Key Techniques and Challenges in NeRF-Based Dynamic 3D Reconstruction

LU Ping[1,2], FENG Daquan[3], SHI Wenzhe[1,2], LI Wan[3], LIN Jiaxin[3]

(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. Beijing XingYun Digital Technology Co., Ltd., Beijing 100176, China;
3. Shenzhen University, Shenzhen 518060, China)

DOI: 10.12142/ZTECOM.202503008

https://kns.cnki.net/kcms/detail/34.1294.TN.20250721.1711.002.html, published online July 22, 2025

Manuscript received: 2024-01-12

Abstract: This paper explores the key techniques and challenges in dynamic scene reconstruction with neural radiance fields (NeRF). As an emerging computer vision method, the NeRF has wide application potential, especially in excelling at 3D reconstruction. We first introduce the basic principles and working mechanisms of NeRFs, followed by an in-depth discussion of the technical challenges faced by 3D reconstruction in dynamic scenes, including problems in perspective and illumination changes of moving objects, recognition and modeling of dynamic objects, real-time requirements, data acquisition and calibration, motion estimation, and evaluation mechanisms. We also summarize current state-of-the-art approaches to address these challenges, as well as future research trends. The goal is to provide researchers with an in-depth understanding of the application of NeRFs in dynamic scene reconstruction, as well as insights into the key issues faced and future directions.

Keywords: neural radiance fields; 3D computer vision; dynamic scene reconstruction

Citation (Format 1): LU P, FENG D Q, SHI W Z, et al. Key techniques and challenges in NeRF-based dynamic 3D reconstruction [J]. ZTE Communications, 2025, 23(3): 71 - 80. DOI: 10.12142/ZTECOM.202503008

Citation (Format 2): P. Lu, D. Q. Feng, W. Z. Shi, et al., "Key techniques and challenges in NeRF-based dynamic 3D reconstruction," *ZTE Communications*, vol. 23, no. 3, pp. 71 - 80, Sept. 2025. doi: 10.12142/ZTECOM.202503008.

1 Introduction

Dynamic 3D reconstruction is an important research topic in the field of computer vision, and its applications cover a wide range of fields, such as virtual reality, medical imaging, and industrial automation[1-4]. Dynamic scenes typically involve moving objects or environments, and 3D reconstruction algorithms for such scenes primarily focus on reconstructing non-rigid objects. This often includes addressing general deformations, joint motions, and the capture and reconstruction of human movements. The challenges of dynamic scene 3D reconstruction can be subdivided into related subproblems, including motion estimation, feature extraction and matching, data alignment and fusion, motion removal, and segmentation. These subproblems are intricately interconnected. In recent years, as static scene 3D reconstruction algorithms have matured, research on algorithms for reconstructing dynamic scenes has emerged as a prominent and challenging research focus. Dynamic 3D reconstruction techniques based on the neural radiance field (NeRF) have attracted extensive attention[5-9]. As an emerging computer vision method, NeRFs fully demonstrate their powerful potential in 3D scene reconstruction[10-14]. The purpose of this paper is to deeply explore the key techniques and challenges in 3D reconstruction based on NeRFs.

First, we introduce the fundamentals and working mechanisms of NeRFs to provide researchers with a foundational understanding. NeRFs draw on the ideas of deep learning and neural networks and apply them to 3D reconstruction tasks, bringing new possibilities to dynamic 3D reconstruction by learning the ability to recover 3D information from multi-view images[15-22].

This is followed by an in-depth discussion of the key technical challenges in performing 3D reconstruction in dynamic environments. These challenges include, but are not limited to, viewpoint and illumination variations of moving objects, object identification and modeling, real-time requirements, data acquisition and calibration challenges, the complexity of motion estimation, and effective evaluation mechanisms for reconstruction results. These issues are the core challenges in dynamic 3D reconstruction and require in-depth research and innovative solutions.

Finally, we summarize the current state-of-the-art approaches and trends to address these challenges, and look

This work was supported by ZTE Industry-University-Institute Cooperation Funds under Grant No. 2023ZTE03-04.

ahead to future research directions. By exploring these key techniques and challenges, this work is expected to promote further development in dynamic 3D reconstruction, and provide support and insights for the realization of more accurate, efficient, and widely used 3D reconstruction techniques.

2 Scene Reconstruction with NeRFs

2.1 NeRFs

In 3D reconstruction, NeRFs, as an implicit representation technique, can depict 3D models through implicit functions learned by neural networks. This method has valuable applications in areas like image generation, viewpoint generation, and re-illumination. This section begins by reviewing and introducing the methods that utilize neural networks as implicit representations for scene geometry, before presenting the concept of NeRFs. A prevalent technique for employing neural networks to implicitly represent 3D geometry is the occupancy network[23 - 24]. This approach employs a neural network to predict the binary occupancy of each point in space, essentially training a binary classification network for 3D space, as illustrated in Fig. 1. The key advantage of this method lies in its use of continuous functions to describe 3D space. In comparison to prior approaches such as voxels and meshes, it excels in describing complex geometric shapes without necessitating additional spatial storage.

Apart from directly classifying space into two categories based on model existence, there exists another implicit representation method that portrays the 3D model through the regression of a signed distance function (SDF)[25 - 26]. This approach allows for the continuous representation of 3D models, enabling the modeling of even those with intricate topologies.

Building upon the SDF method, researchers have enhanced and applied it to represent models with intricate details. One notable example is the Pixel Aligned Implicit Function (PIFu) method[26], which captures the details of a 3D model by projecting spatial points onto a pixel-aligned feature space, enabling high-resolution reconstruction, e.g., of a dressed human model. However, these methods often rely on known 3D shapes as supervisory information, which is challenging to obtain in many applications. Consequently, subsequent research has aimed to relax this constraint by directly utilizing images as supervision. For example, some studies introduced differentiable drawing techniques, incorporating rendering steps into neural networks to train the network based on errors in image rendering. NIEMEYER et al.[27] employed a placeholder network as the representation structure for 3D model geometry, determining ray-model surface intersection points using numerical methods. Each intersection point served as input for the neural network to predict the corresponding color value. SITZMANN et al.[28] predicted color and feature vectors for each 3D spatial coordinate, proposing a differentiable drawing function composed of recurrent neural networks to locate the object surface. However, these methods often struggled with complex shapes, limited to handling simple structures with low geometric complexity, yielding overly smooth drawing results. Against this backdrop, MILDENHALL et al.[29] introduced NeRF, a novel representation method that uses only input images as supervisory information. NeRF can accurately fit implicit functions for high-resolution geometric shapes, achieving photo-realistic viewpoint synthesis results for complex scenes. The overall process of this algorithm is depicted in Fig. 1. NeRF employs a multi-layer perceptron to express a 5D vector function, describing both geometric and color information of a 3D model.

