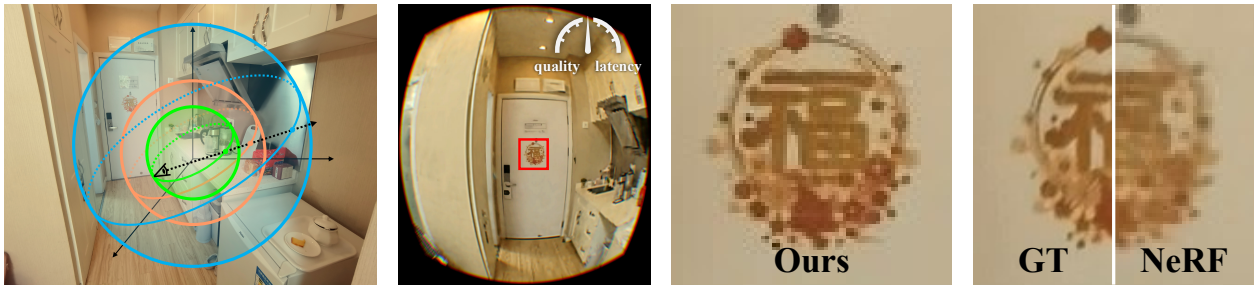
FoV-NeRF: Foveated Neural Radiance Fields for Virtual Reality

Nianchen Deng*, Zhenyi He*, Jiannan Ye, Budmonde Duinkharjav, Praneeth Chakravarthula, Xubo Yang†, and Qi Sun†



(a) scene representation (b) overall quality (c) foveal quality (d) foveal references

Fig. 1: *Illustration of our gaze-contingent neural radiance field for VR.* (a) Visualization of our egocentric-viewing-tailored coordinate and neural scene representation. (b) Our neural encoding and synthesis method matches human visual and stereoscopic acuity and also balances the quality and run-time perceptual latency. Beyond fast performance, (c) and (d) zoom into the foveal region where the human vision has highest sensitivity to imagery quality and stereopsis. **Our** method shows superior quality than alternative neural rendering (**NeRF** as in (d), [33]) approaches, with the full quality rendered ground truth, **GT**, as reference (d).

Abstract— Virtual Reality (VR) is becoming ubiquitous with the rise of consumer displays and commercial VR platforms. Such displays require low latency and high quality rendering of synthetic imagery with reduced compute overheads. Recent advances in neural rendering showed promise of unlocking new possibilities in 3D computer graphics via image-based representations of virtual or physical environments. Specifically, the neural radiance fields (NeRF) demonstrated that photo-realistic quality and continuous view changes of 3D scenes can be achieved without loss of view-dependent effects. While NeRF can significantly benefit rendering for VR applications, it faces unique challenges posed by high field-of-view, high resolution, and stereoscopic/egocentric viewing, typically causing low quality and high latency of the rendered images. In VR, this not only harms the interaction experience but may also cause sickness. To tackle these problems toward six-degrees-of-freedom, egocentric, and stereo NeRF in VR, 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. We then jointly optimize the latency/performance and visual quality while mutually bridging human perception and neural scene synthesis to achieve perceptually high-quality immersive interaction. We conducted both objective analysis and subjective studies to evaluate the effectiveness of our approach. We find that our method significantly reduces latency (up to 99% time reduction compared with NeRF) without loss of high-fidelity rendering (perceptually identical to full-resolution ground truth). The presented approach may serve as the first step toward future VR/AR systems that capture, teleport, and visualize remote environments in real-time.

Index Terms—Virtual Reality; Gaze-Contingent Graphics; Neural Representation; Foveated Rendering

1 Introduction

The unprecedented virtual and augmented reality (VR and AR) hardware revolutions have enabled realistic immersion and natural interaction. An essential need for future VR/AR is to teleport different users and different environments for collaborative scenarios. The recent advancement of neural radiance field (NeRF) methods encode a full radiance field reconstruction via image-based shots [33]. The high quality and flexibility show promising applications for serving as future VR content creation and consumption representations.

Unlike traditional screens, practical VR displays require high field-of-view (FoV), low latency, stereoscopic, and egocentric viewing. These requirements are fundamental to ensure realism and safety without strong simulator sickness. However, the inference from NeRF representation, compared with forward-rendering, demands heavy computation and causes high latency and/or low quality while consumed with immersive viewers.

Acceleration without compromising perceptual quality is the 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 [16, 39, 52]. This is achieved by harnessing high FoV head-mounted displays and high precision eye-tracking technologies. However, existing forward foveated rendering approaches rely on 3D assets such as triangularized geometries. Acquiring such information for physical world objects, such as humans, typically contain noise or low-quality texture. This problem, in comparison, can be harnessed by image-based approaches.

