera path does not offer much variation in view point (as is often the case with a hand-held monocular video). Failures usually occur where there are thin structures, specular surfaces and low-texture regions and the quality of reconstruction significantly diminishes at low baselines.

2.1.3. SLAM Frameworks

Simultaneous Localization and Mapping (SLAM) algorithmsFMLPAM, ORB-SLAM and LSD-SLAM concentrate on real-time performance, allowing them to be used in AR, robotics and mobile mapping. Visual SLAM systems are systems that track camera movement in an incremental manner, and keep sparse or semi-dense maps [16]. In spite of their practical effectiveness, SLAM systems focus on localization, but not high-fidelity geometry. They tend to be incomplete with their reconstructions, vulnerable to drift over long sequences and infrequently include elaborate appearance models. Therefore, though SLAM can be used in pose estimation, monocular sequence photorealistic reconstruction cannot be done solely by SLAM.

2.1.4. Depth Estimation of a single image.

Initial depth estimation systems based on deep learning train depth priors which predict plausible depth using one RGB frame, with either supervised or self-supervised training. This kind of priors come in handy in monocular configurations with a poor explicit geometric cue.

The depth of single-image however is never fully clear: there is no definition of scale, and fine details are usually over-smoothed on prediction. These priors are useful in providing information to guide reconstruction but not to substitute multi-view information on geometry.

Table 1: Comparison of Classical Monocular Reconstruction Techniques

Method	Core Principle	Strengths	Limitations	Typical Computational Cost
SfM	Keypoint matching → pose estimation → triangulation	Accurate global poses	Sparse output; fails on low-texture scenes	Moderate–high
MVS	Depth estimation across multiple views	Dense detailed surfaces	Needs accurate poses; sensitive to lighting	High
SLAM	Incremental tracking + mapping	Fast; real-time; robotics/AR	Drift; limited density; less photorealistic	Low–moderate
Single-image depth	Learned priors from one image	Fast; dense output	Ambiguous scale; limited accuracy	Low

2.2. Neural Rendering Foundations

This subsection introduces early neural rendering work that established the conceptual basis for implicit radiance fields and neural scene representations.

2.2.1. Neural Volumes

Neural Volumes express experiences in terms of voxel grids with each voxel holding a learned feature vector stored in it. A neural decoder is used to encode voxel features to colors, and differentiable compositing is used to generate new views [17]. Voxel-based methods have the disadvantage of cubic memory expansion, which prevents resolution or scalability to large or fine scenes. However, Neural Volumes showed that view-consistent synthesis could be obtained using differentiable volumetric rendering and train geometric structure directly using images.

2.2.2. Scene Representation Networks

Continuous implicit neural scene representations were introduced by SRNs, whereby an MLP takes the 3D coordinates as inputs to a feature or radiance-related quantity. These networks allowed resolution-independent representation of complicated geometry with ease[18]. The primary contribution of SRN was that scenes do not necessarily have to be discretized into voxel grids, a notion upon which radiance-field models were subsequently built.

2.2.3. DeepVoxels

DeepVoxels suggested a differentiable 3D feature grid which extends to camera views by learned basis functions. The method preserves viewpoint dependent appearance and geometry information by storing persistent volumetric features. Such an architecture combined classical voxel grids and neural implicit functions with view-synthesis and learned 3D features[19].

2.2.4. Fundamentals of Differentiable Rendering

Gradients can be used to propagate rendering losses to scene representations using differentiable rendering methods. First methods investigated differentiable renderer based on rasterization and volumetric methods. The seminal observation computing image synthesis in a differentiable form directly affected radiance-field models which are trained through gradient-based optimization solely on 2D supervision.

2.3. Volumetric Rendering and Implicit Neural Representations

2.3.1. Continuous Radiance Fields

The differentiable rendering methods enable the gradient of rendering losses to flow back to the scene representations. First methods investigated differentiable renderer based on rasterization and volumetric methods [20]. The seminal observation computing image synthesis in a differentiable form directly affected radiance-field models which are trained through gradient-based optimization solely on 2D supervision.

2.3.2. Ray Sampling and Ray Marching

The rendering process is accomplished by projecting rays of each pixel in the camera into the scene and sampling a number of 3D points along the ray. Each sample is predicted by a neural network in terms of density and color[21]. The hierarchical sampling strategies narrow down the sampling on the regions of high density, enhancing accuracy and minimizing the cost of computation.

2.3.3. Density and Color Prediction

The value of the predicted density is the likelihood of a point to contribute radiance along the ray. Surfaces are usually associated with high density. The prediction of colors can be based upon the direction of viewing in which specular or reflective phenomena are captured. This general direction of construction enables the model to capture fine geometry and photometric variations without explicit shape representations.

2.3.4. Differentiable volumetric integration

Last pixel color is calculated using accumulated transmittance and rays using accumulated transmittance and accumulated rays. The whole process of integration is differentiable, implying that the changes in subtle differences between images can influence the alteration of the radiance-field parameters. The differentiable volume-rendering equation is the mathematical base of the radiance-field-based techniques and allows them to learn detailed representations of scenes directly on the basis of image observations.