NeRF relies solely on input images as supervisory information. This innovative approach excels at fitting precise implicit functions in high-resolution geometric shapes, consequently attaining photo-realistic viewpoint synthesis results for complex scenes. NeRF represents the 3D scene as a differentiable and continuous radiation field F_θ:

$$F_\theta(\boldsymbol{x},\boldsymbol{d}) = [\sigma,\boldsymbol{c}] \quad (1),$$

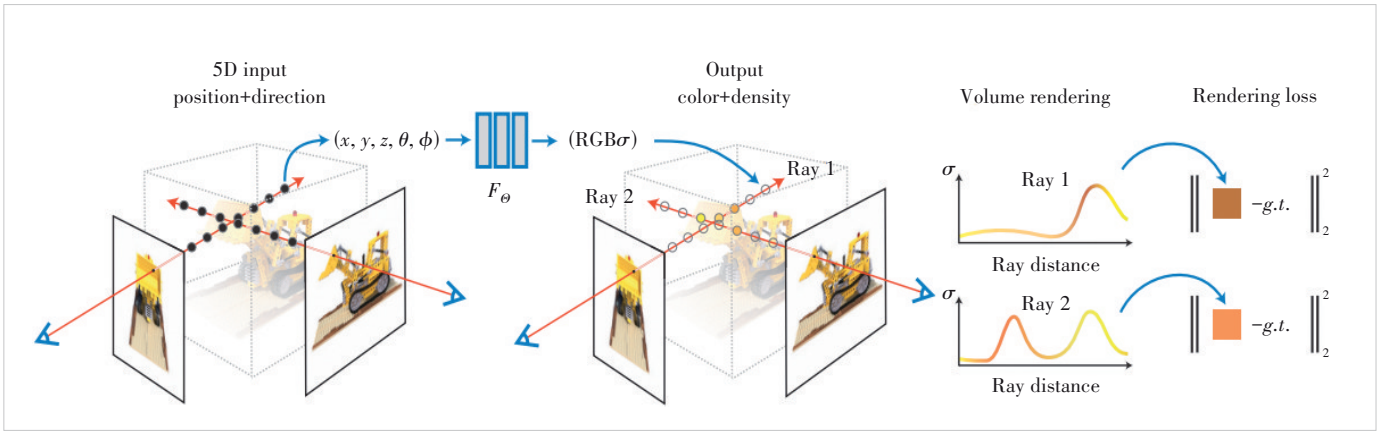

Figure 1. An overview of neural radiance field scene representation and differentiable rendering procedure

where $\boldsymbol{x} = (x, y, z)$ denotes the coordinates of a point in 3D space; $\boldsymbol{d} = (d_x, d_y, d_z)$ denotes the normalized viewing direction; θ is the set of variables that parameterize the model. e.g., a multilayer perceptron (MLP); σ denotes the density estimate at the point $\boldsymbol{x}$, which is the probability that the ray terminates at the point. Assuming that the position of the current camera center is $o \in R^3$ and connecting any pixel on the image with the center, we can get the view direction $d \in R^3$. We parameterize a ray extending from the camera center o, with the view direction d as follows:

$$l(t) = o + td, t \in (-\infty, +\infty) \tag{2}$$

According to the formula of the volume rendering, the color value of the pixel can be expressed as

$$\boldsymbol{C} = \int_{t_n}^{t_f} T(t)\sigma(\boldsymbol{l}(t))\boldsymbol{c}(\boldsymbol{l}(t), \boldsymbol{d})\mathrm{d}t \tag{3}$$

where

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\boldsymbol{l}(s))\mathrm{d}s\right) \tag{4}$$

The transmittance, as defined in Eq. (4), quantifies the probability that a ray, traveling between points t_n and t, is absorbed, scattered, or reflected by objects encountered along its path.

Thanks to the differentiable volume rendering process, this technique can be seamlessly integrated into the training of the aforementioned neural network. This enables a training process that relies exclusively on image color values as the supervisory signal. Furthermore, to prevent the loss of high-frequency information in the synthesized image, NeRF employs positional encoding for input variables[30]. Specifically, this encoding process involves mapping the variables to their Fourier features. Inspired by positional encoding techniques in natural language processing (e.g., those used in Transformers), NeRF adopts a similar approach to encode input coordinates. This approach employs a set of basis functions, which can either be fixed or learned[31]. The spatial embeddings generated by these basis functions simplify the MLP's task of learning the mapping from a location to specific values, as they effectively partition the input space. The positional encoding method used in NeRF is defined as:

$$\boldsymbol{x} \mapsto [\cos(\boldsymbol{M}\boldsymbol{x}), \sin(\boldsymbol{M}\boldsymbol{x})] \tag{5}$$

In Eq. (5),

$$\boldsymbol{M} = [\boldsymbol{I} \quad 2\boldsymbol{I} \quad 2^2\boldsymbol{I} \quad \cdots \quad 2^{p-1}\boldsymbol{I}]^{\top} \tag{6}$$

where $\boldsymbol{x}$ represents the input coordinate, and p stands for a hyperparameter that governs the frequencies utilized, with its value dependent on the target signal resolution. The "soft" binary encoding of the input coordinates is employed, facilitating the network's access to higher frequencies within the input.

2.2 Dynamic Scene Radiance Field Reconstruction

The initial work on NeRF focused exclusively on static scenes. Given that dynamic scenarios are far more prevalent in real-world applications, one of the most critical directions in NeRF advancements is the modeling of radiation fields for dynamic scenes, a branch closely aligned with the demand for realistic 3D scene representation[29, 32–33]. Dynamic scene NeRFs are 3D scene representations learned from a set of posed images. They are formulated to address the challenge of rendering photo-realistic images from unseen viewpoints, and adopt implicit representation based on coordinates, which then maps spatial points to density and color[34–35]. Recent research in this field is extensive. Based on the different reconstruction objects, we elaborate on them from human-based and scene-based reconstruction perspectives, as shown in Table 1.

