Foveated Neural Radiance Fields for Real-Time and Egocentric Virtual Reality

NIANCHEN DENG, Shanghai Jiao Tong University
ZHENYI HE, New York University
JIANNAN YE, Shanghai Jiao Tong University
PRANEETH CHAKRAVARTHULA, University of North Carolina at Chapel Hill
XUBO YANG, Shanghai Jiao Tong University
QI SUN, New York University

Fig. 1. Illustration of our gaze-contingent neural synthesis method. (a) Visualization of our active-viewing-tailored egocentric neural scene representation. (b) Our method allows for extremely low data storage of high-quality 3D scene assets and achieves high perceptual image quality and low latency for interactive virtual reality. Here, we compare our method with the state-of-the-art alternative neural synthesis [Mildenhall et al. 2020] and foveated rendering solutions [Patney et al. 2016; Perry and Geisler 2002] and demonstrate that our method significantly reduces more than 99% time consumption (from 9s to 20ms) and data storage (from 100MB mesh to 0.5MB neural model) for first-person immersive viewing. The orange circle indicates the viewer’s gaze position. We also show zoomed-in views of the foveal images generated by various methods in (c) and show that our method is superior compared to existing methods in achieving high perceptual image fidelity.

Traditional high-quality 3D graphics requires large volumes of fine-detailed scene data for rendering. This demand compromises computational efficiency and local storage resources. Specifically, it becomes more concerning for future wearable and portable virtual and augmented reality (VR/AR) displays. Recent approaches to combat this problem include remote rendering/streaming and neural representations of 3D assets. These approaches have redefined the traditional local storage-rendering pipeline by distributed computing or compression of large data. However, these methods typically suffer from high latency or low quality for practical visualization of large immersive virtual scenes, notably with extra high resolution and refresh rate requirements for VR applications such as gaming and design.

Tailored for the future portable, low-storage, and energy-efficient VR platforms, we present the first gaze-contingent 3D neural representation and view synthesis method. We incorporate the human psychophysics of visual- and stereo-acuity into an egocentric neural representation of 3D scenery. Furthermore, we jointly optimize the latency/performance and visual quality, while mutually bridging human perception and neural scene synthesis, to achieve perceptually high-quality immersive interaction. Both objective analysis and subjective study demonstrate the effectiveness of our approach in significantly reducing local storage volume and synthesis latency (up to 99% reduction in both data size and computational time), while simultaneously presenting high-fidelity rendering, with perceptual quality identical to that of fully locally stored and rendered high-quality imagery.

1 INTRODUCTION

Efficient storage is a core challenge faced by today’s computational systems, especially due to the fast-growing high-resolution acquisition/display devices, the demand for 3D data quality, and machine learning. Thus, streaming services and remote cloud-based rendering (such as Google Stadia and Nvidia Shield) have rapidly emerged in recent years. These storage-friendly systems have shown an unprecedented potential to reshape the traditional computer graphics pipeline by completely relieving the end user’s computational workload and storage pressure, thus advancing device portability. Future
spatial and immersive computing platforms such as VR and AR are potentially major beneficiaries due to the extensive demands of portability and energy-efficiency\(^1\).

However, existing low-storage services have only been successfully deployed in relatively latency-insensitive platforms such as TV-based gaming consoles due to the inevitable latency from transmission (more than 100 ms\(^2\)). In comparison, VR and AR require remarkably low latency and significantly higher fidelity/speed for a sickness-free and natural experience. As a consequence, there is a demand for massive high-quality spatial-temporal data. Unfortunately, no solution exists today to deliver high-quality and dynamic/interactive immersive content at low latency and with low local storage.

Recent years have witnessed a surge in neural representations as an alternative to compact 3D representations such as point clouds and polygonal meshes, enabling compact assets. Application of neural representations includes voxelization [Sitzmann et al. 2019a], distance fields [Mildenhall et al. 2020; Park et al. 2019; Sitzmann et al. 2020], and many more. However, existing solutions suffer from either high time consumption or low image fidelity with large scenes.

Acceleration without compromising perceptual quality is also an ultimate goal of real-time rendering. Gaze-contingent rendering approaches have shown remarkable effectiveness in presenting imagery of perceptual quality identical to full resolution rendering [Guentner et al. 2012; Patney et al. 2016; Tursun et al. 2019]. This is achieved by harnessing high field-of-view(FoV) head-mounted displays and high precision eye-tracking technologies. However, existing foveated rendering approaches still require locally stored full 3D assets, resulting in long transmission time and storage volume, whose availability is unfortunately limited in wearable devices.

