Toward Next-generation Volumetric Video Streaming with Neural-based Content Representations

Kaiyan Liu*, Ruizhi Cheng*, Nan Wu*, and Bo Han
George Mason University
{kliu23,rcheng4,nwu5,bohan}@gmu.edu

Abstract

Striking a balance between minimizing bandwidth consumption and maintaining high visual quality stands as the paramount objective in volumetric content delivery. However, achieving this ambitious target is a substantial challenge, especially for mobile devices with constrained computational resources, given the voluminous amount of 3D data to be streamed, strict latency requirements, and high computational load. Inspired by the advantages offered by neural radiance fields (NeRF), we propose, for the first time, to deliver volumetric videos by utilizing neural-based content representations. We delve deep into potential challenges and explore viable solutions for both video-on-demand (VOD) and live video streaming services, in terms of the end-to-end pipeline, real-time and high-quality streaming, rate adaptation, and viewport adaptation. Our preliminary results lend credence to the feasibility of our research proposition, offering a promising starting point for further investigation.

CCS Concepts

• Information systems → Multimedia streaming; **• Computing methodologies → Mixed / augmented reality**.

Keywords

Volumetric Video Streaming and Neural Radiance Fields

ACM Reference Format:
Kaiyan Liu*, Ruizhi Cheng*, Nan Wu*, and Bo Han. 2023. Toward Next-generation Volumetric Video Streaming with Neural-based Content Representations. In *The 1st ACM Workshop on Mobile Immersive Computing, Networking, and Systems (ImmerCom '23), October 6, 2023, Madrid, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3570361.3613469

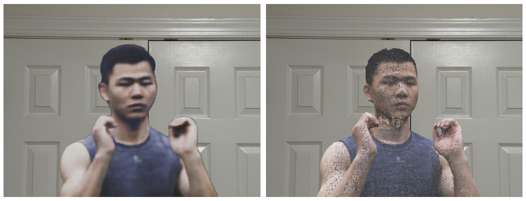

Figure 1: Comparison of volumetric content rendered on HoloLens, with NeRF (left) and point cloud (right).

1 Introduction

Holographic communication [9], a major beneficiary of 3D content delivery, harnesses volumetric content to construct holograms that depict 3D objects or scenes, thereby offering an immersive experience for users. A key characteristic of volumetric content is its provision of six degrees of freedom (6DoF) in movement, enabling users to not only change viewing angles but also freely navigate in the 3D space.

While there have been increasing efforts in recent years to optimize volumetric content delivery and enhance its quality of experience(QoE) [19, 23, 29, 73–76], existing work still falls short in several areas. For example, traditional representation methods with point clouds and meshes [46, 54] have limitations when it comes to representing dynamic elements [32, 41] and lighting effects [10, 63], owing to their discrete nature. Thus, these techniques often fail to achieve photo-realistic rendering quality, affecting the QoE.

The latest advancements in implicit neural representations, such as neural radiance fields (NeRF) [38], have gained popularity as an attractive alternative for representing volumetric content with high visual quality [16, 37, 47], as shown in Figure 1. NeRF is a neural-based method for generating high-quality images through novel view synthesis. Rather than relying on discrete points or polygons, it leverages a multilayer perceptron (MLP) to depict a scene as a continuous function, enabling the rendering of photo-realistic images for an immersive viewing experience. It maps a continuous space of 3D position and viewing direction to a density and view-dependent radiance, leading to the creation of a 2D image through volume rendering, a process that aggregates colors along each ray.

Given that the vanilla NeRF is computationally intensive and primarily suited for static scenes, recent work concentrates on optimizing the performance of NeRF [6, 15, 26, 70]

[*] These authors contributed equally to this work.

ImmerCom '23, October 6, 2023, Madrid, Spain

ACM ISBN 979-8-4007-0339-3/23/10.
https://doi.org/10.1145/3570361.3613469

and extending it for dynamic scenes [13, 44, 45, 56, 57, 64]. Nevertheless, current endeavors have yet to address the networking and systems challenges associated with streaming volumetric content with NeRF.

In this paper, we propose to deliver volumetric content by leveraging neural-based representations. Our goal is to make NeRF-based volumetric video streaming systems practical by boosting network efficiency and enhancing QoE. Despite extensive research towards NeRF for dynamic scenes, the design and implementation of NeRF-based volumetric video streaming systems still pose the following challenges: (1) the trade-offs between model size, inference time, and visual quality; (2) the feasibility of incorporating rate and viewport adaptation into NeRF models; and (3) the stringent real-time requirement for live video streaming.

Our research aims to provide novel insights and potential strategies for real-time, high-quality volumetric video streaming systems by leveraging the power of NeRF. The motivation to utilize NeRF for representing volumetric content comes from the inherent capability of NeRF to synthesize photo-realistic 3D scenes directly from 2D images, an attribute particularly valuable for outdoor scenes [35], where content capturing via RGB-D cameras is less effective. Driven by these remarkable capabilities, our work delves into the potential challenges and solutions in applying NeRF-based methods for video-on-demand (VOD) and live streaming services. In summary, our work has the following contributions.