1) Human-based reconstruction

The dynamic 3D reconstruction of the human body is, to a certain extent, associated with the application requirements of remote presentation, virtual reality, augmented reality, virtual

Table 1. An overview of the human- and scene-based reconstruction methods

Object	Method	Data Attribute	Required Data	3D Representation	Year
Human-based	Neural body[36]	Multi-view	I+P1+S	V	2021
	Neural actors[37]	Multi-view	I+P1+S	P2+VD	2021
	HVTR[38]	Multi-view	P1+S	V	2022
	NDR[39]	Monocular	I+P1	P2+VD	2022
	HumanNeRF[40]	Multi-view	I+P1	P2+VD	2022
	GM-NeRF[41]	Multi-view	I+P1+S	P2+VD	2023
Scene-based	NeRFlow[45]	Multi-view	I+P1	P2+VD	2021
	NeRFPlayer[46]	Multi-view	I+P1	V	2023
	Dynamic-NeRF[47]	Monocular	I+P1+M	P2+VD	2021
	TiNeuVox[48]	Multi-view	I+P1	V	2022
	NRNeRF[49]	Monocular	I+P1	V	2022
	D-NeRF[50]	Multi-view	I+P1	P2+VD	2021
	NRNeRF[51]	Monocular	I+P1	V	2022
	Torf[52]	Multi-view	I+P1	P2+VD	2022
	Neural 3D[53]	Multi-view	I+P1	P2+VD	2022
	DynIBaR[54]	Monocular	I+P1	P2+VD	2023

GM-NeRF: generic model-based neural radiance field
HVTR: hybrid volumetric-textural rendering
I: Images
M: object masks
NDR: neural dynamic reconstruction
NRNeRF: non-rigid neural radiance field
P1: camera poses (exact or approximate)
P2: 3D position
S: skinned multi-person linear prior model
V: neural volumetric
VD: 2D viewing direction

fitting, and similar domains. PENG et al.[36] presented a novel method, named Neural Body, for dynamic human 3D reconstruction. This approach introduces a neural mixed weight domain to generate a deformation field, combining the mixed weight field with 3D human bones to achieve commendable results in dynamic 3D reconstruction of the human body. However, it should be noted that this method relies on bone-driven deformation templates, which exhibit limited universality and necessitate considerable time for reconstruction. Particularly, in cases involving non-rigid deformations with intricate clothing, the reconstruction effectiveness tends to be suboptimal.

Similar to Neural Body, Neural Actors (NA)[37] and hybrid volumetric-textural rendering (HVTR) [38] use the skinned multi-person linear (SMPL) model to represent deformation states. They utilize proxies to explicitly carve out the surrounding 3D space into the canonical pose embedded in NeRF. To facilitate the recovery of high-fidelity details in both geometry and appearance, they employ additional 2D texture maps defined on the SMPL surface as additional conditioning for the NeRF MLP. CAI et al.[39] introduced a template-free method termed neural dynamic reconstruction (NDR), which proficiently reconstructs dynamic scenes from monocular videos. This method leverages color and depth information to optimize surface deformation and employs a neural invertible network to ensure cyclic consistency between any two frames. Additionally, topology-aware networks are employed to model topology variables, effectively addressing challenges related to topology changes. Nonetheless, it's worth noting that the NDR method exhibits subpar reconstruction performance for dynamic scenes with rapid motion and demands a substantial number of computing resources. Another method, named HumanNeRF[40], demonstrates how to train a NeRF for a specific participant based on monocular input data, utilizing a skeleton-driven motion field refined by a general non-rigid motion field. CHEN et al.[41] proposed an effective general framework called the generic model-based neural radiance field (GM-NeRF) for synthesizing free-viewpoint images. Specifically, they first registered the appearance codes of multi-view 2D images onto a geometric proxy through a geometry-guided attention mechanism. This helps mitigate the misalignment between inaccurate geometric priors and the pixel space. Building upon this, they further performed neural rendering and partial gradient backpropagation to achieve efficient perceptual supervision and enhance the perceptual quality of the synthesis.

While the aforementioned approaches yield promising results in portrait scenarios, their applicability declines when dealing with highly non-rigid deformations, particularly for articulated human motion captured from a single view. To address this, methods explicitly leverage human skeleton embeddings. The Neural Articulated Radiance Field (NARF)[42] is trained on pose-annotated images. Joint objects are decomposed into multiple rigid object parts, with their local coordinate systems and global shape variations located at the top. A converged NARF enables novel view rendering via pose manipulation, depth map estimation, and body part segmentation. In contrast, A-NeRF[43] learns actor-specific volumetric neural body models in a self-supervised manner from a monocular camera. This method combines dynamic NeRF volumes with the explicit controllability of articulated human skeletons and reconstructs poses and radiance fields through a comprehensive analysis approach. Once trained, the radiance field can be used for novel view synthesis and motion retargeting. They demonstrate the benefits of using the learned non-surface model, which enhances the accuracy of human pose estimation in monocular videos through photometric reconstruction loss. A-NeRF is trained on monocular video, while Animatable NeRFs (ANRF)[44] is a skeleton-driven approach used for reconstructing human body models from multi-view videos. Its core component is a novel motion representation called the neural blend weight field, which is combined with the 3D human skeleton to generate a deformation field. Similar to several general non-rigid NeRF approaches, ANRF maintains a canonical space and estimates bidirectional correspondences between multi-view inputs and canonical frames. The reconstructed animatable human body model can be used for free-viewpoint rendering and re-rendering under new poses. Additionally, human meshes can be extracted from ANRF by applying marching cubes to the volume density of discretized canonical space points. The method achieves high visual accuracy for the learned human body model, and the authors suggest addressing complex non-rigid deformations on observed surfaces, such as those caused by looseness. The authors recommend future work to improve the handling of complex non-rigid deformations on observed surfaces, such as those caused by loose clothing.