To overcome these open challenges mentioned above, we present the first gaze-contingent and interactive neural synthesis for high-quality 3D scenes. Aiming at the next-generation wearable, low-latency, and low-storage VR/AR platforms, our compact neural model predicts stereo frames given users’ real-time head and gaze motions. The predicted frames are perceptually identical to the traditional real-time rendering of high-quality scenes. With wide FoV and high-resolution eye-tracked VR displays, our method significantly reduces 3D data storage (from 100MB mesh to 0.5MB neural model) compared with traditional rendering pipelines; it also remarkably improves responsiveness (from 9s to 20 ms) compared with alternative view neural synthesis methods. That is, we enable instant immersive viewing without long loading/transmission waits while preserving visual fidelity. Tailored to future lightweight, power and memory-efficient wearable display systems, expected to be active all-day long akin to today’s everyday-use mobile phones, our method achieves minimal local storage for high-quality 3D scenes.

To achieve immersive viewing with low storage, high visual quality and low systematic latency, we introduce spatial-angular human psychophysical characteristics to neural representation and synthesis. Specifically, we represent 3D scenes with concentric spherical coordinates. Our representation also allows for depicting the foveated color sensitivity and stereopsis during both the training and inference phases. Furthermore, we derive an analytical spatial-temporal perception model to optimize our neural scene representation towards imperceptible loss in image quality and latency.

We validate our approach by conducting numerical analysis and user evaluation on commercially available AR/VR display devices. The series of experiments reveal our method’s effectiveness in delivering an interactive VR viewing experience, perceptually identical to an arbitrarily high quality rendered 3D scene, with significantly lowered data storage. We will open source our implementation and dataset. We make the following major contributions:

- A low-latency, interactive, and significantly low-storage immersive application. It also offers full perceptual quality with instant high resolution and high FoV first-person VR viewing.
- A scene parameterization and vision-matched neural representation/synthesis method considering both visual- and stereo-acuity.
- A spatiotemporally analytical model for jointly optimizing systematic latency and perceptual quality.

2 RELATED WORK

2.1 Image-based View Synthesis

Image-based rendering (IBR) has been proposed in computer graphics to complement the traditional 3D assets and ray propagation pipelines [Gortler et al. 1996; Levoy and Hanrahan 1996; Shum and Kang 2000]. Benefiting from the uniform 2D representation, it delivers pixel-wise high quality without compromising performance. However, one major limitation of IBR is the demand for large and dense data acquisition, resulting in limited and sparse view ranges. To address this problem, recent years have witnessed extensive research on synthesizing novel image-based views instead of fully capturing or storing them. Examples include synthesizing light fields [Kalantari et al. 2016; Li and Khademi Kalantari 2020; Mildenhall et al. 2019] and multi-layer volumetric videos [Broxton et al. 2020]. However, despite being more compact, the synthesis is usually in a locally interpolated fashion than globally applicable with highly dynamic VR camera motions.

2.2 Implicit Scene Representation for Neural Rendering

To fully represent a 3D object and environment without large explicit storage, neural representations have drawn extensive attention in computer graphics and computer vision communities. With a deep neural network that depicts a 3D world as an implicit function, the neural networks may directly approximate an object’s appearance given a camera pose [Mildenhall et al. 2020; Park et al. 2017; Sitzmann et al. 2020b, 2019a,b]. Prior arts also investigated implicitly representing shapes [Park et al. 2019] and surfaces [Mescheder et al. 2019] to improve the visual quality of 3D objects. Inspired by [Park et al. 2019], signed distance functions have been deployed for representing raw data such as point cloud [Atzmon and Lipman 2020; Gropp et al. 2020] with better efficiency [Chabra et al. 2020]. The input information ranged from 2D images [Choi et al. 2019; Lin et al. 2020a; Yariv et al. 2020], 3D dataset [Oechsle et al. 2019; Saito et al. 2019; Zhang et al. 2020], time-varying vector field [Niemeyer et al. 2019] to local features [Liu et al. 2020; Tretschk et al. 2020].

---

\(^1\)https://www.vrexchange.com/blog/virtual-reality-data-storage-concern

\(^2\)https://www.pcgamer.com/heres-how-stadias-input-lag-compares-to-native-pc-gaming/
Current implicit representations primarily focus on locally "outside-in" viewing of individual objects, with low field-of-view, speed, and resolution. For large scenes, the lack of coverage may cause the quality to drop. However, in virtual reality, the cameras are typically first-person and highly dynamic, requiring low latency, high resolution, and 6DoF coverage.