- We first explore the research challenges tied to NeRF-based volumetric content delivery in VOD services and propose potential solutions (§3.2-§3.4). For example, high-resolution image rendering is still demanding due to the increased latency for processing more pixels. We propose to leverage foveated rendering [3, 39], which reduces the total number of to-be-rendered pixels and decreases the overall computation load. On the other hand, delivering NeRF models over the Internet to represent volumetric content may be bandwidth-intensive. Thus, we propose to explore model compression [60], rate adaptation with scalable neural networks [7, 69], and viewport adaptation to alleviate bandwidth consumption.
- We then investigate NeRF-based live volumetric video streaming, a promising avenue for next-generation services such as telesurgery [8] and remote collaboration [61] (§3.5). The main challenges stem from the need for real-time and continuous learning since future frames in a live setting are unknown. Given that the real-time training of NeRF models remains challenging, we propose an acceleration approach that involves offline pre-training of the model for the initial scene, followed by frame-specific fine-tuning based on pixel alterations between subsequent frames.
- Finally, we study the feasibility of NeRF-based volumetric content delivery that utilizes state-of-the-art NeRF models for dynamic scenes. Our preliminary results indicate that there exists a trade-off between model size, inference time, and visual quality. In addition, current methods commonly train a singular NeRF model for all frames in the video, and the model size stays the same despite using chunks with varying numbers of frames, making it less suitable for video streaming applications. These results underscore the importance of further optimizations to make NeRF-based volumetric video streaming practical.

2 Background

Traditional Volumetric Content Representations mainly utilize geometry structures such as point clouds [19, 29], meshes [10, 72], and voxels [12, 42]. Point clouds are effective for non-manifold structures [11], yet their absence of spatial connectivity [50, 58] may cause holes. Meshes excel in offering surface detail and efficient rendering via rasterization [5, 27]. However, their reliance on a fixed topology [32, 41] hampers modeling topological changes. Furthermore, they often struggle to model occlusions and optical effects [10, 63], restricting their potential for generating photo-realistic 3D models. Comparatively, voxels surpass point clouds with their regular structure and editing efficiency [68], and provide internal features and facilitate volumetric operations for topological flexibility [17, 36] compared to meshes. However, typical voxelization strategies, which map voxels to occupancy fields [36] or signed distance functions [17], still demand significant memory, confining their application to simple geometric shapes [38, 43].

Neural Radiance Fields. Beyond traditional geometric representations of volumetric content, the advent of neural networks has introduced more innovative methods. With MLP models, NeRF [38] leverages the plenoptic function [1] to construct an implicit, continuous representation of a volumetric scene. For rendering, NeRF utilizes a differentiable version of ray marching [22] that involves querying the neural network at multiple positions along each camera ray to generate color and density values. The inherent differentiability of this approach facilitates the optimization of scene representation, effectively narrowing the gap between 2D image pixels and 3D properties of the scene [21, 33]. Hence, NeRF serves as an efficient method [35, 59, 71] for synthesizing novel views from 2D images, effectively capturing the dynamic interplay of light and color within the 3D space.

Learning-based Immersive Content Representations. In addition to NeRF, there are several other learning-based approaches to represent immersive content [4, 31, 32]. Neural volume [31] conducts volume rendering for view synthesis, similar to NeRF. It utilizes an encoder-decoder network architecture wherein the decoder generates a volume containing RGB and opacity values. MVP [32] is a follow-up of neural volume [31]. It combines the neural volume and traditional

3D mesh to represent volumetric content, enabling practitioners to strike a balance between rendering quality and latency. Despite these viable alternatives, NeRF, with its relatively easy implementation and high-quality representation capabilities, is our primary focus in this paper.

Streamable NeRF. The original NeRF is designed mainly for static scenes, making it not applicable for streaming. To adapt NeRF for free-view volumetric videos, early studies either directly integrated the time dimension as an additional input to NeRF [13, 67] or employed a secondary MLP to model and learn deformations for each video frame [44, 45, 49]. However, these methods bear several limitations, including slow rendering speed [13, 67], difficulty in representing large-scale motion or dynamic events such as topological changes [13, 44, 49], and large model size alongside lengthy training periods [44, 49]. To address these challenges, recent efforts proposed several innovative solutions, for example, utilizing latent codes to represent the frames with the goal of reducing the model size and training time [25], as well as dynamically detecting foreground objects to accommodate the representation of large movements [30, 57, 66].

3 NeRF-based Volumetric Content Delivery

In this section, we start with outlining the end-to-end pipelines for streaming volumetric videos by delivering NeRF models. Subsequently, we delve into an in-depth exploration of the research challenges associated with VOD services. Finally, we pivot to discussing the distinct research challenges inherent to live video streaming.

3.1 End-to-end Pipeline

Figure 2 depicts the end-to-end pipeline of VOD and live video streaming scenarios. The setups encompass three components: the client, the client's edge server (referred to as "edge"), and the video content server (referred to as "server"). On the client side, the user wears an MR headset to watch videos to gain a truly immersive experience. Given the resource constraints of mobile headsets, the user is assisted by an edge that executes volume rendering based on a trained NeRF model. The data exchange between the client and the edge is as follows. During streaming, the client sends the headset's 6DoF pose to the edge. The edge then creates the input parameters of the NeRF model based on the received pose. After that, it performs volume rendering with NeRF's output and sends the rendered image back to the user's headset. The main variation between different setups resides in the communication between the server and the edge.