3. Original NeRF: Formulation and Pipeline

The original NeRF has introduced a new view synthesis paradigm based on neural network and differentiable rendering to learn a continuous volumetric representation of the scene. It is necessary to understand this baseline and evaluate its monocular-video adaptations.

3.1. NeRF Architecture

3.1.1. Multi-Layer Perceptron (MLP) as a Continuous Scene Function

NeRF represents a fixed 3D scene by a continuous function Ftheta (x,d) F.theta(x, d) F.theta(x, d) that takes a 3D point as an input xxx and a viewing angle d as an input and returns both a color and a volume density. This operation is implemented by a deep MLP-based apparatus with fully connected layers that allow the storage of fine-grained geometric and photometric data in network weights[22]. The implicit formulation is very high-spatial-resolution with no huge memory overhead as opposed to discrete voxel grids or point clouds.

3.1.2. Positional Encoding for High-Frequency Detail

A key innovation of NeRF is the positional encoding scheme, that is, the input coordinates are first mapped to a sequence of sinusoidal functions and then input to the MLP. This increases the range of modeling of high-frequency variations like edges, textures, and sharp geometry by the network. In the absence of positional encoding, normal MLPs will overfit smooth reconstructions because they have small spectral representation capabilities.

3.1.3. Prediction of RGB Color and Volume Density

The network outputs two key quantities:

- Density (σ\sigmaσ): it encodes geometry and describes the likelihood of light being absorbed or scattered at the point, and therefore has a value at each point.
- Radiance (ccc) : this is the RGB color emitted in the direction of viewing.

This is where the foundation of the NeRF differentiable rendering is based and the experience of the learned implicit field is connected to the camera pixels [23].

3.2. Volumetric Rendering

3.2.1. Ray Marching Through the Radiance Field

On every pixel of the image, NeRF launches a camera ray and samples a collection of points along the ray. The color and density of the MLP are sampled at every point and then incorporated along the ray based on the classical volume rendering principles. The resultant integration will give the desired pixel color with the existing scene representation.

3.2.2. Hierarchical Sampling Strategy

NeRF uses a hierarchical architecture to obtain better rendering results with reduced efficiency and accuracy: a coarse network is used to estimate the coarse density distribution, and then a fine network is used to concentrate the high-resolution samples on high-density regions. This coarse to fine technique not only speeds up the training process but also increases the visual fidelity.

3.2.3. Coarse-to-Fine Network Pairing

The coarse and fine MLPs have identical functional structure except that they learn scene details at different sampling granularities. Their results are summation, to create the final render, which allows the creation of finer results without the expensive costs of the sampling.

3.3. Strengths and Limitations

3.3.1. Strengths: Photorealistic Novel-View Synthesis

NeRF has demonstrated record-breaking realism in generating new perspectives that surpass previous neural rendering techniques and the classical 3D reconstruction approaches through its ability to reproduce the whole illumination, effects depending on the viewpoint, and the complex geometry.

3.3.2. Limitations: Computer and Practical Constraints

NeRF is computationally intensive, which needs massive training time and dense multi-view supervision. It is implicitly represented and thus geometry and appearance are closely related, thus not easily generalized to new scenes.

3.3.3. Video Reconstruction Implications on Monocular Video

Since the original NeRF operates on multi-view data with correct camera pose, it cannot be used directly on monocular-only input, which has scale ambiguity and inadequate parallax. These limitations inspired a large number of initial extensions to enable NeRF to be used with sparse or monocular measurements.

4. Early Extensions of NeRF Relevant to Monocular Video

The first version of NeRF showed excellent quality novel-view synthesis but was extremely dependent on dense multi-view images (as well as large-scale computing power). Researchers came up with a set of changes and improvements in order to generalize it to monocular video and sparse-view inputs. The first extensions were aimed at enhancing the data efficiency, the integration of priors, and speed of inference, establishing the foundation of the monocular reconstruction using neural radiance fields.

4.1. NeRF for Sparse Views

4.1.1. PixelNeRF

PixelNeRF trains the radiance-field network conditioned on image features of training images. Using 2D CNN features, PixelNeRF is also able to generalize on novel viewpoints using less training images, which is especially helpful in sparse-view or monocular systems. The network employs feature pooling and projecting rays to direct the volumetric reconstruction to enhance performance in a situation where the multi-view coverage is restricted.[24]

4.1.2. DietNeRF

DietNeRF solves the sparse-ray supervision problem by introducing regularization terms which provide perceptual consistency and semantic priors. This motivates the network to make plausible geometry predictions in the even cases that have scarce image data. DietNeRF prevents the tendency to overfit to sparse views by punishing deviations of the statistics of known image images and feature distributions [25].