2.3 Gaze-Contingent Rendering

Both rendering and neural synthesis suffer from the heavy computational load, thus systematic latency and lags. With a full mesh representation, foveated rendering has been proposed to accelerate local rendering performance, including mesh- [Guenter et al. 2012; Patney et al. 2016] or image-based [Kaplanyan et al. 2019; Sun et al. 2017] methods. The accelerations are typically achieved through high field-of-view displays (such as VR head-mounted displays) and eye-tracking technologies. However, these all require full access to the original 3D assets or per-frame images. Consequently, the systems usually require large local storage, long transmission waits, or constant frame-based transmissions with latency. Our method bridges human vision and deep synthesis. It accelerates neural synthesis beyond local rendering, enabling both computational and storage efficiency. It is also suitable for future cloud-based virtual reality with instant full-quality viewing and interaction.

2.4 Panorama-Based 6DoF Immersive Viewing

Panorama images and videos much fit VR applications thanks to their 360 FoV coverage. However, the main challenge is their single projection center in each frame, limiting free-form camera translation, thus 6DoF immersive viewing experience. Recently, extensive research has been proposed to address this problem. Serrano et al. [2019] presented a depth-based dis-occlusion method that dynamically reprojects to 6DoF cameras, enabling natural viewing with motion parallax. Pozo et al. [2019] jointly optimizes capture and viewing processes. Machine learning approaches have advanced the robustness of various geometric and lighting conditions [Attal et al. 2020a,b; Lin et al. 2020b]. As a spherical image, panoramas are designed for capturing physical worlds with flexible reprojections to displays. The 6DoF viewing typically suffers from trade-offs among performance, achievable resolution, and dis-occlusion effects due to lack of knowledge of the 3D space. With an orthogonal mission, our model represents and reproduces arbitrary 3D virtual worlds for low storage VR. It lowers memory usage by replacing traditional mesh- or volume-based data than predicting image-space occlusions.

3 METHOD

Based on a concentric spherical representation and trained network predicting RGBDs on the spheres (Section 3.1), our system predominantly comprises two main steps in real-time: visual-/stereocuity adaptive elemental images synthesis with ray marching for fovea, mid-, and far-peripheries; followed by image-based rendering to composite displayed frames (Section 3.2). For desired precision-performance balance, we further craft an analytical spatial-temporal model to optimize the determination of our intra-system variables, including coordinates and neural-network parameters (Section 3.3).

3.1 Egocentric Neural Representation

The recent single-object-oriented "outside-in" view synthesis methods [Mildenhall et al. 2020; Sitzmann et al. 2019a] typically represent the training targets using uniform voxelization. However, immersive VR environments introduce open challenges for such parameterization due to the rapid variation of the egocentric (first-person) viewing (e.g., Figure 1a). As a consequence, the neural representation on large virtual environments typically suffer from ghosting artifacts, low resolution, or slow speed (e.g., the white cube and red wall are blurred in Figure 1b and Figure 1c).

Egocentric coordinates. To tackle this problem, we are inspired by the recent panorama dis-occlusion methods [Attal et al. 2020b; Broxton et al. 2020; Lin et al. 2020b] to depict the rapidly varying first-person views with concentric spherical coordinates. This representation enables robust rendering at real-time rates and allows for 6DoF interaction to navigate complex immersive environments. As visualized in Figure 2, our representation is parameterized with the number of concentric spheres (N) and their respective radii (r = {ri}, i ∈ [1, N]). Our network aims at predicting the function C(q) = (r, g, b, d) of any position q(r,θ,φ) on these N spherical surfaces. Here, (r, g, b) and d are the predicted color and density, respectively. We then employ a ray marching through this intermediate function to produce images of specific views.

Training and synthesis. For each scene, we approximate the function C by a multilayer perceptron (MLP) network, similar to [Mildenhall et al. 2020]. The network is composed of Nm fully-connected layers of Nc channels, each using ReLU as the activation function. We further determine Nm, Nc and N via a spatial-temporal optimization in Section 3.3. Given rays of pixels in an image, the network predicts the colors and densities along these rays on every sphere. The ray marching weighted integrates predictions in back-to-front order to obtain actual colors of corresponding pixels. Similar to [Mildenhall et al. 2020], we apply a trigonometric input encoding and compute the L2-distance between the volume-rendered and mesh-rendered images as the training loss function. We discuss the specifics of training data creation and implementations in Section 4.
3.2 Gaze-Aware Synthesis

Our concentric spherical coordinate, as described in Section 3.1, addresses the view variance problem in VR. However, it still suffers from significant rendering latency (several seconds per frame). This is one of the core challenges causing neural representation to be unsuitable for VR. Instead of the typical single image prediction, we synthesize multiple elemental images to enable real-time responsiveness. The elemental images are generated based on the viewer’s head and gaze motions and are adapted to the retinal acuity in resolution and stereo.