2) Scene-based reconstruction

DU et al.[45] proposed NeRFlow to learn dynamic 4D spatio-temporal scenes. NeRFlow consists of two separate modules: a radiation field (top) trained by neural rendering, and a flow field (bottom) trained using 3D keypoint correspondence. The two fields are then kept consistent, which enables the radiation field to acquire prior information from earlier states. SONG et al.[46] used a feature flow approach to model dynamic radiation scenes. The authors mainly used time-dependent sliding windows for points in 4D space to generate flow features, and then decomposed the dynamic scene into predicted static fields, deformation fields, and new scene decomposition fields via a point-by-point probabilistic method. Finally, the expectation of the decomposition fields was fed into NeRF for modeling. However, since local feature channels were used to model each frame in the scene, which enables streaming but limits the representation of temporally distant repetitive activities, it might be used multiple times to reconstruct the same action, resulting in a waste of time. GAO et al.[47] proposed DynamicNeRF, an algorithm for generating novel views from any viewpoint of a monocular dynamic scene video and any input time step. The algorithm takes a monocular video with N

frames and a binary mask of the foreground object for each frame as input, and models the time-varying structure and the appearance of the scene using continuous and differentiable functions.

However, some authors believe that the key problem in solving dynamic scene rendering lies in the encoding of temporal information. FANG et al.[48] proposed TiNeuVox, which uses a combination of optimizable explicit voxel features and temporal information encoding to quickly generate dynamic scenes. They first input the point coordinates and temporal coding into a deformation network to obtain the offset coordinates, then interpolated the voxels according to the offset coordinates to obtain the voxel features, and finally connected the original coordinates, temporal coding, and voxel features and fed them into a NeRF network to obtain the colors and densities. However, they did not consider the relationship between neighboring frames, so there is a slight problem with the coherence of the video. In addition, ABOU-CHAKRA et al.[49] introduced an online method for generating dynamic scenes. Inspired by particle dynamics, they proposed a new particle coding that enables the intermediate features of NeRF to move in conjunction with the geometry they represent. As a result, the authors have achieved automated generation of dynamic scenes by back-projecting rendering losses to particle positions and encoding particle parameters.

Another class of methods introduces additional deformation fields to predict the motion of points by mapping their coordinates to a normative space where large motion or geometric changes can be captured and learned. PUMAROLA et al.[50] proposed a method to extend NeRFs to the dynamic domain, D-NeRF, which allows a single camera to reconstruct and draw a new image as it moves under both rigid and non-rigid motion images of the scene. Therefore, it is necessary to include time as an additional input to the system and to divide the learning process into two main phases: one phase encodes the scene into a canonical space and the other maps this canonical representation to a deformed scene at a specific time.

Other methods improve dynamic neural rendering in various ways, e.g., distinguishing between foreground and background. TRETSCHK et al.[51] proposed non-rigid NeRF (NRNeRF), a reconstruction and new view synthesis method for general non-rigid dynamic scenes. The method takes an RGB image of a dynamic scene (e.g., from monocular video recordings) as input and creates high-quality representations of spatio-temporal geometry and appearance. Meanwhile, quality enhancement using depth information can improve dynamic neural rendering. ATTAL et al.[52] noted that neural networks can represent and accurately reconstruct the radiance field of a static 3D scene (e.g., NeRF). However, dynamic scene approaches for monocular video capture rely on data-driven priors to reconstruct dynamic content. To address this, the authors replaced this a priori information with time-of-flight (TOF) camera measurements and introduced a neural representation based on a continuous-wave TOF camera image formation model. Instead of using processed depth maps, the method models the raw TOF sensor measurements to improve the reconstruction quality and to avoid the problems of low reflectivity regions, multipath interference, and the limited explicit depth range of the sensor. Additionally, setting keyframes to produce sharper results is another effective approach. LI et al.[53] proposed a new 3D video synthesis method that compactly and expressively represents multi-viewpoint video recordings of dynamic real scenes, allowing for high-quality viewpoint synthesis and motion interpolation.

The state-of-the-art method based on temporally varying NeRFs, also known as Dynamic NeRFs, has demonstrated impressive results in this task. However, for long videos with complex object motions and uncontrolled camera trajectories, the method may result in blurry or inaccurate renderings. To address this issue, LI et al.[54] proposed a novel approach. Instead of encoding the entire dynamic scene within the weights of an MLP, this method employs a volume-image-based rendering framework. This framework synthesizes new viewpoints by aggregating features from nearby views in a scene-motion-aware manner, overcoming these limitations. The system retains the capability of previous methods to model complex scenes and view-dependent effects. Still, it can also synthesize realistic new views for long videos with complex dynamic scenes and unconstrained camera trajectories.

3 Database and Evaluation

3.1 Common Database

We present the common database in this section in Table 2.

1) DNA-Rendering[55] is a large-scale, high-fidelity repository for neural actor rendering, represented by neural implicit fields of human actors. This dataset contains data from 500 individuals, with 527 distinct sets of clothing, 269 types of daily actions, and 153 types of special performances, including relevant interactive objects for some actions. Additionally, a pro-

Table 2. Information on commonly used datasets for dynamic 3D reconstruction

Name	Object	Cases	Cameras	Resolution	Year
DNA-Rendering	Human-based	439	60	4K	2023
ZJU_MoCap	Human-based	9	23	1K	2021
ENeRF-Outdorr	Scene-based	8	18	4K	2022
NVIDIA	Scene-based	12	4 scenes with monocular; 8 scenes with 12 cameras	960×540	2020

fessional multi-view system was constructed to capture data, which contains 60 synchronous cameras with a max resolution of 4 096×3 000 and a frame rate of 15 frames per second.

2) ZJU_MoCap[36]. This dataset captures nine dynamic human videos using a multi-camera system that has 21 synchronized cameras. All sequences have a length ranging from 60 to 300 frames. In these videos, humans perform complex motions, including twirling, Taichi, arm swinging, warmup, punching, and kicking.

3) ENeRF-Outdoor[56] is a dynamic dataset of multi-purpose outdoor scenes, collected by 18 synchronized cameras. Each sequence generally has about 1 000 frames and complex motions.