4.1.3. RegNeRF

RegNeRF explicitly regularizes to enhance stability in the sparse-view configurations. Some of these techniques are depth smoothness constraints and consistency between neighboring rays. These adjustments minimize artifacts present in the network like floating geometry or noise that are often present when the network has inadequate viewpoints.

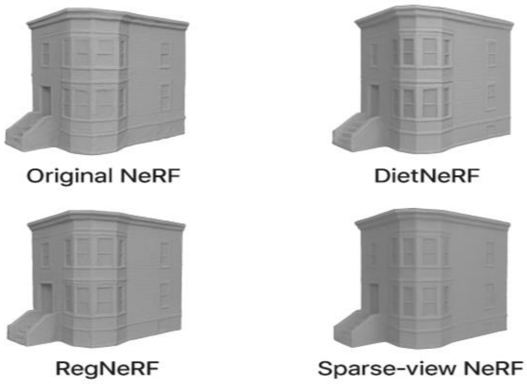

Fig 2: Comparison of sparse-view NeRF reconstructions versus original NeRF

Original NeRF exhibits degraded quality when multi-view data is limited, while regularized variants (DietNeRF, RegNeRF) show improved geometry and fewer artifacts.

4.2. NeRF with Priors for Monocular Video

Incorporating priors is critical for enabling monocular reconstruction, where geometric information is ambiguous [26].

4.2.1. Depth-Prior-Assisted NeRF Variants

Early models used geometrical constraints by taking the first step of monocular depth estimates or sparse depth maps and training NeRF. Depth priors provide predictions of densities using a ray, which reduces scale ambiguity and enhances consistency of the scene.

4.2.2. Priors of Learning Shape and Appearance

Other techniques make use of massive datasets to acquire generic shape and texture priors so that they can better generalize to unknown scenes. Priors conditioned networks can recover plausible geometry using very little monocular input to compensate for missing multi-view information. Research examined the use of one image to initialize NeRF representations. Some of the techniques used are the use of predicted depth, semantic features or pre-trained encoders to bootstrap volumetric representation. This strategy will cut down the amount of data and enable monocular sequences to gradually increase the refinement of the scene.

4.3. Fast Reconstruction Extensions

Reducing computational cost is essential for practical monocular-video reconstruction, particularly for real-time or interactive applications [27].

4.3.1. FastNeRF

Ray batching and network factorization of FastNeRF optimizations will also enable the rendering and training process to be faster. The changes ease the computational bottleneck and therefore, reconstruction of sequential frames becomes simpler.

4.3.2. KiloNeRF

KiloNeRF splits the scene into smaller sub-networks where each sub-network occupies a section of the volume. These sub-networks can be compared in parallel to give the power to conclude the results more quickly, without reducing the quality of reconstruction, and comes in handy in the examination of longer sequences of monocular video.

4.3.3. PlenOctrees

PlenOctrees is a volumetric representation of precomputer hierarchy octree representation. This is made possible by the caching of the network outputs to a multi-resolution tree structure, which enables rendering to be rendered with much greater speed, and synthesis almost in real-time at high fidelity. It is the speed-ups that make NeRF the appropriate choice when applied to monocular video, where a series of successive frames can be inputted to it, and can be gradually fused into a coherent 3D world. Combined with these original extensions of NeRF, two bottlenecks of NeRF became addressed, namely computational efficiency and data sparsity. They made the first application of learned priors, regularization and faster representations so that they could learn monocular video reconstruction without necessarily needing to use dense multi-view data in addition to producing photorealistic quality.

5. Dynamic Scene and Non-Rigid NeRF Extensions

The original NeRF architecture represents fixed scenes thus it is inappropriate to be used to reconstruct dynamic or deformable scenes which are prevalent in monocular video. This was overcome by early extensions, which added temporal and motion-aware representations to the radiance-field representation. Such approaches allow the reconstruction of moving objects and non-rigid scenes with the high-fidelity view synthesis [28].

5.1. Deformable NeRF (D-NeRF)

D-NeRF proposes a deformation-sensitive radiance field, which is a model with the movement of scene points over time. The coordinates of the 3D space are time-conditioned to deformations that provide a mapping between the points in a canonical reference frame to the location of the point at a given timestamp. Color and density in the network are predicted according to the spatial location and the state of time, and with the help of it, dynamic geometry can be reconstructed. It can be used to synthesize novel-views of non-rigid objects, e.g. moving humans or morphing shapes, and forms a basis in the processing of monocular video sequences in which objects might move with the scene in relation to a fixed camera.

5.2. Neural Scene Flow Fields

This subsection highlights methods that jointly model motion and geometry using scene flow.NSFF builds on top of NeRF by integrating 3D scene flow estimation, so that the network can simultaneously learn both per-point motion and radiance. Not only color and density but also a 3D flow vector, that is, displacement of points over time, is predicted by each ray sample. NSFF is able to synthesize temporally coherent frames by combining volumetric rendering with motion-sensitive features, and can learn the complicated dynamics of rigid motion, deformation or articulated motion. This bi-modeling of geometry and motion is especially useful in monocular video, when multi-view information on depth is missing, but replaced at least partially by time.