VOD Service. Current research in computer vision and graphics communities generally trains a single NeRF model for all video frames. In this setup, the server transmits the trained model to the edge prior to streaming, and during streaming, there is no data transmission between the server and the edge, as shown in Figure 2 (a). However, we argue that an ideal setup should divide the video into several chunks, each represented by a separate NeRF model, as demonstrated in Figure 2 (b). This setup is driven by the following two considerations. First, training a model for an entire video could result in a large model size, particularly for long videos. Furthermore, given that the model is not divisible (§3.3), this could prolong startup time, negatively impacting QoE [28]. Second, in real-world scenarios, users may watch only portions of a video, which can lead to inefficiencies if a single model is used for the entire video.

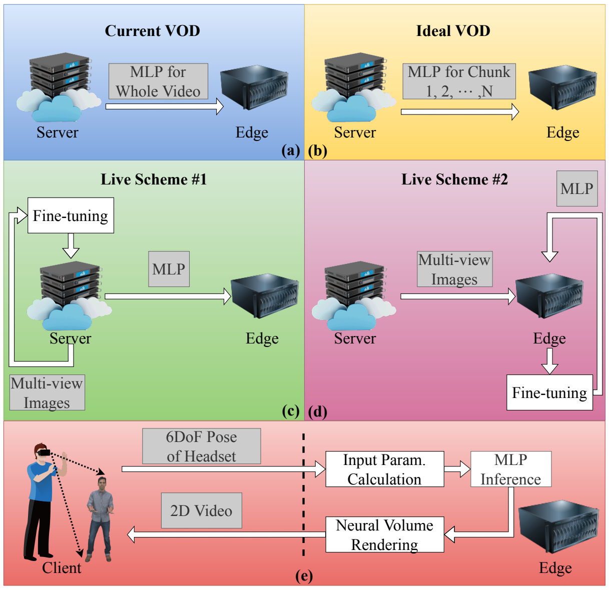

Figure 2: End-to-end pipelines of VOD and live video streaming for NeRF-based volumetric content delivery. Top: Data delivery from the server to the edge for VOD, where the server trains an MLP for the whole video (a) and trains an MLP for each chunk of the video (b), respectively. Middle: Data delivery from the server to the edge for live video streaming, where the fine-tuning is conducted on the server (c) and the edge (d), respectively. Bottom: Data exchange between the client and the edge for VOD and live video streaming (e).

Live Video Streaming. Applying NeRF for live video streaming requires continuous learning (fine-tuning) of the NeRF model since the future frame is unknown. This process can be conducted on either the edge or the server, as shown in Figures 2 (c) and (d). This choice presents a trade-off between end-to-end latency and bandwidth requirements, warranting further exploration. Fine-tuning on the server could potentially decrease training latency, as servers usually possess better computational resources. However, transmitting the fine-tuned model over the Internet may demand higher bandwidth than images, introducing additional latency if the network is congested or has limited bandwidth. Conversely, while offloading the fine-tuning task to the edge device might reduce bandwidth requirements, the computational resources at the edge might be insufficient for rapid model fine-tuning, leading to high end-to-end latency.

3.2 Real-time and High-quality Streaming

Real-time and high-quality NeRF-based volumetric video streaming demands a delicate balance between (model) transmission latency, rendering latency, and visual quality. This presents two substantial challenges as follows.

Delivery of NeRF models. A common strategy to accelerate frame rendering involves a trade-off between the storage footprint and frame rendering latency. For instance, Muller *et al.* [40] simplified the MLP model for real-time rendering by employing multi-level hash tables to encode low-dimension inputs into high-dimensional features, which preserves necessary information for high-quality rendering. This strategy reduces the rendering latency at the cost of increased storage usage, incurring the challenge of high bandwidth requirements for transmitting NeRF models (*e.g.,* 245.2 MB for a 200-frame video processed by the model of Peng *et al.* [47], as shown in §4.2) from the server to the edge. Moreover, the inherent requirement of large models for high-resolution content could further increase the bandwidth needed for model transmission.

To address the above issue, we propose to leverage advanced compression techniques to reduce the model size for storage and transmission, without significantly affecting visual quality. For instance, the vector-quantized auto-decoder compression method, proposed by Takikawa *et al.* [60] for static scenes, could be adapted to dynamic videos. However, the computation overhead associated with decompression operations is typically high [60]. Thus, it is critical to facilitate real-time decompression while minimizing the usage of computation resources, to preserve sufficient computational capacity for frame rendering on the edge.