4) NVIDIA[57]. This dataset collects dynamic scenes using two methods: a) Moving monocular camera: Short-term dynamic events (about 5 s) are captured by a hand-held monocular camera (Samsung Galaxy Note 10) with a frame rate of 60 frames per second and a resolution of 1 920×1 080. Sequences are subsampled if the object motion is not salient, and therefore, the degree of the scene motion is significantly larger than that of the camera's ego motion, making quasi-static dynamic reconstruction inapplicable. Four dynamic scenes are captured, including human activity, human-object interactions, and animal movements; b) Stationary multi-view cameras: Eight scenes are captured by a static camera rig with 12 cameras (GoPro Black Edition).

3.2 Evaluation Metrics

The synthesis of novel views through NeRF employs visual quality assessment metrics as benchmarks. These metrics aim to evaluate the quality of individual images with (full-reference) or without (no-reference) ground truth images. To date, peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM)[58], and learned perceptual image patch similarity (LPIPS)[59] are the most commonly used metrics in NeRF related literature.

1) PSNR is one of the important metrics for measuring image quality. The formula for calculating PSNR is as follows:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right) \tag{7},$$

where MAX is the maximum possible range of pixel values in the image (usually 255 for 8-bit images), and MSE is the average of squared differences between corresponding pixels. A higher PSNR value indicates better image quality, making it a widely used standard for evaluating image reconstruction quality in image processing and compression. It is important to note that PSNR may not fully align with human perception of image quality. Therefore, in certain applications, other metrics such as SSIM or LPIPS are employed to more comprehensively assess image quality.

2) SSIM consists of three contrast similarity modules, namely: luminance, contrast, and structure. Luminance modules can be written as:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{8},$$

where μ_x and μ_y are the average gray values of images I_x and I_y, respectively; C_1 is a constant. Contrast modules can be written as:

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{9},$$

where σ_x and σ_y are the standard deviations of images I_x and I_y, respectively; C_2 is a constant. Structure modules can be written as:

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{10}.$$

Finally, SSIM can be formulated as:

$$\mathrm{SSIM}(x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{11}.$$

3) LPIPS. The LPIPS distance is used to measure the average feature distances between two images, which is calculated from the weighted pixel-level MSE of the multilayer feature maps.

$$\mathrm{LPIPS}(x,y) = \sum_l^L \frac{1}{H_l W_l} \sum_{h,w}^{H_l,W_l} \left\| w_l \odot \left(x_{hw}^l - y_{hw}^l\right) \right\|_2^2 \tag{12},$$

where x_{hw}^l and y_{hw}^l are the features of the reference and evaluation images at pixel width w, pixel height h, and layer l. H_l and W_l are the height and width of the feature maps of the corresponding layers.

4 Challenges

3D reconstruction in dynamic environments involves a series of complex and important technical challenges. One is perspective and illumination changes of moving objects. In an ever-changing dynamic scene, the positions and orientations of objects are constantly changing over time, and thus their appearance and perspective change significantly at different moments. In addition, lighting conditions may vary constantly across time and space, further complicating the accurate capture and modeling of moving objects. Object recognition and modeling is another challenging area. The need to reliably identify and track multiple moving objects in dynamic environments requires highly accurate object recognition and modeling techniques to efficiently handle complex scenes and ensure the accuracy and consistency of 3D models.

Data acquisition can be challenging in dynamic environments where the position and orientation of moving objects may change over time. Therefore, it is important to address the complexity of data acquisition and ensure the accuracy and consistency of the sensors. Another major problem faced in dynamic 3D reconstruction compared with static 3D reconstruction is the need to accurately estimate camera and object motion. Finally, in order to validate and evaluate the quality of 3D reconstruction results, it is necessary to develop evaluation methods and metrics applicable to dynamic environments, especially for monocular dynamic scene reconstruction. How to evaluate the quality of image reconstruction from different viewpoints at the same moment is a very urgent problem in engineering practice. Comprehensive consideration of these technical challenges and continuous improvement of algorithms and methods are the key to realizing high-quality 3D reconstruction in dynamic environments.

Fig. 2 shows a 3D reconstruction process for dynamic and static distributed scenes, which is also the basic process used in practice. Taking this process as an example, we will specifically introduce the key issues and challenges in dynamic 3D reconstruction. When the input is a video stream from a common capture device, extensive preprocessing is required. The video stream is first decomposed into video frames, and then the existing or improved instance segmentation algorithms are applied to the RGB image of these frames to generate static and dynamic masks. Due to the input requirements and limitations of the existing terminal memory and graphics, we need to reasonably select key frames from the full set of video frames, including the key information of the dynamic scene, continuous changes in motion, significant changes in illumination, accurate camera viewpoints, frames with overlapping regions, the depth, and optical flow of the correctly calculated. The effective high-precision frame information can maximize the quality and accuracy of the 3D reconstruction of the dynamic scene and help produce better results.

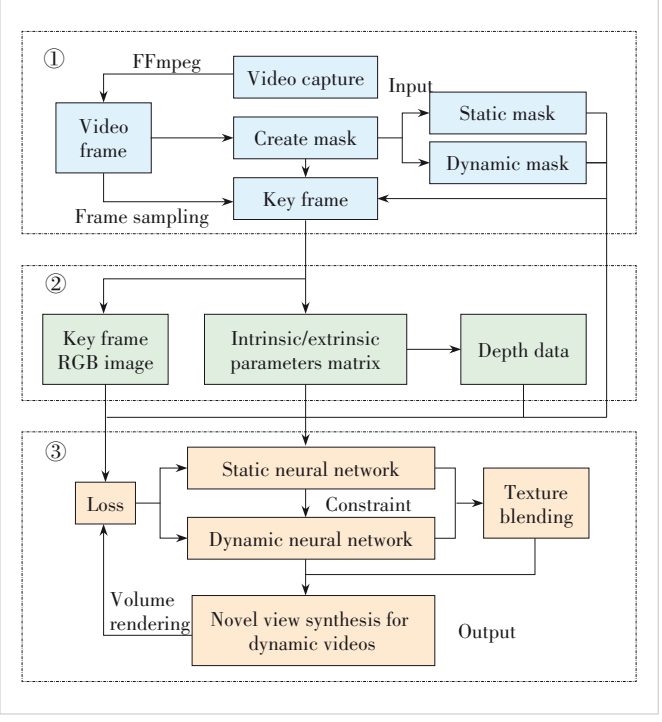

Figure 2. A novel viewpoint synthesis framework based on dynamic and static distributions