5.3. Dynamic NeRF-Based Methods for Monocular Video

5.3.1. NeRF-T

NeRF-T adds to NeRF the concept of temporal conditioning in order to process sequential frames of one camera. It is a time-combining method of dynamic scene representations that allows tracking and reconstruction of moving objects continuously. NeRF-T smooth and temporally consistent novel-view synthesis is enabled by conditioning the radiance field on the temporal input [29].

5.3.2. Single-Camera View Based Dynamic View Synthesis.

A number of studies investigated dynamic video reconstruction of monocular videos using single frame depth priors, optical flow or learned motion embeddings. The techniques enable the network to learn motion patterns and eliminate ambiguity due to the lack of multiple viewpoints. The dynamic radiance fields generate consistent geometry and naturalism of appearance change between frames.

5.3.3. Challenges in Monocular Dynamic Reconstruction

Despite these advancements, reconstructing dynamic scenes from monocular video remains challenging:

- Motion blur degrades feature consistency and density prediction.
- Occlusions create missing observations for parts of moving objects.
- Lack of multi-view redundancy limits disambiguation of geometry and motion.

The above constraints indicate that early dynamic NeRF methods can generate moving scenes in an impressive manner, but the quality of results still depends on attentive time modeling, regularization, and prior information about object dynamics. In general, deformable and motion-aware versions of NeRF represent significant advances in the NeRF towards monocular video reconstructions of dynamic scenes. They demonstrate that radiance-field models can be adapted to support temporal conditioning and scene-flow reasoning can be implemented and hence can be applied to generalize to non-static scenes in order to achieve more generalized applications in robotics, AR/VR, and video-based 3D content capture.

6. Optimization and Training Strategies

Neural radiance field training on monocular video sequences presents special difficulties as it has limited viewpoints, scale ambiguity, motion and is very expensive to compute. This part examines optimization methods, regularization methods and efficiency methods that have been derived to facilitate strong monocular reconstruction.

6.1. Regularization Techniques

Regularization is critical for guiding NeRF training on sparse or monocular data to avoid overfitting and improve reconstruction fidelity.

6.1.1. Sparse-Depth and Semantic Consistency

Asparse depth measurements or semantic priors are also provided to give a certain constraint in the volumetric density prediction. Varieties of NeRFs give better reconstructions by imposing consistency between predicted geometry and known depths (which are sparse, either SfM, SLAM or single-image depth networks). Semantic regularity encourages geometrical smoothness between areas marked in much the same way and appearance stabilizes additional learning in underconstrained settings [30].

6.1.2. Pose Refinement (iNeRF)

iNeRF suggested the use of pose refinement by minimizing the difference between the images that are generated and the original images. In order to mix camera pose and radiance-field parameters, the technique can be used to process videos taken with a monocular, using coarse or erroneous initial poses in order to form high quality 3D reconstructions. Pose refinement eradicates misalignment artifacts and improves consistency of sequence time.

6.1.3. Depth Smoothness and Prepared Enforcement.

Smoothing predicted depth using smoothness, or learned, priors smooths unrealistic geometric variation across surfaces. The use of monocular sequences and networks trained with assistance of sentencing big gradients or variations of expected surface distributions can result in more consistent reconstructions regardless of poor multi-view coverage.

6.1.4. Camera Pose Estimation for Monocular Video

Accurate camera poses are essential for NeRF training; this subsection addresses strategies for monocular sequences.

6.1.5. COLMAP Limitations

Multi-view results: Camera poses Multi-view images COLMAP, a popular SfM pipeline, can give multi-view scenes, however monocular videos are usually affected by either low baseline, motion blur or no texture areas. In these scenarios, SfM can only fail or generate drifted poses, and some alternative or improved pose estimation techniques must be used.

6.1.6. Self-Calibrated NeRF

Self-calibrated variants of NeRF do the joint optimization of unknown camera pose and scene parameters. With the help of differentiable rendering, such approaches will optimize pose estimates by training the network; that is, they do not rely on external SfM outputs. This comes in handy especially in monocular video sequences where the external pose estimation is unreliable.

6.1.7. Pose and Radiance Fields Joint Optimization.

Pose-refinement Pose-refinement Joint optimization frameworks combine radiance-field training with pose refinement. Photometric reconstruction, geometric priors, and temporal smoothness are the common types of loss functions. This combined method lowers error cumulative due to misaligned frames, and enhances the stability of monocular reconstructions.

6.2. Memory and Efficiency Solutions

Efficiency techniques address the high computational and memory demands of NeRFs, especially when applied to long monocular video sequences.

6.2.1. Multi-Resolution Hashing

Before Instant-NGP, the idea of multi-resolution hashing allowed the representation of features in a compact and multi-scale way. Networks save scene features in hash-based grids, and query the network in areas of interest by each ray, reducing memory resources, and maintaining detail.