Frame Rendering. Despite the advancements in NeRF acceleration, further improvements are needed to achieve real-time rendering of high-resolution images for dynamic scenes. For example, the state-of-the-art design by Peng *et al.* [47], while efficient for rendering 512×612 images, may fail to maintain real-time rendering of high-definition content (*e.g.,* with a resolution of 1920×1080). We propose to employ foveated rendering [3, 39] inspired by the characteristics of the human visual system (HVS) to solve this problem. The HVS features a high-resolution foveal area and a peripheral region where resolution gradually decreases [18]. Accordingly, foveated rendering reduces computation overhead without perception loss by focusing computation resources on rendering the high-resolution fovea region in detail, while reducing the rendering quality in the peripheral vision. By reducing the total number of pixels to be rendered, foveated rendering can significantly decrease the overall computational load. Our proposed foveated rendering can selectively march more rays in the foveal region and fewer in the peripheral regions.

3.3 Rate Adaptation

In NeRF-based volumetric video streaming, where the NeRF models are transmitted over the Internet, rate adaptation becomes a crucial aspect. This method, similar to traditional 2D video streaming, may require substantial bandwidth (*e.g.,* streaming at 30 FPS with a resolution of 512×612 [47] necessitates ~300 Mbps). Under dynamic network conditions, a common strategy is to adjust image resolution based on the predicted available bandwidth [20, 34]. However, in the case of NeRF-based streaming, altering the model size dynamically is impractical due to the intrinsic design of NeRF.

The original NeRF utilizes an MLP model that is not directly capable of handling different rendering resolutions through the adjustment of model size. In NeRF, all the model parameters contribute to the 3D scene reconstruction [14]. Due to the intricate interconnection of weight parameters within the model, omitting even a small portion could potentially disrupt the process and result in reconstruction failure. A straightforward extension of traditional rate adaptation schemes to NeRF-based video streaming involves training different MLP models to handle various output resolutions by modifying their depth and width. However, this approach could lead to substantial increases in memory and storage requirements, making it less than ideal.

To address this challenge, we propose to extend the scalable video coding (SVC) [55] to NeRF by exploring scalable neural networks [7, 69], such as slimmable networks [24] and progressive networks [53]. These networks are designed to be segmented into multiple executable sub-networks of varying widths and depths, each trained to accommodate a particular rendering resolution. For example, a narrower sub-network would manage low-resolution outputs, whereas a wider sub-network, with the narrower ones incorporated, would accommodate high-resolution content. By dynamically adjusting the network size, the enhanced NeRF model could support various rendering resolutions.

3.4 Viewport Adaptation

Viewport adaptation is a prevalent strategy in immersive video streaming, aiming at bandwidth reduction by delivering mainly video content that the viewer is anticipated to consume, rather than the entire scene [19, 51]. In bandwidth-hungry NeRF-based volumetric video streaming, viewport adaptation is a critical component. There are two basic requirements for its application in immersive video streaming: content segmentation for selective transmission and viewport prediction for content selection. The latter has been extensively studied in immersive video streaming, such as

360° video streaming [51] and point-cloud-based volumetric video streaming [19]. The former, however, is non-trivial for NeRF-based video streaming, for which a single MLP is typically trained to represent the whole scene. However, as illustrated in §3.3, to gain high-quality reconstruction, we need to transmit all parameters of the MLP model, undermining the bandwidth-saving benefit of viewport adaptation.

A straightforward solution is to partition the whole scene into multiple voxels and represent each voxel with an MLP [40, 52]. However, this solution may be confined only to the bounded scene [52]. Moreover, determining the optimal number of cells is non-trivial. A fine-grained segmentation strategy may incur high segmentation and storage overhead, while a limited number of cells might reduce the effectiveness of viewport adaptation. An innovative approach involves utilizing attention mechanisms [62] to assign weights to NeRF parameters. During streaming, this adaptability allows us to rank and stream high-weight parameters for the viewer’s specific viewport, optimizing bandwidth consumption.

3.5 Live Video Streaming

Different from VOD services, live volumetric video streaming enables more exciting use cases, such as telesurgery [8] and remote collaboration [61]. In this section, we outline several challenges and potential solutions related to NeRF-based live volumetric video streaming.

Real-time and Continuous Learning. Live video streaming presents the complex challenge of real-time, continuous training of NeRF models for novel view synthesis since the content of future frames remains unknown. Training NeRF models is notoriously time-consuming [2, 48], intensifying this challenge. To mitigate this issue, we propose a solution hinged on the observation that the variation of content between frames may be limited. Hence, once the initial scene model is trained, there is no need for retraining from scratch. For subsequent frames, we can fine-tune the pre-trained model by feeding features extracted from the altered content [71], potentially expediting continuous training.

Even though we still need to train the models for the initial scene, its impact on QoE is likely to be limited due to the following two reasons. First, in real-world scenarios, the initial scene is generally known before streaming commences, such as the recording studio, allowing us to pre-train the NeRF model offline. Second, recent advancements have significantly accelerated the NeRF training process [16].

Viewport Adaptation. In the context of live video streaming, we could conduct viewport adaptation on the transmitted multiple-view images. For each frame, we could potentially transmit only the content within the user’s predicted viewport, effectively reducing the number of delivered pixels. This approach could potentially decrease both the size of the transmitted images and the model. Consequently, it could alleviate the bandwidth demand for live streaming schemes, as shown in Figure 2 (c) and (d). However, a caveat to consider is that the transmitted multi-view images are used to fine-tune the MLP, which is trained on the previous frames. Given that the user’s viewport is likely to change over time, the content in the current frame might not be present in the previous ones, making it difficult to fine-tune the model for the current frame. Devising an effective strategy to overcome this challenge necessitates further investigation.