There has been a surge in image-based neural representa-

- *Nianchen Deng and Jiannan Ye are with School of Software, Shanghai Jiao Tong University. E-mail: {dengnianchen, wsyhdyjn}@sjtu.edu.cn.*
- *Zhenyi He is with Department of Computer Science, New York University. E-mail: zh719@nyu.edu.*
- *Budmonde Duinkharjav is with Immersive Computing Lab, New York University. E-mail: budmonde@gmail.com.*
- *Praneeth Chakravarthula is with Computer Science, UNC Chapel Hill. E-mail: cpk@cs.unc.edu*
- *Xubo Yang is with School of Software, Shanghai Jiao Tong University and Peng Cheng Laboratory. E-mail: yangxubo@sjtu.edu.cn.*
- *Qi Sun is with Tandon School of Engineering, New York University. E-mail: qisun@nyu.edu.*

* *These authors contributed equally to this work.*

† *The corresponding authors.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

tions/rendering as an alternative to traditional 3D representations such as polygonal meshes. Examples include voxelization [47], distance fields [33, 38, 45], multi-layer panorama imagery [2, 26], and many more. However, existing solutions suffer from high time consumption (i.e., latency), low image fidelity, or inconsistent robustness of complex scenes with occlusion. Recent accelerations include depicting the inference as basis functions [56] or fast integration via learning gradients [27]. Whereas, no method encodes the human visual system and optimizes for the egocentric and stereoscopic viewing in VR.

We present the first gaze-contingent neural radiance representation and foveated synthesis approach. By encoding human vision's spatial/stereoscopic acuity and temporal sensitivity, we significantly improve responsiveness (from 9s to 20 ms) from alternative neural synthesis methods. This is achieved without loss of visual fidelity compared with the rendered or captured ground truth.

We achieve this by first representing 3D radiance fields with concentric spherical coordinates. The tailored representation both optimizes for egocentric viewing and minimizes inference running time. Our representation also allows for depicting the foveated color and stereopsis sensitivities during the training and inference phases. To this end, we devise retina-matched perceptual models and VR-tailored representation coordinates toward high visual quality as well as low systematic latency. Lastly, we derive an analytical spatial-temporal perception model from optimizing our neural scene representation toward imperceptible loss in image quality and latency.

We validate our system by conducting psychophysical experiments, numerical analysis, and case studies on commercially available VR display devices. The series of experiments reveals our method's effectiveness and advantages: it delivers immersive, high-FoV, and high-fidelity neural radiance fields that are perceived identical to an arbitrarily high-quality rendered 3D scene with perception optimized quality-latency performance. To encourage third-party reproduction and extensions in the community, we will open-source our implementation and dataset. The code and dataset are accessible at https://github.com/dengnianchen/fovnerf. In summary, we make the following major contributions:

- A low-latency and high-fidelity immersive application, offering full perceptual quality with instant high resolution and high FoV first-person VR viewing;
- A 3D neural radiance field representation tailored for egocentric immersive viewing and accelerated inference;
- A human-vision-matched neural synthesis method considering both visual- and stereo- acuity;
- A spatio-temporal analytical model for jointly optimizing systematic latency and perceptual quality for human observers.

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 [14, 22, 44]. Benefiting from the flexible 2D representation, it delivers pixel-wise high quality without compromising rendering performance. However, a major limitation of IBR is its sparse and small viewing range, narrowing the applicability in the smooth VR viewing experience where the head/gaze continuously moves. To address this problem, recent research has been exploring synthesizing novel image-based views instead of fully capturing them. Examples include synthesizing light fields [18, 24, 32] and multi-layer volumetric videos [7]. However, the synthesis usually suffers from the trade-off between optimal quality and fast performance.

2.2 Neural Scene Representation and Rendering

To fully represent a 3D object and environment, neural representations have drawn extensive attention. 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 [33,37,46–48]. Prior arts also investigated implicitly representing shapes [38] and surfaces [31] to improve the visual quality of 3D objects. Inspired by [38], signed distance functions have been deployed for representing raw data such as point cloud [4, 15] with better efficiency [9]. The input information ranged from 2D images [12, 25, 57], 3D context [36, 41, 59], time-varying vector field [35] to local features [28, 51]. 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.