3.2.1 Adaptive visual acuity. The human vision is foveated. Inspired by this, we significantly accelerate the computation by integrating the characteristic of spatially-adaptive visual acuity without compromising perceptual quality.

Specifically, given device-tracked camera position (x), direction (R), and gaze position (v_g), we train two orthogonal networks (i.e., function Cs) to synthesize foveal (I_f(x, R, v_g), 0 – 20 deg), and peripheral elemental images (i.e., two Cs to deploy for ray marching) for each eye. The periphery includes mid- (I_m(x, R, v_g), 0 – 45 deg) and high- eccentricity visual fields (I_p(x, R), 0 – 110 deg), see Figure 3. Note that I_p is independent from the gaze direction v_g. Specifically, to incorporate the display capabilities and perceptual effects, we define I_f/I_m/I_p as of 125^2/256^2/230 × 256 resolutions respectively. That is, the I_f has the highest spatial resolution of 6.4 pixels per degree (PPD), higher than those of I_m (5.7 PPD) and I_p (2.33 PPD).

3.2.2 Adaptive stereo-acuity. Head-mounted VR displays require stereo rendering, thus the elemental images I_f^{[l,r]}, I_m^{[l,r]}, and I_p^{[l,r]} for each (left and right) eye. This, however, doubles the inference computation cost that is critical for latency- and frame-rate-sensitive VR experience (please refer to Section 5.3 for breakdown comparisons).

To further accelerate stereo rendering towards real-time experience, we accommodate the inference process with the adaptive stereo-acuity in the periphery. In fact, besides the visual acuity, psychophysical studies have also revealed human’s significantly declined stereopsis while receding from the gaze point [Mochizuki et al. 2012]. Inspired by this characteristic, we perform the computation with I_f^{[l,r]}, I_m^{[l,r]}, and I_p^{[l,r]} instead of inferring 6 elemental images, where c indicates the view at the central eye (the midpoint of the left and right eyes). Figure 4 visualizes the stereopsis changes from the adaptation using an anaglyph.

3.2.3 Real-time frame composition. With the obtained elemental images as input, an image-based rendering in the fragment shader is then executed to generate final frames for each eye (Figure 3d). The output frames are displayed on the stereo VR HMDs. Two adjacent layers are blended using a smooth-step function across 40% of the inner layer. This enhances visual consistency on the edges between layers [Gunter et al. 2012]. To accommodate the mono-view I_m and I_p, they are shifted towards each eye according to approximated foveal depth range. Lastly, we enhance the contrast following the
mechanism of [Patney et al. 2016] to further preserve peripheral elemental images’ visual fidelity due to its low PDD. The details are visualized in Figure 5.

3.3 Latency-Quality Joint Optimization
As a view synthesis system based on sparse egocentric representation (the $N$ spheres) and neural networks (the $N_{m}$, $N_{c}$), the method inevitably introduces approximation errors. On the other hand, these variables also determine the online computational time that involves inferring function $C$ and ray marching the $N$ spheres. However, VR strictly demands both quality and performance. We present a spatial-temporal model that analytically depicts the correlations and optimizes the variables for an ideal user experience.

**Precision loss of a 3D scene.** As shown in Figure 2, under the egocentric representation, a 3D point $q$ is re-projected as the nearest point on a sphere that connects it to the origin point:

$$ q'(N, r, q) = r_{k} \frac{q}{||q||}, k = \arg\min_{j \in [1, N]} ||q|| - r_{j} ||. \tag{1} $$

Similar to volume-based representation, the multi-spherical system is also defined in the discrete domain. The sparsity thus naturally introduces approximation error that compromises the synthesis quality. To analytically model the precision loss, we investigate the geometric relationship among the camera, the scene, and the representation. As illustrated in Figures 1a and 2, for a sphere (located at origin point) with radius $r$, its intersection (if exists) with a directional ray $\{x, v\}$ is

$$ p(r, x, v) = x + \left( (x \cdot v)^2 - ||x||^2 + r^2 \right)^{\frac{1}{2}} - x \cdot v \right) v, \tag{2} $$

where $x$ and $v$ are the ray’s origin point and normalized direction, respectively. Inversely, given a view point $x$ observing a spatial location $q$, the ray connecting them has the direction $v(q, x) = \frac{q - x}{||q - x||}$. This ray may intersect with more than one sphere. Among them, the closest one to $q$ is:

$$ \hat{q}(x, N, r, q) = p(r_{k}, x, v(q, x)), k = \arg\min_{j \in [1, N]} ||q|| - r_{j} ||. \tag{3} $$

In the 3D space, the offset distance $||q' - q||$ indicates the precision loss at $q$ from the representation. By integrating over all views and scene points, we obtain:

$$ E_{scene}(N, r) = \int \int ||q'(N, r, q) - \hat{q}(x, N, r, q)|| dqdx, \tag{4} $$

$$ \forall (x, q) \text{ pair without occlusion in between}. $$

By integrating all 3D vertices $q$ and camera positions $x$ in our dataset sampling, $E_{scene}$ depicts how the generic representation precision of a scene, given a coordinate system defined by $N$ and $r$.