6.2.2. Acceleration-Grids (PlenOctrees)

PlenOctrees use hierarchical octree grids to convert pre-trained NeRFs that enable lookups and ray integration in a volume fashion. This method offers considerable rendering and training speed-ups, and is applicable to monocular-video reconstruction in which one frame should be effectively processed at a time. All these optimization strategies overcome three main problems of monocular NeRF training, namely, underconstrained geometry (through regularization), untrustworthy poses (through joint or self-calibrated optimization), and computational cost (through memory-efficient and accelerated representations). With the combination of these approaches, it became possible to use works that allowed efficient and high-fidelity 3D reconstruction using monocular sequences.

7. Evaluation Metrics, Datasets, and Benchmarks

Testing using monocular-video 3D reconstruction needs the devised techniques to be assessed with standardized datasets, geometry and appearance metrics and benchmarks. This section is a summary of the most frequently used evaluation protocols and datasets.

7.1. Evaluation Metrics

Selecting appropriate metrics ensures quantitative comparison of reconstruction quality.

7.2. Geometric Metrics

- Chamfer Distance (CD): Measures This averages closest-point distance between predicted and ground-truth point clouds. The lower the values the better geometry it represents[31].
- Earth Mover Distance (EMD): Maps similarity globally, which is a geometrical distance that minimizes the cost to align the predicted and ground-truth points[32].
- Depth Error Metrics: Mean absolute error (MAE) or root-mean-square error (RMSE) of predicted and actual depths.

7.3. Appearance Metrics

- Peak Signal-to-Noise Ratio (PSNR): The photometric similarity of rendered and ground-truth images; high PSNR means that view synthesis is good[33].
- Structural Similarity Index (SSIM): Perceptual similarity; resistant to luminance changes.
- Learned Perceptual Image Patch Similarity (LPIPS): Learned high-level perceptual differences between synthesized and reference images.

7.4. Datasets for Monocular and Sparse-View Evaluation

Datasets provided diverse benchmarks for static, dynamic, and single-view scenarios.

7.5. Synthetic Datasets

- Blender Synthetic Scenes: NeRF paper original; they are images of inanimate objects with ground-truth depth and camera pose[34].
- ShapeNet-based Variants: The Variants also provide rendered synthetic objects to evaluate sparse-views.

7.6. Real-World Datasets

- LLFF (Local Light Field Fusion): Forward-facing real-world scenes: these are collected with handheld cameras; these can be used to synthesize novel views with limited baseline.
- Tanks & Temples : Multi-view scans have high quality and can be used on real scenes; it is possible to judge the reconstruction of geometry.
- Dynamic Scene Datasets: Smaller datasets Small sequences of deformable objects or human motion recorded with multi-view rigs are often used as that.

7.7. Benchmarks

Works on NeRF and extensions had found results on these datasets in a manner to compare geometry and novel-view synthesis performance [35].

- Original NeRF : Has been shown to be state-of-the-art with respect to PSNR, SSIM and LPIPS on both synthetic and real forward-facing scenes.
- PixelNeRF and DietNeRF: Compared to baseline NeRF on LLFF, sparse-view reconstruction tasks were benchmarked with regularization and feature conditioning which yielded better quality in low-view settings.
- Dynamic NeRF Extensions (D-NeRF, NSFF, NeRF-T): D-NeRF and NeRF-T were assessed in terms of temporal consistency and geometry prediction on synthetic deformable sequences and small dynamic real-world captures.

Table 2: Evaluation Metrics and Datasets

Method	**Dataset**	**Views**	**PSNR ↑**	**SSIM ↑**	**Chamfer ↓**	**Notes**
NeRF (2020)	Blender	Dense	32–34	0.95	0.004	Static, synthetic
NeRF (2020)	LLFF	Moderate	27–29	0.92	N/A	Real forward-facing scenes
PixelNeRF (2021)	LLFF	Sparse	25–28	0.91	N/A	Sparse-view conditioning
DietNeRF (2021)	LLFF	Sparse	26–29	0.92	N/A	Sparse-view + regularization
D-NeRF (2020)	Synthetic dynamic	Dense	31–33	0.94	0.005	Deformable scenes
NSFF (2020–21)	Synthetic dynamic	Dense	30–32	0.93	0.006	Motion-aware synthesis

This part underlines the assessment of NeRF and its monocular or sparse-view variants based on the synthesis and real dataset combinations, traditional geometric measures (Chamfer, EMD) and image-synthesis ones (PSNR, SSIM, LPIPS). The benchmarks that were made at this time laid the groundwork to future monocular-video NeRF research.

8. Applications

Neural radiance fields with their initial extensions had already been proving useful in practice in a variety of applications. These applications were in the early phases but they showed how NeRF could revolutionize work in immersive visualization, digital preservation, robotics, and visual effects. This part outlines the main fields where NeRF and its methods were utilized.