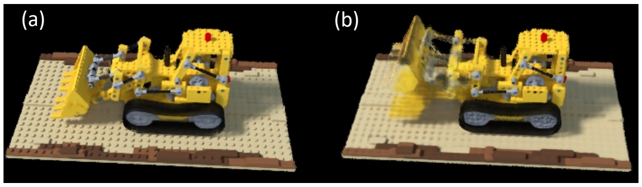

Figure 3: Qualitative comparisons of reconstructions at resolution 400 with 10 (a) and 50 (b) training input frames, respectively (a monocular synthetic dataset [49]).

4 Preliminary Results

4.1 Experiment Setup

We reproduce two state-of-the-art NeRF models, the dynamic MLP-maps [47] and Tensor4D [56], which are specialized for dynamic 3D reconstruction and rendering with dense and sparse input views, respectively. We leverage the RGB datasets associated with the aforementioned models that capture foreground dynamic scenes within bounded regions. In particular, for the dynamic MLP-maps model, we utilize the NHR [65] dataset that contains images at a resolution of 1224×1024, captured by an array of up to 80 synchronized cameras. For the Tensor4D model, we employ a synthetic dataset from D-NeRF [49] that consists of images with 800×800 resolution, captured by a single monocular camera.

To explore the trade-offs between model size, inference time, and visual quality, we experiment with two different settings. In the first setting, we train the model corresponding to different chunks by partitioning the input frames. In the second setting, we generate images at different resolutions, with the downsampling ratios set to 2 and 4, utilizing the same pre-trained models. These models are implemented in vanilla PyTorch and executed on a machine with an NVIDIA GeForce RTX 3060 GPU and an Intel Core i7-12700K CPU.

4.2 Experimental Results

Model Size. In the context of volumetric video streaming based on NeRF representations, the server transmits the pre-trained MLP model to the edge. Therefore, the size of the model crucially determines the initial startup time, influencing the QoE. In this paper, we train the vanilla MLP-maps model with 200 frames of a video in the NHR dataset, resulting in a large model size of 245.2MB. We then investigate the

Sparse-view in Tensor4D [56]		Dense-view in MLP-maps [47]	
Resolutions	Time (s)	Resolutions	Time (s)
200×200p	7	306×256p	0.02
400×400p	28	612×512p	0.07
800×800p	115	1224×1024p	0.20

Table 1: Comparison of rendering time per frame at different resolutions in sparse and dense view cases.

impact of using chunks with varying frame numbers on the model size and reconstruction quality. We train the "lego" model following the Tensor4D configuration with chunk sizes of 10 and 50 (*i.e.,* number of input frames). Despite the variation in chunk size, the model size remains consistent at 197.9MB. Figure 3 shows the reconstructed images at resolution 400 for a monocular synthetic dataset, trained with 10 and 50 input frames, respectively. This figure indicates that smaller chunk sizes yield higher quality reconstructions, featuring detailed aspects and no artifacts, despite the model size remaining the same. This suggests that a model trained with a smaller number of inputs potentially retains more detailed information without compressing as much data.

Inference Time. Table 1 shows a comparison of rendering time per frame at varying resolutions, both in sparse and dense view cases. The table illustrates that rendering high-resolution images typically requires more time, highlighting the need for a balance between video quality and rendering time. Furthermore, a noteworthy observation from the table is that the rendering time for the dense view using the MLP-maps model exhibits a significant improvement over the sparse view on the Tensor4D model. This enhancement potentially paves the way for achieving real-time rendering at the cost of requiring dense views with more cameras.

Visual Quality. Figures 4 and 5 depict reconstructed images at various resolutions and the ground truth for the monocular synthetic dataset with the Tensor4D model and the NHR dataset with the MLP-maps model. The qualitative comparison of these images reveals that, generally, as the resolution increases, the reconstructed images become clearer with enhanced details, such as the granularity of the "lego", and the defined facial features and clothing folds. Despite the models producing photo-realistic images, there are still visible discrepancies compared to the ground truth. Therefore, there is still significant room for improvement in terms of visual quality for NeRF-based volumetric content representation.

Discussion. Our preliminary results demonstrate the inherent trade-offs between model size, inference time, and visual quality in NeRF-based volumetric video streaming. We observe a considerable discrepancy in rendering time with sparse and dense view inputs, even though the perceived difference in the visual quality of the reconstructed images is small. In the case of dense-view inputs, the MLP-maps model is already capable of achieving real-time rendering. However, a prevalent issue is that current models tend to train a singular NeRF model on all frames in the video. This approach results in the NeRF model with a high storage cost, making it less suitable for video streaming applications. Therefore, additional design and optimization are required to achieve practical NeRF-based volumetric content delivery.

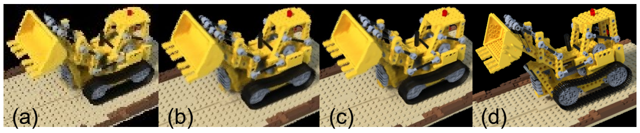

Figure 4: Qualitative comparisons of (a) – (c): reconstructions at resolutions of 200, 400, and 800, respectively. (d): the ground truth at a resolution of 800 (a monocular synthetic dataset [49]).