In comparison, a neural representation aims at predicting projected image given a $x$ and $R$. Thus, we further extend Equation (4) to image space to analyze the error given a set of camera’s projection matrix $M(x, R)$ as

$$ E_{image}(N, r, x, R) = \int \int ||M(x, R) \cdot (q - \hat{q}(x, N, r, q)|| dqdx. \tag{5} $$

From Figure 2 and eq. (5), we observe: given a fixed min/max range of $r$, $N$ is negatively correlated to $E_{image}$; with a fixed $N$, the correlation between distribution of $r_{j}$ and scene content (i.e., distribution of $q$) also determines $E_{image}$.

However, for neural scene representation, infinitely enlarging $N$ would significantly raise the challenges in training precision and inference performance. Likewise, increasing representation densities (i.e., higher $N$) and/or network complexities (i.e., higher $N_{m}$, $N_{c}$) naturally improves the image output quality (lower Equation (5)). However, this significantly increases the computation during ray marching, causing quality drop stretched along time. In the
performance-sensitive VR scenario, the latency breaks the continuous viewing experience and may cause simulator sickness. Thus, with content-aware optimization, we further optimize the system towards an ideal quality-speed balance.

**Spatial-temporal modeling.** Inspired by [Albert et al. 2017; Li et al. 2020], we perform spatial-temporal joint modeling to determine the optimal coordinate system \((N, N_m, N_c)\) for positional precision and network complexity \((N_m, N_c)\) for color precision that adapt to individual computational resources and scene content. This is achieved via latency-precision modeling in the spatial-temporal domain:

\[
E(N, N_m, N_c) = \sum_t \int C_{N_m, N_c}(q) \times \| M(x_t, v_t) \cdot q - M(x_{t-1}, v_{t-1}) \cdot q(r_k, x_{t-1}, v_{t-1}) \| dq,
\]

where \(l \triangleq l(N, N_m, N_c)\) is the system latency with a given coordinate and network setting. \(C_{N_m, N_c}(q)\) is the four \((r, g, b, a)\) output channels’ L1-distance between a given network setting and the highest values \(N = 64, N_m = 8, N_c = 256\). For simplicity, we assumed uniformly distributed \(r\) with a fixed range of the spherical coverage.

As suggested by Albert et al. [2017], the latency for a foveated system shall reach below ~50ms for undetectable artifacts. Given our test device’s eye-tracking latency ~12ms and the photon submission latency ~14ms ([Albert et al. 2017]), the synthesis and rendering latency shall be less than \(L_0 = 24ms\). Thus, we determine the optimal \((N, r)\) to balance latency and precision as

\[
\arg\min_{N, N_m, N_c} E(N, N_m, N_c),
\]

\[
s.t. \, l(N, r) < L_0.
\]

Figure 6 visualizes an example of the optimization mechanism for an foveal image \(I_f\). The optimized results \((N, N_m, N_c)\) for individual networks are used for training. The optimization outcomes are detailed in Section 4. The visual quality is validated by psychophysical study (Section 5.1) and objective analysis (Section 5.2). The latency breakdown of our system is reported in Section 5.3.

**4 IMPLEMENTATION**

In this section, we elaborate on how we collect data for training, the specific parameters we choose for the two orthogonal neural networks, and the software/hardware environment we use.

**Data generation.** For each scene, we generate two datasets separately for training foveal and periphery networks. The foveal dataset is composed of 256 × 256 images rendered at different views with 20 deg field-of-view, while the periphery dataset uses 45 deg field-of-view to rendering. The view translations \(x\) are sampled every 5 cm on each axis. The rotations \(R\) are sampled so that adjacent samples are seamlessly connected with no overlap.

**Optimized network parameters.** Guided by the latency-quality joint optimization described in Section 3.3, we implemented the foveal network with \(N_m = 4, N_c = 128\) and \(N = 32\). For the periphery neural network, the values are \(N_m = 4, N_c = 96\) and \(N = 16\). These optimized parameters achieve a proper balance between quality and latency, as shown in Figure 8 and Table 1.