8.1. AR/VR and View Synthesis

NeRF enabled high-quality novel-view synthesis, a key requirement for immersive applications.Photorealistic simulation of spaces and environments under arbitrary viewpoints is critical in augmented and virtual reality. The interpolative nature of NeRF made it possible to generate full-fledged virtual worlds with little or no geometry modeling using a small number of images. There was also evidence from forward-facing datasets like LLFF that pre-trained NeRFs could be used to render smooth and consistent camera movements, which is why they could be used in lightweight AR/VR applications. Slim-view versions (PixelNeRF, DietNeRF) also decreased the amount of data required, and fast scene acquisition became possible for consumer-level VR applications. Cultural heritage NeRF projects were used to recreate statues, sculptures, and small architectural sites given a small set of photographs. Radiance-field-based methods captured geometry and appearance and therefore provided detailed digital representations that could be viewed through an interactive display or stored as archival items. Continuous representation helped prevent artifacts of discretization found in more conventional photogrammetry, giving it visually correct representations that could be used in preservation and in exhibition.

8.2. Robotics and Perception

NeRF was used in robotic perception to aid in scene comprehension and robot navigation.Although original formulations of NeRF were computationally expensive, early experiments showed learned radiance fields to be useful in enhanced 3D perception in robotics. Volumetric representations of sparse images have the potential to improve obstacle mapping, environment modeling and planning by view. Monocular robots or drones with extensions that used depth priors and pose-refinement to infer both the geometry and appearance of the scene could be made to run on single-camera sequences, possibly leading to less sensor demands in navigation and inspection tasks.

8.3. Digital Content Creation / Visual Effects (VFX)

NeRF enabled novel workflows for CGI and post-productionThe strict power to produce lifelike novel perspectives of the taken images offered new visual effects and content generation resources. Applications ranged from recreating real life scenery to be used in a film to create an integration with computer generated content, to simulating camera motion during post production, to creating reference art to be used in animation and compositing. The artists were also able to render a scene with minimal photography and generate elastic, view compatible images without the heavy manualization of models and simplifying the production chain.

NeRF and variants of sparse-view and motion-aware NeRF had already proven that it can be versatile in many fields. Key contributions included:

- Realistic novel-view synthesis for AR/VR experiences.
- Non-invasive digitization for cultural heritage preservation.
- Enhanced 3D perception and scene understanding for robotics.
- Efficient digital content generation and VFX workflows.

Although limited by the cost of computation and data needs, these initial applications demonstrated that neural radiance fields have the potential to help close the gap between computer vision and graphics and real world practical tasks.

9. Limitations and Anticipated Future Directions for Monocular Video NeRF

While neural radiance fields demonstrated remarkable capabilities for high-fidelity view synthesis and 3D reconstruction, early works faced significant challenges, particularly when applied to monocular video sequences. Simultaneously, literature anticipated research directions aimed at addressing these limitations. This section provides a critical review of the constraints observed in early NeRF studies and highlights predicted avenues for improvement.

9.1. Limitations of NeRF Approaches

This subsection summarizes major challenges faced by NeRF models trained on monocular or sparse-view video sequences.

9.1.1. Heavy Training Time

NeRF needs millions of ray evaluations and deep MLP evaluations per scene. Even with small-scale datasets, hours to days can be spent on training, and it is therefore restricted to implementation with long video-sequences or real-time applications.

9.1.2. Difficulty in Generalization

Original NeRF models are environment-specific, where training is needed every time the new environment is introduced. This limitation was partly solved with early sparse-view and prior-based extensions, although the problem of generalization between scenes was still very poor. Networks tend to fit images that they have seen, and therefore cannot transfer to unseen perspectives or novel monocular video sequences.

9.1.3. Relatively Limited Real-Time Capability.

Trained NeRFs were also computationally expensive to render novel views because of a dense ray sampling and MLP evaluation. Monocular video applications could achieve near real-time performance with coarse-to-fine sampling and early acceleration techniques (e.g. PlenOctrees), but not with other techniques.

9.1.4. Pose Dependence

NeRF presupposes the known camera poses in order to reconstruct the volumetric representation in an accurate way. Monocular video sequences can also have noisy or floating poses particularly on hand held capture. Although pose refinement techniques (iNeRF, self-calibrated NeRF) reduced this problem, the use of correct poses was an inherent limitation.

9.1.5. Problems in Dynamic Scenes.

NeRF original assumption of statistic scenes restricts its applicability to moving or non-rigid objects. NSFF, deformable NeRFs and temporal conditioning techniques enhanced performance, however, monocular display increases ambiguity in motion, occlusion and depth estimation. Also reconstruction quality is complicated by the motion blur and the absence of multi-view redundancy.