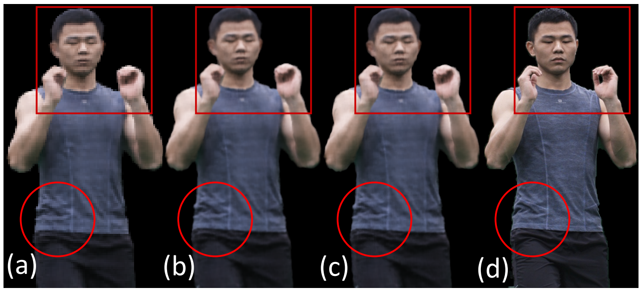

Figure 5: Qualitative comparisons of (a) – (c): reconstructions at resolutions of 256, 512, and 1024, respectively. (d): the ground truth at a resolution of 1024 (NHR dataset [65]).

5 Conclusion

In this paper, we charted an ambitious research agenda, focusing on neural-based volumetric video streaming. This approach harnesses the strengths of NeRF, aiming for photo-realistic visual quality. To reduce bandwidth consumption and ultimately enhance the QoE, we delved into the unique challenges related to NeRF-based volumetric content delivery in VOD and live video streaming services and proposed potential solutions to these problems. Our preliminary results suggest a delicate balance that needs to be maintained between model size, inference time, and visual quality. We hope that our work will inspire future research endeavors in NeRF-based volumetric video streaming, ultimately delivering immersive content with high visual quality, efficient bandwidth usage, and low end-to-end latency.

Acknowledgment

We thank the anonymous reviewers for their insightful comments. This work was supported in part by the U.S. NSF under Grants 2212296 and 2235049 and a Google Research Scholar Award.

References

[1] Edward H Adelson, James R Bergen, et al. 1991. The Plenoptic Function and the Elements of Early Vision. *Computational models of visual processing* 1, 2 (1991), 3–20.

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2022. Mip-nerf 360: Unbounded Anti-aliased Neural Radiance Fields. In *Proceedings of IEEE/CVF CVPR.*

[3] Behnam Bastani, Eric Turner, Carlin Vieri, Haomiao Jiang, Brian Funt, and Nikhil Balram. 2017. Foveated Pipeline for AR/VR Head-Mounted Displays. *Information Display* 33, 6 (2017), 14–35. https://doi.org/10.1002/j.2637-496X.2017.tb01040.x

[4] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. 2020. Immersive Light Field Video with a Layered Mesh Representation. In *Proceedings of ACM SIGGRAPH.*

[5] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. 2019. Learning to Predict 3D Objects with an Interpolation-based Differentiable Renderer. *Advances in neural information processing systems (NIPS)* (2019).

[6] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. 2023. Mobilenerf: Exploiting the Polygon Rasterization Pipeline for Efficient Neural Field Rendering on Mobile Architectures. In *Proceedings of the IEEE/CVF CVPR.*

[7] Junwoo Cho, Seungtae Nam, Daniel Rho, Jong Hwan Ko, and Eunbyung Park. 2022. Streamable Neural Fields. In *Proceedings of IEEE/CVF ECCV.*

[8] Paul J Choi, Rod J Oskouian, and R. Shane Tubbs. 2018. Telesurgery: Past, Present, and Future. *Cureus* 10, 5 (2018), e2716.

[9] Alexander Clemm, Maria Torres Vega, Hemanth Kumar Ravuri, Tim Wauters, and Filip De Turck. 2020. Toward Truly Immersive Holographic-type Communication: Challenges and Solutions. *IEEE Communications Magazine* 58, 1 (2020), 93–99. https://doi.org/10.1109/MCOM.001.1900272

[10] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-Quality Streamable Free-Viewpoint Video. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–13.

[11] Leila De Floriani, Franco Morando, and Enrico Puppo. 2003. Representation of non-manifold objects. In *Proceedings of ACM Symposium on Solid Modeling and Applications.* https://doi.org/10.1145/781606.781656

[12] Robert A Drebin, Loren Carpenter, and Pat Hanrahan. 1988. Volume Rendering. *ACM Siggraph Computer Graphics* 22, 4 (1988), 65–74.