The obtained information is then fed into the neural network for processing. In static NeRF, only the position information is input to derive static color values and density features. Since a dynamic scene exhibits two different attributes, static and dynamic, at the same sampling point under different viewpoints and at different times, the static output is used as part of the input to the dynamic neural network. This network is trained with the spatiotemporal information, thereby constraining the overall convergence. Our research team is constantly conducting experiments, and one of the key challenges lies in the sparsity of the dynamic data, which makes it difficult to achieve high precision results. Therefore, there is an urgent need to explore additional constraints and methods to improve modeling accuracy. These constraints can cover a number of aspects, including but not limited to depth constraints, temporal continuity, motion modeling, optical flow coherence, and multi-sensor fusion. By introducing these constraints, we hope to make a bigger breakthrough in reconstruction quality, which in turn will enable more accurate capture and reproduction of complex dynamic scenes. Finally, voxel rendering is performed on the dynamic and static data obtained from training, in order to complete the model reconstruction based on NeRF rendering and to generate new viewpoints over time.

5 Development Trends

The field of 3D reconstruction with NeRF still faces a series of problems, and the following are what we consider as possible future research directions: 1) Developing robust machine learning and deep learning models with generalized modeling capabilities for dynamic scenes, including improving model robustness to handle challenges such as noise, occlusion, and incomplete data; 2) Exploring more effective constraints and implicit modeling in which the physical and geometric properties of the scene are better captured; 3) Advancing multimodal fusion, including images, point clouds, sound, etc., which helps improve the understanding and modeling of dynamic scenes and makes reconstruction results more comprehensive and accurate; 4) Promoting self-supervised learning to reduce the dependence on labeled data. Especially in the absence of large-scale labeled data, self-supervised learning methods can improve the performance of dynamic 3D reconstruction; 5) Conducting semantic modeling of dynamic scenes.

6 Conclusions

In this paper, we delve into the key techniques and challenges of dynamic 3D reconstruction based on NeRFs. Dynamic 3D reconstruction is an important research direction in the field of computer vision with a wide range of applications. As an emerging computer vision method, NeRF has a strong potential in 3D scene reconstruction.

This paper first introduces the basic principles and working mechanism of NeRF, which provides readers with a foundation for understanding this novel approach. NeRF draws on deep learning and neural networks and applies them to the 3D reconstruction task to learn to recover 3D information from multi-view images. It then presents an in-depth discussion of the key technical challenges facing 3D reconstruction in dynamic environments, including viewpoint and illumination variations of moving objects, object recognition and modeling, real-time requirements, data acquisition and calibration challenges, complexity of motion estimation, and effective mechanisms for evaluating reconstruction results. These core challenges require in-depth research and innovative solutions. In addition, this paper summarizes current technology trends and approaches to address these challenges and outlines future research directions. It emphasizes the importance of directions such as robustness of machine learning and deep learning models, more efficient constraints, multimodal fusion, self-supervised learning, and semantic modeling of dynamic scenes.

Finally, this paper emphasizes that the field of dynamic 3D reconstruction will continue to thrive with the rise of the metaverse. These research directions will help to continuously improve the performance and applicability of dynamic 3D reconstruction, drive innovation and development in this field, and create more exciting applications and possibilities. We look forward to making more breakthroughs in this challenging and opportune field and contributing more support and insight to the future of 3D reconstruction technology.

References

[1] GONZÁLEZ IZARD S, SÁNCHEZ TORRES R, ALONSO PLAZA Ó, et al. Nextmed: automatic imaging segmentation, 3D reconstruction, and 3D model visualization platform using augmented and virtual reality [J]. Sensors, 2020, 20(10): 2962. DOI: 10.3390/s20102962

[2] LI H M. 3D indoor scene reconstruction and layout based on virtual reality technology and few-shot learning [EB/OL]. [2024-01-02]. https://onlinelibrary. wiley. com/doi/full/10.1155/2022/4134086? msockid=271754324752654020fd45de467c6460

[3] TANG F L, WU Y H, HOU X H, et al. 3D mapping and 6D pose computation for real time augmented reality on cylindrical objects [J]. IEEE transactions on circuits and systems for video technology, 2020, 30(9): 2887 - 2899. DOI: 10.1109/TCSVT.2019.2950449

[4] SAMAVATI T, SORYANI M. Deep learning-based 3D reconstruction: a survey [J]. Artificial intelligence review, 2023, 56(9): 9175 - 9219. DOI: 10.1007/s10462-023-10399-2

[5] PALAZZOLO E, BEHLEY J, LOTTES P, et al. ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals [C]// International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 7855 - 7862. DOI: 10.1109/IROS40897.2019.8967590

[6] STIER N, ANGLES B, YANG L, et al. LivePose: online 3D reconstruction from monocular video with dynamic camera poses [EB/OL]. [2024-01-02]. https://openaccess. thecvf. com/content/ICCV2023/papers/Stier_LivePose_Online_3D_Reconstruction_from_Monocular_Video_with_Dynamic_Camera_ICCV_2023_paper.pdf

[7] NOVOTNY D, ROCCO I, SINHA S, et al. KeyTr: keypoint transporter for 3D reconstruction of deformable objects in videos [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 5585 - 5594. DOI: 10.1109/CVPR52688.2022.00551

[8] CHEN X T, SRA M. IntoTheVideos: exploration of dynamic 3D space reconstruction from single sports videos [C]//The 34th Annual ACM Symposium on User Interface Software and Technology. ACM, 2021: 14 - 16. DOI: 10.1145/3474349.3480215

[9] WANG B, JIN Y, CHEN Y X, et al. Gaze tracking 3D reconstruction of object with large-scale motion [J]. IEEE transactions on instrumentation and measurement, 2023, 72: 7002612. DOI: 10.1109/TIM.2023.3251419

[10] REMONDINO F, KARAMI A, YAN Z Y, et al. A critical analysis of NeRF-based 3D reconstruction [J]. Remote sensing, 2023, 15(14): 3585. DOI: 10.3390/rs15143585