**Ray marching bounding.** We bound the ray marching within region that \(r\) are sampled in, thus has impact on \(E_{\text{scene}}\) in Equation (4). As shown in Figure 7a, results with infinite upper bounding \((r \in [r_{\text{near}}, +\infty))\) incur blurry and ghosting artifacts due to larger \(E_{\text{scene}}\). In contrast, limited bounding significantly improves the quality (Figure 7b).

Environment. The system was implemented through OpenGL framework using CUDA and accelerated by TensorRT with Intel(R) Xeon(R) Gold 6230R CPU @ 2.10GHz (256GB RAM) and one NVIDIA GTX 3090 graphics card. For each scene, we trained around 200 epochs for experiment use (about a day).
5 EVALUATION

With a variety of scenes as shown in Figure 8, we conduct a subjective study (Section 5.1), and calculate objective measurements (Section 5.2) to evaluate the perceptual quality. Further, we analyze the intra- and inter-system performance in Section 5.3.

5.1 User Study

We conducted a psychophysical experiment to investigate how users perceive our solution (OURS) compared with existing neural view synthesis methods ([Mildenhall et al. 2020], NeRF) and foveation rendering of the original mesh scene ([Perry and Geisler 2002], F-GT).

Stimuli. For precise comparison and accommodating the low frame rate of alternative solutions (as to be compared in Section 5.3), each group of stimuli consisted of static stereo images rendered via F-GT or synthesized via OURS/NeRF with the same views, as shown in Figure 8. They were generated with the same and randomly defined gaze fixation across conditions. The resolution of the image per eye was 1440 × 1600. During the study, the target gaze position was indicated as a green cross on the stimuli images. We used the gas and minecraft scenes for the diverse indoor/outdoor environment. Each condition from an individual scene consists of 3 different gaze positions.

The foveal kernel size in F-GT was designed to match our display size and eye-display distances. For generating NeRF condition, we retrained the model from [Mildenhall et al. 2020] on our dataset with $N_m = 4$, $N_c = 128$, and $N = 16$. The bottom right of each scene figure Figure 8 shows the resulting NeRF images.
Setup. Each participant wore an eye-tracked HTC Vive Pro Eye headset and remained seated to examine the stimuli during the experiment. Ten users participated in and completed the study (3 females, $M = 23.7$, $SD = 1.49$). No participants were aware of the research, the experimental hypothesis, nor the number of methods. All participants had a normal or corrected-to-normal vision.

Task. The task was a two-alternative-forced-choice (2AFC). Each trial consists of a pair of stimuli generated from two of the three methods (OURS / NeRF / F-GT) with a sampled view and gaze position. Each stimulus appeared for 300ms on display. A forced 1.5sec break (black screen) was introduced between conditions to clear the vision. During the study, the participants were instructed to fix their gazes on the green cross. To prevent the fixation from shifting away and ensure accuracy, we tracked the users’ gaze throughout the experiment. Whenever the gaze is more than 5 deg away from the target, a trial was dropped immediately with a black screen informing the participant. After each trial, the participants were instructed to select which of the two stimuli appeared with higher visual quality using a keyboard. Before each experiment, a warm-up session with 6 trials was provided to familiarize the participants with the study procedure. The orders of conditions among trials were randomized and counter-balanced. The entire experiment of each user consists of 36 trials, 6 trials per pair of ordered conditions. To minimize the effect of accumulated fatigue, we also enforced breaks between trials (at least 2 seconds) and after each scene (60 seconds). Meanwhile, the participants were allowed to take as much time as needed.

Results. Figure 9 plots the results considering both methods and their orders in the 2AFC experiments. Here we analyzed and reported the results regardless of the orders. Among all three conditions, we observed close-to-random-guess among trials that compared F-GT and OURS (53.3% voted for OURS, $SD = 14.9$%). Meanwhile, a significantly higher ratio of voting F-GT over NeRF was observed (92.5%, voted for F-GT, $SD = 11.4$%, binomial test showed $p < 0.005^{***}$). The preference applies to OURS vs. NeRF as well (91.7% voted for OURS, $SD = 14.8$, binomial test showed $p < 0.005^{***}$).

Discussion. The close-to-random-guess indicated the statistically similar perceptual quality between OURS and foveated ground truth (F-GT). Given the perceptual identity between foveated and full-resolution rendering, the subjective study reveals that OURS can achieve similar quality as a rendering with locally fully stored high-quality 3D data.

Meanwhile, both of the two conditions showed significant quality preference than NeRF. That is, with immersive, high FoV, and first-person viewing constraints, OURS synthesizes gaze-contingent retinal images with superior perceptual quality than alternative synthesis solutions. The latter typically aims to reconstruct the full image over the entire field-of-view. To further validate the perceptual quality at individual eccentricities across the whole visual field, we analyze with objective metrics in the following section.