[13] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. 2021. Neural Radiance Flow for 4D View Synthesis and Video Processing. In *Proceedings of the IEEE/CVF ICCV.*

[14] Emilien Dupont, Hyunjik Kim, SM Ali Eslami, Danilo Jimenez Rezende, and Dan Rosenbaum. 2022. From Data to Functa: Your Data Point is a Function and You Can Treat it Like One. In *Proceedings of International Conference on Machine Learning (ICML).*

[15] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance Fields without Neural Networks. In *Proceedings of the IEEE/CVF CVPR.*

[16] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. 2023. Learning Neural Volumetric Representations of Dynamic Humans in Minutes. In *Proceedings of IEEE/CVF CVPR.*

[17] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. 2020. Local Deep Implicit Functions for 3D Shape. In *Proceedings of the IEEE/CVF CVPR.*

[18] Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D graphics. *ACM Transactions on Graphics* 31, 6 (2012), 1–10. https://doi.org/10.1145/2366145.2366183

[19] Bo Han, Yu Liu, and Feng Qian. 2020. ViVo: Visibility-aware Mobile Volumetric Video Streaming. In *Proceedings of ACM MobiCom.* https://doi.org/10.1145/3372224.3380888

[20] Junchen Jiang, Vyas Sekar, and Hui Zhang. 2012. Improving Fairness, Efficiency, and Stability in HTTP-based Adaptive Video Streaming with FESTIVE. In *Proceedings of ACM CoNEXT.* https://doi.org/10.1145/2413176.2413189

[21] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. 2020. Sdfdiff: Differentiable Rendering of Signed Distance Fields for 3D Shape Optimization. In *Proceedings of the IEEE/CVF CVPR.*

[22] James T Kajiya and Brian P Von Herzen. 1984. Ray Tracing Volume Densities. *ACM SIGGRAPH computer graphics* 18, 3 (1984), 165–174.

[23] Kyungjin Lee, Juheon Yi, Youngki Lee, Sunghyun Choi, and Young Min Kim. 2020. GROOT: a real-time streaming system of high-fidelity volumetric videos. In *Proceedings of ACM MobiCom.* https://doi.org/10.1145/3372224.3419214

[24] Changlin Li, Guangrun Wang, Bing Wang, Xiaodan Liang, Zhihui Li, and Xiaojun Chang. 2021. Dynamic Slimmable Network. In *In Proceedings of IEEE/CVF CVPR.*

[25] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. 2022. Neural 3D Video Synthesis from Multi-view Video. In *Proceedings of IEEE/CVF CVPR.*

[26] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. 2020. Neural Sparse Voxel Fields. *Advances in Neural Information Processing Systems (NIPS)* 33 (2020), 15651–15663.

[27] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. 2019. Soft Rasterizer: A Differentiable Renderer for Image-based 3D Reasoning. In *Proceedings of the IEEE/CVF CVPR.*

[28] Xi Liu, Florin Dobrian, Henry Milner, Junchen Jiang, Vyas Sekar, Ion Stoica, and Hui Zhang. 2012. A Case for a Coordinated Internet Video Control Plane. In *Proceedings of ACM SIGCOMM.*

[29] Yu Liu, Bo Han, Feng Qian, Arvind Narayanan, and Zhi-Li Zhang. 2022. Vues: practical mobile volumetric video streaming through multiview transcoding. In *Proceedings of ACM MobiCom.* https://doi.org/10.1145/3495243.3517027

[30] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. 2023. Robust Dynamic Radiance Fields. In *Proceedings of the IEEE/CVF CVPR.*

[31] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Transactions on Graphics (ToG)* 38, 4 (2019), 1–14.

[32] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of Volumetric Primitives for Efficient Neural Rendering. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–13.

[33] Matthew M Loper and Michael J Black. 2014. OpenDR: An Approximate Differentiable Renderer. In *Proceedings of IEEE/CVF ECCV.*

[34] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *Proceedings of ACM SIGCOMM.* https://dl.acm.org/doi/10.1145/3098822.3098843

[35] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Proceedings of IEEE/CVF CVPR.*

[36] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *Proceedings of the IEEE/CVF CVPR.*

[37] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. 2022. KeypointNeRF: Generalizing Image-based Volumetric Avatars using Relative Spatial Encoding of Keypoints. In *Proceedings of IEEE/CVF ECCV*.

[38] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[39] Bipul Mohanto, ABM Tariqul Islam, Enrico Gobbetti, and Oliver Staadt. 2022. An integrative view of foveated rendering. *Computers & Graphics* 102 (2022), 474–501. https://doi.org/10.1016/j.cag.2021.10.010

[40] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Transactions on Graphics (ToG)* 41, 4 (2022), 1–15.

[41] Richard A Newcombe, Dieter Fox, and Steven M Seitz. 2015. Dynamicfusion: Reconstruction and Tracking of Non-rigid Scenes in Real-time. In *Proceedings of the IEEE/CVF CVPR*. 343–352.

[42] Fakir S. Nooruddin and Greg Turk. 2003. Simplification and Repair of Polygonal Models using Volumetric Techniques. *IEEE Transactions on Visualization and Computer Graphics* 9, 2 (2003), 191–205.

[43] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning Continuous Signed Distance Functions for Shape Representation. In *Proceedings of the IEEE/CVF CVPR*.

[44] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF CVPR*.

[45] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. 2021. HyperNeRF: a Higher-dimensional Representation for Topologically Varying Neural Radiance Fields. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–12.

[46] Jingliang Peng, Chang-Su Kim, and C-C Jay Kuo. 2005. Technologies for 3D Mesh Compression: A Survey. *Journal of visual communication and image representation* 16, 6 (2005), 688–733.

[47] Sida Peng, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2023. Representing Volumetric Videos as Dynamic MLP Maps. In *Proceedings of IEEE/CVF CVPR*.

[48] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *Proceedings of IEEE/CVF CVPR*.