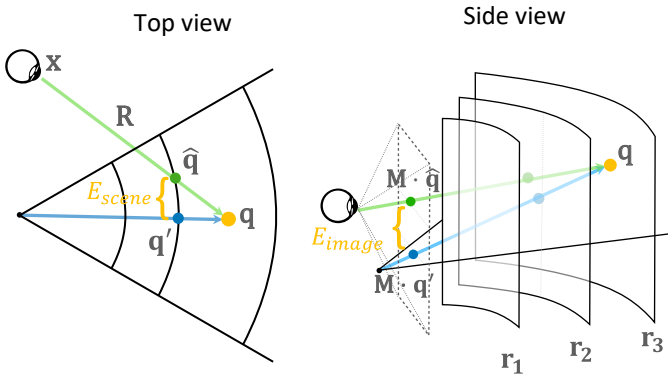


Fig. 2: *Coordinate system and variable annotations.* We partially visualize our 3D full spherical representation with an example of $N = 3$ spheres. Variables and equations are annotated at corresponding positions.

2.3 Panorama-Based 6DoF Immersive Viewing

Panorama images and videos are directly applicable to VR platforms 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 natural viewing experience. Recently, extensive research has been proposed to address this problem. Serrano et al. [43] presented a depth-based dis-occlusion method that dynamically reprojects to 6DoF cameras, enabling natural viewing with motion parallax. Pozo et al. [40] jointly optimizes capture and viewing processes. Machine learning approaches have advanced the robustness of various geometric and lighting conditions [2, 3, 26]. 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/translation ranges, and dis-occlusion artifacts due to the insufficient none-light-of-sight capture. With an orthogonal mission, our model represents and reproduces arbitrary 3D virtual or physical worlds. In Section 5.2, we compare our method with a panoramic-imagery-based view synthesis approach considering visual quality and allowable translation.

2.4 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 and/or enhancing perceptual cues, including mesh- [11, 16, 20, 39], image- [19, 21], and optically-based [10, 49] methods. The accelerations are typically achieved through high-FoV displays (such as VR headsets) and eye-tracking technologies [29]. However, these all require full access to the original 3D assets. Our method bridges human vision and neural rendering. It accelerates and extends neural synthesis beyond local rendering, enabling computational efficiency and perceptually high fidelity.

3 Method

Based on a concentric spherical representation and correspondingly trained network predicting RGBDs (Section 3.1), our system predomi-

nantly comprises two main rendering steps in run-time: synthesizing visual-/stereoscopic-acuity-adaptive elemental images 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 representation sparsity and neural-network complexities (Section 3.3).

3.1 Egocentric Neural Representation and Training

The recent single-object-oriented "outside-in" view synthesis methods [33, 47] typically represent the training targets using uniform voxelization. However, immersive VR environments introduce unique and open challenges for such parameterization due to the commonly egocentric (first-person) and "inside-out" viewing perspective (e.g., Figure 1a). As a consequence, the neural representation on large virtual environments typically suffer from ghosting artifacts, low resolution, or slow speed (Figure 1c).

Egocentric coordinate. To tackle this problem, we are inspired by the recent panoramic imagery dis-occlusion methods [3, 7, 26]: we depict the rapidly varying first-person views with concentric spherical coordinates. This representation has been shown to allow for robust rendering at real-time rates and 6DoF interaction to navigate inside complex immersive environments. As visualized in Figure 2, our representation is parameterized with the number of concentric spheres per neural network (N) and their respective radii ($\mathbf{r} = \{r_i\}, i \in [1,N]$). Under this spherical system, a given 3D spatial position can be represented as $\mathbf{q} = \mathbf{q}(r,\theta,\phi)$ where θ and ϕ are two angular numbers. Similar to [33], the run-time rendering goal is predicting a 4D vector (r,g,b,d) for each $\mathbf{q}$, followed by a view-dependent ray marching through this intermediate function to synthesize individual pixel. Here, (r,g,b) and d are the color and density, respectively.

Neural representation. Existing neural rendering for "outside-in" viewing independently train neural networks to predict the 4D vector for individual $\mathbf{q}$. This is due to the high variations in view points other than the viewing targets. However, egocentric viewing is the opposite: despite the translation for 6DoF viewing, the observers' may change viewing targets frequently by rotating the head and gaze (See Figure 2). That is, given a neural network's capability, a local egocentric neural representation may encode less viewing changes but more spatial variances. To achieve this aim, we design the network to infer an array of vectors per viewing ray:

$$\mathscr{C}(\mathbf{x},\mathbf{v}) \triangleq (\mathscr{R},\mathscr{G},\mathscr{B},\mathscr{D}). \tag{1}$$

Here, $\mathbf{x}/\mathbf{v}$ defines a ray's origin/direction. $\mathscr{R},\mathscr{G},\mathscr{B},\mathscr{D}$ are N dimensional vectors representing the R/G/B/intensities of the N intersecting points ($\mathbf{q}(r_i,\theta_i,\phi_i), i \in [1,N]$) between the ray and the concentric spheres. The system can then render individual color channels via integrating over $\mathscr{R}/\mathscr{G}/\mathscr{B}$ with $\mathscr{D}$ as weights. Denser spherical sampling (i.e., higher N) lead to more precise quality but slower run-time performance.