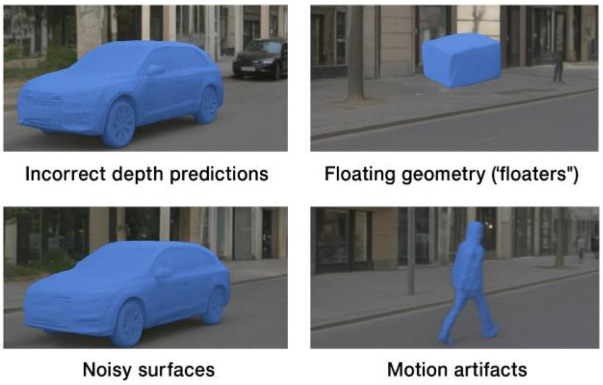

Fig 3: Common Failure Cases In Nerf: Incorrect Depth Predictions, Floating Geometry ("Floaters"), Noisy Surfaces, and Motion Artifacts When Applied To Monocular Video Sequences

A number of directions to address the limitations of NeRF and make it feasible to conduct practical monocular-video reconstruction were suggested in pre-2022 research articles. It was expected that incorporation of geometric, depth and semantic priors would alleviate the use of dense multi view information. These techniques were monocular depth estimation, large dataset shape priors and semantic consistency constraints. The other significant concern was to accelerate network assessment. Various methods were proposed to minimize the computational cost and allow longer sequences to be processed by hierarchical ray sampling, octree-based caching (PlenOctrees), network factorization and multi-resolution feature grids. It was also noted that researchers were in need of networks that can memorize new scenes without undergoing complete retraining. Monocular sequence few-shot or zero-shot generalization Feature-conditioned or meta-learning methods, including PixelNeRF, were suggested. Temporal consistency improvement, motion through occlusions and a sound treatment of non-rigid motion were identified to be key attributes of dynamic monocular sequences. It was early predicted that motion priors and time smoothness constraints would be added to stabilize the reconstruction. Lastly, the small network architecture was proposed to minimize memory and latency to be able to run on resource-bound devices. Suggested solutions were network pruning, quantization and modular MLP architectures.

10. Conclusion

Neural Radiance Fields (NeRF) has radically changed the future of 3D reconstruction by providing a continuous and differentiable volumetric representation that integrates both geometry and appearance modeling. NeRF and its initial variants showed impressive performance with respect to producing photorealistic new images, reconstructions of both static and dynamic scenes, and the use of small amounts of multi view data. The advances gave an appropriate avenue to a monocular-video reconstruction which previously was limited by depth ambiguity, scale uncertainty, and the lack of viewpoint coverage. The classical approaches like Structure-from-Motion (SfM), Multi-View Stereo (MVS) and SLAM provided the foundation with the capability to provide pose estimation and sparse-to-dense geometry reconstruction, but failed to work in cases of monocularity, low texture, occlusions and dynamic motion. NeRF has addressed most of these shortcomings by learning implicit representations of a scene, and early sparse-view and monocular extensions (PixelNeRF, DietNeRF, RegNeRF) extended the scope of volumetric neural rendering. D-NeRF and NSFF as well as NeRF-T are dynamic scene modeling techniques, which allow reconstruction of moving or deformable objects, a key issue in monocular video capture.

Regardless of these developments, the NeRF research manifested its major limitations. The use of high computational cost, intensive training time and requirement to employ precise camera pose limits real-time and general-purpose deployment. Monocular input increased scale ambiguity and depth uncertainty and dynamic and non-rigid scene reconstruction was still vulnerable to motion blur and occlusion. These limitations were pointed out in early evaluation protocols, where synthetic and real-world datasets were evaluated based on geometric and photometric metrics to measure reconstruction fidelity and novel-view synthesis quality. Literature also was providing expected future pathways to fight these challenges. The combination of geometric, depth and semantic priors would help eliminate dependence on dense multi-view information, and hierarchical sampling, octree-based caching and multi-resolution feature grids would help to provide quicker inference. Zero- or few-shot adaptable networks were anticipated to enhance the performance of different scenes, which are generalizable. It was identified that temporal consistency, motion priors and lightweight architectures are important in the strong dynamic scene reconstruction and practical application to resource-constrained devices.

In conclusion, the original publications on the NeRF made both the promise and the limitations of the neural radiance fields in reconstructing monocular-video known. They also served as a guide on how future studies should be conducted, pointing at the lack of efficiency, generalization, dynamic modeling, and scalability. This literature contributed to both expanding our insight into the neural volumetric depiction of a scene and showed viable directions on how the technology might be utilized in AR/VR, robotics, digital heritage, and visual effects. Consequently, the initial works continue to be critical in the development of the course of research in 3D reconstruction and neural representations that continue to influence future innovations and application in real-life applications.

References

[1] T. Georgiou, A. et al., "A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision," *Int. J. Multimedia Inf. Retr.*, vol. 9, no. 3, pp. 135–170, 2020.

[2] A. Bapat, *Towards High-Frequency Tracking and Fast Edge-Aware Optimization*, Ph.D. dissertation, Univ. North Carolina at Chapel Hill, 2019.

[3] M. B. Alatise and G. P. Hancke, "A review on challenges of autonomous mobile robot and sensor fusion methods," *IEEE Access*, vol. 8, pp. 39830–39846, 2020.