[11] CHEN H S, GU J T, CHEN A P, et al. Single-stage diffusion nerf: a unified approach to 3D generation and reconstruction [EB/OL]. (2023-04-13) [2024-01-02]. https://arxiv.org/abs/2304.06714

[12] XU H Y, ALLDIECK T, SMINCHISESCU C. H-nerf: neural radiance fields for rendering and temporal reconstruction of humans in motion [EB/OL]. (2021-10-26) [2024-01-02]. https://arxiv.org/abs/2110.13746

[13] XU J K, PENG L, CHEN H R, et al. MonoNeRD: NeRF-like representations for monocular 3D object detection [C]//International Conference on Computer Vision. IEEE, 2023: 6791 - 6801. DOI: 10.1109/ICCV51070.2023.00627

[14] LI S X, LI C J, ZHU W B, et al. Instant-3D: instant neural radiance field training towards on-device AR/VR 3D reconstruction [C]//The 50th Annual International Symposium on Computer Architecture. ACM, 2023: 1 - 13. DOI: 10.1145/3579371.3589115

[15] KIRSCHSTEIN T, QIAN S, GIEBENHAIN S, et al. NeRSemble: multi-view radiance field reconstruction of human heads [EB/OL]. (2023-05-04) [2024-01-02]. https://arxiv.org/abs/2305.03027

[16] CHEN J, YI W, MA L, et al. GM-NeRF: learning generalizable model-based neural radiance fields from multi-view images [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2023: 20648 - 20658

[17] GAO H, LI R, TULSIANI S, et al. Monocular dynamic view synthesis: a reality check [J]. Advances in neural information processing systems, 2022, 35: 33768 - 33780

[18] LI T Y, SLAVCHEVA M, ZOLLHOEFER M, et al. Neural 3D video synthesis from multi-view video [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 5511 - 5521. DOI: 10.1109/CVPR52688.2022.00544

[19] ZHANG J Z, LUO H M, YANG H D, et al. NeuralDome: a neural modeling pipeline on multi-view human-object interactions [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 8834 - 8845. DOI: 10.1109/CVPR52729.2023.00853

[20] WEI Y, LIU S H, RAO Y M, et al. NerfingMVS: guided optimization of neural radiance fields for indoor multi-view stereo [C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 5590 - 5599. DOI: 10.1109/ICCV48922.2021.00556

[21] DENG F W, HUANG S J. High-precision 3D structure optical measurement technology for 5G power modules [J]. ZTE technology journal, 2024,

30(5): 75 – 80. DOI: 10.12142/ZTETJ.202405011
[22] FENG D Q, ZHANG S L, LYU X Y, et al. Metaverse: concept, architecture, and suggestions [J]. ZTE technology journal, 2024, 30(S1): 3 – 15. DOI: 10.12142/ZTETJ.2024S1002
[23] MESCHEDER L, OECHSLE M, NIEMEYER M, et al. Occupancy networks: learning 3D reconstruction in function space [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 4455 – 4465. DOI: 10.1109/CVPR.2019.00459
[24] CHEN Z Q, ZHANG H. Learning implicit fields for generative shape modeling [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 5932 – 5941. DOI: 10.1109/cvpr.2019.00609
[25] PARK J J, FLORENCE P, STRAUB J, et al. Deepsdf: learning continuous signed distance functions for shape representation [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE, 2019: 165 – 174
[26] SAITO S, HUANG Z, NATSUME R, et al. PIFu: pixel-aligned implicit function for high-resolution clothed human digitization [C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 2304 – 2314. DOI: 10.1109/ICCV.2019.00239
[27] NIEMEYER M, MESCHEDER L, OECHSLE M, et al. Differentiable volumetric rendering: learning implicit 3D representations without 3D supervision [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 3501 – 3512. DOI: 10.1109/cvpr42600.2020.00356
[28] SITZMANN V, ZOLLHÖFER M, WETZSTEIN G. Scene representation networks: continuous 3D-structure-aware neural scene representations [C]//The 33rd International Conference on Neural Information Processing Systems. ACM, 2019: 1121 – 1132
[29] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: representing scenes as neural radiance fields for view synthesis [EB/OL]. (2020-03-19) [2024-01-02]. https://arxiv.org/abs/2003.08934
[30] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//The 31st International Conference on Neural Information Processing System. ACM, 2017: 6000 – 6010
[31] TANCIK M, SRINIVASAN P, MILDENHALL B, et al. Fourier features let networks learn high frequency functions in low dimensional domains [C]//The 34th International Conference on Neural Information Processing Systems. ACM, 2020: 7537 – 7547
[32] YU A, YE V, TANCIK M, et al. PixelNeRF: neural radiance fields from one or few images [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 4576 – 4585. DOI: 10.1109/cvpr46437.2021.00455
[33] ZHANG K, RIEGLER G, SNAVELY N, et al. Nerf++: analyzing and improving neural radiance fields [EB/OL]. (2020-10-15) [2024-01-02]. https://arxiv.org/abs/2010.07492
[34] WANG Z, WU S, XIE W, et al. NeRF--: neural radiance fields without known camera parameters [EB/OL]. (2021-02-14) [2024-01-02]. https://arxiv.org/abs/2102.07064
[35] BARRON J T, MILDENHALL B, TANCIK M, et al. Mip-nerf: a multiscale representation for anti-aliasing neural radiance fields [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2021: 5855 – 5864. DOI: 10.1109/ICCV48922.2021.00580
[36] PENG S, ZHANG Y, XU Y, et al. Neural body: implicit neural representations with structured latent codes for novel view synthesis of dynamic humans [J]. IEEE Transactions on pattern analysis and machine intelligence. 2021: 9054 – 9063
[37] LIU L J, HABERMANN M, RUDNEV V, et al. Neural actor [J]. ACM transactions on graphics, 2021, 40(6): 1 – 16. DOI: 10.1145/3478513.3480528
[38] HU T, YU T, ZHENG Z R, et al. HVTR: hybrid volumetric-textural rendering for human avatars [C]//International Conference on 3D Vision (3DV). IEEE, 2022: 197 – 208. DOI: 10.1109/3DV57658.2022.00032
[39] CAI H, FENG W, FENG X, et al. Neural surface reconstruction of dynamic scenes with monocular RGB-D camera [J]. Advances in neural information processing systems, 2022, 35: 967 – 981
[40] WENG C, CURLESS B, SRINIVASAN P P, et al. HumanNeRF: free-viewpoint rendering of moving people from monocular video [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 16189 – 16199. DOI: 10.1109/CVPR52688.2022.01573
[41] CHEN J C, YI W T, MA L Q, et al. GM-NeRF: learning generalizable model-based neural radiance fields from multi-view images [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 20648 – 20658. DOI: 10.1109/CVPR52729.2023.01978
[42] NOGUCHI A, SUN X, LIN S, et al. Neural articulated radiance field [C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 5742 – 5752. DOI: 10.1109/ICCV48922.2021.00571
[43] SU S Y, YU F, ZOLLHOEFER M, et al. A-nerf: surface-free human 3D pose refinement via neural rendering [EB/OL]. (2021-10-29) [2024-01-02]. https://arxiv.org/abs/2102.06199v1
[44] PENG S, DONG J, WANG Q, et al. Animatable neural radiance fields for human body modeling [EB/OL]. (2021-10-07) [2024-01-02]. https://arxiv.org/abs/2105.02872
[45] DU Y L, ZHANG Y N, YU H X, et al. Neural radiance flow for 4D view synthesis and video processing [C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 14304 – 14314. DOI: 10.1109/ICCV48922.2021.01406
[46] SONG L C, CHEN A P, LI Z, et al. NeRFPlayer: a streamable dynamic scene representation with decomposed neural radiance fields [J]. IEEE transactions on visualization and computer graphics, 2023, 29(5): 2732 – 2742. DOI: 10.1109/TVCG.2023.3247082
[47] GAO C, SARAF A, KOPF J, et al. Dynamic view synthesis from dynamic monocular video [C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 5692 – 5701. DOI: 10.1109/ICCV48922.2021.00566
[48] FANG J M, YI T R, WANG X G, et al. Fast dynamic radiance fields with time-aware neural voxels [C]//Proceedings of SIGGRAPH Asia 2022 Conference Papers. ACM, 2022: 1 – 9. DOI: 10.1145/3550469.3555383
[49] ABOU-CHAKRA J, DAYOUB F, SÜNDERHAUF N. Particlenerf: particle based encoding for online neural radiance fields in dynamic scenes [EB/OL]. (2023-03-24) [2024-01-02]. https://arxiv. org/abs/2211.04041
[50] PUMAROLA A, CORONA E, PONS-MOLL G, et al. D-NeRF: neural radiance fields for dynamic scenes [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 10313 – 10322. DOI: 10.1109/cvpr46437.2021.01018
[51] TRETSCHK E, TEWARI A, GOLYANIK V, et al. Non-rigid neural radiance fields: reconstruction and novel view synthesis of a dynamic scene from monocular video [EB/OL]. (2020-12-22) [2024-01-02]. https://arxiv.org/abs/2012.12247
[52] ATTAL B, LAIDLAW E, GOKASLAN A, et al. Törf: time-of-flight radiance fields for dynamic scene view synthesis [C]//The 35th International Conference on Neural Information Processing Systems. ACM, 2021: 26289 – 26301
[53] LI T, SLAVCHEVA M, ZOLLHOEFER M, et al. Neural 3D video synthesis [EB/OL]. (2021-03-03) [2024-01-02]. https://arxiv.org/abs/2103.02597
[54] LI Z Q, WANG Q Q, COLE F, et al. DynIBaR: neural dynamic image-based rendering [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 4273 – 4284. DOI: 10.1109/CVPR52729.2023.00416
[55] CHENG W, CHEN R X, FAN S M, et al. DNA-rendering: a diverse neural actor repository for high-fidelity human-centric rendering [C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023: 19925 – 19936. DOI: 10.1109/ICCV51070.2023.01829