[49] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-nerf: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of IEEE/CVF CVPR*.

[50] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proceedings of IEEE/CVF CVPR*.

[51] Feng Qian, Bo Han, Qingyang Xiao, and Vijay Gopalakrishnan. 2018. Flare: Practical Viewport-Adaptive 360-Degree Video Streaming for Mobile Devices. In *Proceedings of ACM MobiCom*.

[52] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. 2021. Kilonerf: Speeding Up Neural Radiance Fields with Thousands of Tiny Mlps. In *Proceedings of the IEEE/CVF CVPR*.

[53] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive Neural Networks. https://arxiv.org/abs/1606.04671. [accessed on 10-July-2023].

[54] Radu Bogdan Rusu and Steve Cousins. 2011. 3D is here: Point Cloud Library (PCL). In *2011 IEEE international conference on robotics and automation*. 1–4.

[55] Heiko Schwarz, Detlev Marpe, and Thomas Wiegand. 2007. Overview of the Scalable Video Coding Extension of the H.264/AVC Standard. *IEEE Transactions on Circuits and Systems for Video Technology* 17, 9 (2007), 1103–1120.

[56] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. 2023. Tensor4d: Efficient Neural 4D Decomposition for High-fidelity Dynamic Reconstruction and Rendering. In *Proceedings of the IEEE/CVF CVPR*.

[57] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. 2023. NeRFPlayer: A Streamable Dynamic Scene Representation with Decomposed Neural Radiance Fields. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2732–2742.

[58] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. 2018. Splatnet: Sparse Lattice Networks for Point Cloud Processing. In *Proceedings of the IEEE/CVF CVPR*.

[59] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. 2021. A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose. *Proceedings of Advances in Neural Information Processing Systems (NIPS)* (2021).

[60] Towaki Takikawa, Alex Evans, Jonathan Tremblay, Thomas Müller, Morgan McGuire, Alec Jacobson, and Sanja Fidler. 2022. Variable Bitrate Neural Fields. In *Proceedings of the ACM SIGGRAPH*.

[61] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence . In *Proceedings of ACM Symposium on User Interface Software and Technology (UIST)*.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*.

[63] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. 2021. Unsupervised Point Cloud Pre-training via Occlusion Completion. In *Proceedings of the IEEE/CVF ICCV*.

[64] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minve Wu, Jingyi Yu, and Lan Xu. 2022. Fourier PlenOctrees for Dynamic Radiance Field Rendering in Real-time. In *Proceedings of the IEEE/CVF CVPR*.

[65] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. 2020. Multi-view Neural Human Rendering. In *Proceedings of the IEEE/CVF CVPR*.

[66] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. 2022. D^{2}NeRF: Self-Supervised Decoupling of Dynamic and Static Objects from a Monocular Video. *Proceddings of Advances in Neural Information Processing Systems (NIPS)* (2022).

[67] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. 2021. Space-time Neural Irradiance Fields for Free-viewpoint Video. In *Proceedings of IEEE/CVF CVPR*.

[68] Yusheng Xu, Xiaohua Tong, and Uwe Stilla. 2021. Voxel-based Representation of 3D Point Clouds: Methods, Applications, and its Potential Use in the Construction Industry. *Automation in Construction* 126 (2021), 103675.

[69] Guo-Wei Yang, Wen-Yang Zhou, Hao-Yang Peng, Dun Liang, Tai-Jiang Mu, and Shi-Min Hu. 2022. Recursive-NeRF: An Efficient and Dynamically Growing NeRF. *IEEE Transactions on Visualization and Computer Graphics* (2022).

[70] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021. Plenoctrees for Real-time Rendering of Neural Radiance Fields. In *Proceedings of the IEEE/CVF ICCV*.

[71] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelNeRF: Neural Radiance Fields from One or Few Images. In *Proceedings*

of IEEE/CVF CVPR.

[72] Emin Zerman, Cagri Ozcinar, Pan Gao, and Aljosa Smolic. 2020. Textured Mesh vs Coloured Point Cloud: A Subjective Study for Volumetric Video Compression. In *Proceedings of IEEE Quality of Multimedia Experience (QoMEX).*

[73] Anlan Zhang, Chendong Wang, Bo Han, and Feng Qian. 2021. Efficient Volumetric Video Streaming Through Super Resolution. In *Proceedings of ACM HotMobile.*

[74] Anlan Zhang, Chendong Wang, Bo Han, and Feng Qian. 2022. YuZu: Neural-enhanced Volumetric Video Streaming. In *Proceedings of USENIX NSDI.* https://www.usenix.org/conference/nsdi22/presentation/zhang-anlan

[75] Ding Zhang, Bo Han, Parth Pathak, and Haoliang Wang. 2021. Innovating Multi-user Volumetric Video Streaming through Cross-layer Design. In *Proceedings of ACM HotNets.*

[76] Ding Zhang, Puqi Zhou, Bo Han, and Parth Pathak. 2022. M5: Facilitating Multi-User Volumetric Content Delivery with Multi-Lobe Multicast over mmWave. In *Proceedings of ACM SenSys.* https://doi.org/10.1145/3560905.3568540