[4] Y. Lu, et al., "A survey of motion-parallax-based 3-D reconstruction algorithms," *IEEE Trans. Syst., Man, Cybern., Part C (Appl. Rev.)*, vol. 34, no. 4, pp. 532–548, 2004.

[5] M. Zollhöfer, et al., "State of the art on monocular 3D face reconstruction, tracking, and applications," *Comput. Graph. Forum*, vol. 37, no. 2, 2018.

[6] M. Kholil, I. Ismanto, and M. N. Fu'Ad, "3D reconstruction using structure from motion (SFM) algorithm and multi view stereo (MVS) based on computer vision," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1073, no. 1, 2021.

[7] D. Maier, A. Hornung, and M. Bennewitz, "Real-time navigation in 3D environments based on depth camera data," in *Proc. 12th IEEE-RAS Int. Conf. Humanoid Robots (Humanoids)*, 2012.

[8] H. Hofer, *Real-time visualization pipeline for dynamic point cloud data*, Ph.D. dissertation, Wien, 2018.

[9] G. Pintore, et al., "State-of-the-art in automatic 3D reconstruction of structured indoor environments," *Comput. Graph. Forum*, vol. 39, no. 2, 2020.

[10] A. R. Kosiorek, et al., "NeRF-VAE: A geometry aware 3D scene generative model," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2021.

[11] B. Mildenhall, et al., "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021

[12] .A. Flint, D. Murray, and I. Reid, "Manhattan scene understanding using monocular, stereo, and 3D features," in *Proc. Int. Conf. Comput. Vis.*, 2011.

[13] C. Russell, R. Yu, and L. Agapito, "Video pop-up: Monocular 3D reconstruction of dynamic scenes," in *Eur. Conf. Comput. Vis.*, Cham: Springer, 2014.

[14] O. Özyeşil, et al., "A survey of structure from motion*," *Acta Numerica*, vol. 26, pp. 305–364, 2017.

[15] G. Vogiatzis and C. Hernández, "Video-based, real-time multi-view stereo," *Image Vis. Comput.*, vol. 29, no. 7, pp. 434–441, 2011.

[16] S. Sumikura, M. Shibuya, and K. Sakurada, "OpenVSLAM: A versatile visual SLAM framework," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019.

[17] L. Yariv, et al., "Volume rendering of neural implicit surfaces," in *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 4805–4815, 2021.

[18] V. Sitzmann, M. Zollhöfer, and G. Wetzstein, "Scene representation networks: Continuous 3D-structure-aware neural scene representations," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[19] V. Sitzmann, et al., "DeepVoxels: Learning persistent 3D feature embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019.

[20] H. Kato, et al., "Differentiable rendering: A survey," *arXiv preprint arXiv:2006.12057*, 2020.

[21] M. Kettunen, et al., "An unbiased ray-marching transmittance estimator," *arXiv preprint arXiv:2102.10294*, 2021.

[22] M. E. Mirici, et al., "Land use/cover change modelling in a Mediterranean rural landscape using multi-layer perceptron and Markov chain (MLP-MC)," *Appl. Ecol. Environ. Res.*, vol. 16, no. 1, 2018.

[23] T. Bardak and S. Bardak, "Prediction of wood density by using red-green-blue (RGB) color and fuzzy logic techniques," *Politeknik Dergisi*, vol. 20, no. 4, pp. 979–984, 2017.

[24] A. Yu, et al., "pixelNeRF: Neural radiance fields from one or few images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021.

[25] A. Jain, M. Tancik, and P. Abbeel, "Putting NeRF on a diet: Semantically consistent few-shot view synthesis," *Supplementary Materials*, 2021.

[26] C. Gao, et al., "Dynamic view synthesis from dynamic monocular video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021.

[27] J. I. Agulleiro and J.-J. Fernandez, "Fast tomographic reconstruction on multicore computers," *Bioinformatics*, vol. 27, no. 4, pp. 582–583, 2011.

[28] E. Tretschk, et al., "Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021.

[29] J. Chen, et al., "Animatable neural radiance fields from monocular RGB videos," *arXiv preprint arXiv:2106.13629*, 2021.

[30] M. Jaritz, et al., "Sparse and dense data with CNNs: Depth completion and semantic segmentation," in *2018 Int. Conf. 3D Vision (3DV)*, IEEE, 2018.

[31] T. Wu, et al., "Density-aware Chamfer distance as a comprehensive metric for point cloud completion," *arXiv preprint arXiv:2111.12702*, 2021.

[32] D. Applegate, et al., "Unsupervised clustering of multidimensional distributions using earth mover distance," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2011.

[33] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?," in *4th Int. Workshop Quality Multimedia Exp.*, 2012.

[34] S. Basak, et al., "Methodology for building synthetic datasets with virtual humans," in *2020 31st Irish Signals and Systems Conf. (ISSC)*, IEEE, 2020.

[35] E. Tretschk, et al., "Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021.