[56] LIN H T, PENG S D, XU Z, et al. Efficient neural radiance fields for interactive free-viewpoint video [C]//Proceedings of SIGGRAPH Asia 2022 Conference Papers. ACM, 2022: 1 – 9. DOI: 10.1145/3550469.3555376
[57] YOON J S, KIM K, GALLO O, et al. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 5336 – 5345
[58] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE transactions on image processing, 2004, 13(4): 600 – 612. DOI: 10.1109/tip.2003.819861
[59] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 586 – 595. DOI: 10.1109/CVPR.2018.00068

Biographies

LU Ping is the vice president and general manager of the Industrial Digitalization Solution Department of Beijing XingYun Digital Technology Co., Ltd. and the executive deputy director of the National Key Laboratory of Mobile Network and Mobile Multimedia Technology, China. His research directions include cloud computing, big data, augmented reality, and multimedia service-based technologies. He has supported and participated in major national science and technology projects. He has published multiple papers and authored two books.

FENG Daquan is currently a distinguished professor and PhD supervisor with the College of Electronics and Information Engineering, Shenzhen University, China. He has authored or coauthored over 80 papers in refereed journals and conferences, with more than 5 000 citations. His research interests include 3D reconstruction, generative artificial intelligence, and immersive communication. He is the winner of the First Prize in Natural Science from the China Institute of Electronics in 2023 and the National Science Funds for the Excellent Young Scientists (NSFC) in 2024. He was a recipient of the Best Paper Awards of IEEE TSC 2023, DCN 2023, and COMCOMAP 2021.

SHI Wenzhe (shi. wenzhe@xydigit. com) is a strategy planning engineer with Beijing XingYun Digital Technology Co., Ltd., a member of the National Key Laboratory for Mobile Network and Mobile Multimedia Technology, China. His research interests include indoor visual AR navigation, SFM 3D reconstruction, visual SLAM, real-time cloud rendering, VR, and spatial perception.

LI Wan received her ME degree in information and communication engineering from the School of Information Engineering, Chang'an University, China in 2020. She is currently pursuing her PhD degree at the College of Electronics and Information Engineering, Shenzhen University, China. Her research interests include computer vision and 3D reconstruction.

LIN Jiaxin received his ME degree in electronic and communication engineering from the College of Electronics and Information Engineering, Shenzhen University, China in 2020. He is currently pursuing his PhD degree at the College of Electronics and Information Engineering, Shenzhen University. His research interests include 3D vision and neural rendering.