5.2 Visual Quality

Complementary to the subjective measurement (Section 5.1), we evaluate the perceived quality in an objective fashion. Specifically, under all four different scenes (indoor/outdoor, artificial/realistic), we compare OURS, F-GT, and NeRF at individual eccentricity ranges (up to 110 deg, the capability of the VR HMD) via partial images. For each eccentricity range, we compare the deep perceptual similarity (LPIPS) [Zhang et al. 2018] across all scenes. LPIPS uses deep neural networks to estimate perceptual similarities between the image provided and a reference image. Smaller values indicate higher perceptual similarity. For each scene, we sample 20 views with gazes at the middle of the display, resulting in 20 data per eccentricity value, 21 eccentricity value per scene (5 deg step size), thus 420 data per scene. We used one-way repeated measures ANOVAs to compare effects across three stimuli on each eccentricity range and together. Paired t-tests with Holm correction were used for all pairwise comparisons between stimuli. All tests for significance were made at the $\alpha = 0.05$ level. The error bars in the graphs show the 95% confidence intervals of the means.

Results. Figure 10 plots LPIPS values across all scenes and eccentricity ranges (5 deg step size). From foveal to near eccentricity ($\leq 40$ deg), we observed significant effects of the stimuli on LPIPS with a "large" effect size ($\eta^2 = 0.26$). That is, OURS shows significantly lower LPIPS than NeRF ($p < .001^{***}$), although higher than F-GT ($p < .001^{***}$). For example, the main effects of stimuli ($F(2,38) = 210.76$, $p < .001^{***}$) was significant on eccentricity $= [0, 25]$ deg in scene gallery (Figure 10c). OURS was significantly lower than NeRF ($t(19) = -19.54$, $p < .001^{***}$) and F-GT ($t(19) = -4.32$, $p < .001^{***}$) both with a "large" effect size (Cohen’s $d > 0.8$). The example observation generally applies to all 4 scenes being validated.

From near- to far- eccentricity ($> 40$ deg), we observed significant effects of the stimuli on LPIPS with a "large" effect size ($\omega^2 = 0.16$). OURS shows significantly higher LPIPS than NeRF. Whereas, comparing with F-GT, we observed significant lower scores ($p < .001^{***}$). For instance, the main effects of stimuli was significant on eccentricity $= 60$ deg in scene bedroom ($F(2,38) = 411.93$, $p < .001^{***}$, Figure 10d). OURS was significantly lower than F-GT ($t(19) = -3.78$, $p < .001^{***}$), and higher than NeRF ($t(19) = 22.75$, $p < .001^{***}$).
Fig. 10. LPIPS visualization of all scenes and comparison of OURS, NeRF, and F-GT over the visual field. X-axis indicates the eccentricity range (0 to 110). Y-axis shows the LPIPS loss [Zhang et al. 2018] with a full resolution rendering as reference. Lower values mean more similar perceptual quality to the reference. The semi-transparent layers indicate first and third quarterlies. The inset figures’ corresponding eccentricity ranges are indicated by the dashed lines.

Discussion. In the foveal and near-periphery, the observation revealed our method’s significantly higher visual quality than alternative neural representation solutions. This is evidenced by the significantly stronger perceptual similarity to F-GT by comparison between OURS and NeRF. The subjective study (Section 5.1) evidenced the marginally lower similarity than F-GT as statistically unnoticeable among users.

In the far-periphery, OURS showed increased LPIPS than F-GT beyond 40 deg. The latter has been shown to display identical perceptual similarity than full resolution rendering in VR [Patney et al. 2016]. Thus, the findings show that OURS doesn’t compromise the peripheral vision’s quality with its significantly enhanced synthesis acuity in the fovea.

Note that the discoveries also agree with the observations from Section 5.1. That is, in addition to the significantly faster performance (99.8% time reduction per frame as in Section 5.3), our method showed superior perceptual quality than NeRF under first-person, high resolution, and immersive viewing of 3D scenes. Meanwhile, our synthesized views are perceptually identical to the traditionally rendered images with a complete 3D mesh. It yet achieves significant data storage-saving with only about 0.5MB neural model (up to 99.5% saving from 3D mesh in our experiment).

5.3 Performance
Virtual and augmented reality demands high frame rates along with high quality to ensure an immersive and comfortable experience. Our neural synthesis method introduces both spatial and angular (stereo) visual perception for real-time performance (Section 3.1). Further, our precision-latency joint optimization for the representation and network also balances quality and performance. Here, we evaluate the performance of each component in our system and compare with existing neural synthesis solutions.

For high resolution (1440 × 1600), high field-of-view (110 deg) stereo images required by VR display, our system completes all computation (including gaze-contingent neural-inference and elemental images composition) in 31.8ms per frame without stereo foveation, as shown in Table 1. Compared with NeRF (the bottom row of
Table 1), our spatial foveation significantly improves the system performance from offline computation (in seconds) to interactive speed (about 30FPS). Our stereo foveation further reduces the computation time and achieves a high performance of more than 50 frames per second, contributing to a temporally continuous viewing experience without loss of perceived resolution and quality.

|                  | OUR method without stereo foveation | OUR full method | NeRF [Mildenhall et al. 2020] |
|------------------|------------------------------------|-----------------|-------------------------------|
| foveal infer (per eye) | 3.8                                |                 | $9 \times 10^3$              |
| periphery infer (per eye) | 11.9                               |                 |                               |
| blending & contrast enhancement | 0.2                                |                 |                               |
| **Time consumption breakdown and comparison.** The numbers show the average time consumption of each component and the overall system per frame. All units are in millisecond. |

6 CONCLUSION

In this paper, we present a gaze-aware neural scene representation and view synthesis method tailored for future portable virtual reality viewing experiences. Specifically, we overcome the limitations of existing neural rendering approaches for high performance, resolution, and fidelity. Our network individually synthesizes foveal, mid-, and far-periphery retinal images, which are then blended to form a wide field-of-view image matching retinal acuity. This is achieved via adapting the neural network model to the foveated visual acuity and stereopsis. Comparing with traditional rendering, our method requires less than 1% data storage yet delivering identically high perceptual quality. Comparing with alternative neural view synthesis approaches, our method creates significantly faster and higher fidelity viewing for high resolution, high FoV, and stereo VR head-mounted-displays.

Limitations and future work. While our method achieves unprecedented performance compared to existing approaches, it suffers from a few limitations. Our multi-spherical representation introduces higher quality when the VR camera is within the first sphere. As shown in Figure 11, when the virtual camera translates largely within a highly occluded scene, our method still synthesizes the accurate views but with declined foveal image quality. This is due to the accumulated error of the concentric spherical representation (Equation (5)). Multiscale coordinates and networks that consider various level-of-details of the 3D space have shown their effectiveness of interpolating geometries [Winkler et al. 2010]. Developing a multiscale network that synthesizes the image from global to local level-of-details would be an interesting direction for future work, extending the applicability of our method to large-scale scenes with strong occlusions.

Currently, we sample the scene fully based on eccentricity, considering acuity and stereopsis. However, we envision that fine-grained visual sensitivity analysis, such as luminance [Tursun et al. 2019] or depth [Sun et al. 2020], would provide more insights on achieving even higher quality and/or faster performance. Also, our multispherical representation is specialized to the scene for optimal quality. Thus, our network is trained per-scene. Exploring potential means of generalized representation may further reduce storage.

For simplicity, the spatial-temporal joint optimization in Section 3.3 connects the output precision and latency to the number of the spheres (N) but not their radii r. This is due to the significant amount of training with parameter sampling. Incorporating the parameters into a single training process may significantly reduce the time consumption for the optimization. With the adaptive training process, a content-aware distribution (i.e., r) of the spheres would further improve the synthesis quality and performance.

REFERENCES

Rachel Albert, Anuj Patney, David Luebke, and Joohwan Kim. 2017. Latency requirements for foveated rendering in virtual reality. *ACM Transactions on Applied Perception (TAP)* 14, 4 (2017), 1–13.

Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. 2020a. MatryODShka: Real-time 6DoF Video View Synthesis using Multi-Sphere Images. In *European Conference on Computer Vision (ECCV)*.

Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. 2020b. MatryODShka: Real-time 6DoF Video View Synthesis Using Multi-sphere Images. In *Computer Vision – ECCV 2020*. Springer International Publishing, Cham, 441–459.

Matan Atzmon and Yaron Lipman. 2020. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2565–2574.

Michael Bronstein, 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. *ACM Trans. Graph.* 39, 4, Article 86 (July 2020), 15 pages. https://doi.org/10.1145/3388569.3392485

Rohan Chakra, Jan E Lensen, Eddy Ilg, Tanner Schmidt, Julian Straash, Steven Lovegrove, and Richard Newcombe. 2020. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *European Conference on Computer Vision*. Springer, 608–625.

Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. 2019. Extreme view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7781–7790.

Steven J Gottler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. 1996. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. 43–54.

Amos Gropp, Lior Yariv, Nir Haim, Matan Atzmon, and Yaron Lipman. 2020. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099* (2020).

Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D Graphics. *ACM Trans. Graph.* 31, 6, Article 164 (Nov. 2012), 10 pages. https://doi.org/10.1145/2366145.2366183