Inspired by NeRF [33], we devise a machine learning approach that encodes $\mathscr{C}$ as a multi-layer perception (MLP) neural network. NeRF predicts each individual intersection point $\{\hat{\mathbf{q}}_i\}$ before calculates the final color, causing slow performance due to the N network inference operation per ray in run-time. Our egocentric-viewing-tailored representation (Equation (1)) concatenate all N encoded coordinates to a vector and feed the coordinates to the MLP module with only one inference. During training, we define the input $\{\mathbf{x},\mathbf{v}\}$ as the N intersecting points $\hat{\mathbf{q}}_\mathbf{i}$. Our MLP module contains N_m fully-connected layers with N_c channels in each layer.

3.2 Gaze-Contingent Synthesis during Run-time

Our concentric spherical coordinate, as described in Section 3.1, addresses the large view target variance problem in the training stage. However, it may still suffer from significant rendering latency (about half a second for each stereo frame). This is another essential challenges causing neural representation to be unsuitable yet for immersive viewing. In this research, we leverage the spatially adaptive human visual- and stereoscopic- sensitivities to unlock fast runtime inference. 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 Monoscopic Acuity

The human retinal ganglion cells, which collect and transmit visual information, are not uniformly distributed. Instead, its density in the visual fields close to the retinal center is much higher than the periphery [54]. This is referred as foveated vision in accelerating immersive rendering [19, 39]. Inspired by this, we significantly accelerate the runtime inference by integrating the characteristic of spatially-adaptive visual acuity without compromising the *perceptual* quality.

Specifically, given device-tracked camera position ($\mathbf{x}$), direction ($\mathbf{R}$), and gaze position ($\mathbf{v_g}$), we synthesize three elemental images cover different eccentricity ranges of field-of-view (FoV) for each eye: the fovea ($I_f(\mathbf{x},\mathbf{R},\mathbf{v_g})$, $0-20$ deg), the mid-eccentricity ($I_m(\mathbf{x},\mathbf{R},\mathbf{v_g})$, $0-45$ deg) and the entire visual field ($I_p(\mathbf{x},\mathbf{R})$, $0-110$ deg) (Figure 3). Note that I_p is independent from the gaze direction $\mathbf{v_g}$. Given the decreasing visual acuity from low to high eccentricities, we devise two orthogonal networks with different hyper parameters: $\mathscr{C}_{foveal}$ for I_f (the high acuity fovea), and $\mathscr{C}_{periph}$ for I_m and I_p (reduced acuity).

To incorporate the display capabilities and aspect ratios, we define the resolutions of $I_f/I_m/I_p$ as $256^2/256^2/230\times256$. That is, the I_f has the highest angular resolution of 12.8 pixels per degree (PPD), higher than those of I_m (5.7 PPD) and I_p (2.33 PPD).

3.2.2 Adaptive Stereoscopic Acuity

Head-mounted VR displays require stereo rendering to provide parallax depth cues. So do the elemental images $I_f^{\{l,r\}}$, $I_m^{\{l,r\}}$, and $I_p^{\{l,r\}}$ for each (*l*eft and *r*ight) eye. The stereoscopic rendering, however, doubles the inference computation that is critical for latency- and frame-rate-sensitive VR experience (please refer to Section 5.5 for breakdown comparisons).

We accommodate the inference process with the adaptive stereo acuity in perception. In fact, besides the spatial visual acuity in monoscopic vision, psychophysical studies have also revealed human's significantly declined stereopsis while receding from the gaze point [34]. Motivated by this characteristic, we perform the computation with $I_f^{\{l,r\}}$, I_m^c, and I_p^c instead of inferring 6 elemental images, where c indicates the view at the midpoint of the left and right eyes. As I_m^c, and I_p^c have zero disparity, obvious misalignment may exist in the blending area of $I_f^{\{l/r\}}$ and I_m^c. To reduce this misalignment, we shift I_m^c and I_p^c according to the vergence of eyes, i.e. the horizontal difference between the left and right gazes, to introduce a disparity that is close to the foveal. Assuming the difference is $\Delta x_g = x_g^l - x_g^r$ (in pixels), then I_m^c and I_p^c should be shifted by $\Delta x_g/2$ for the left eye and $-\Delta x_g/2$ for the right eye. 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. The output frames are displayed on the stereo VR HMDs. Two adjunct layers are blended using a smooth-step function across 40% of the inner layer as shown in Figure 3. This enhances visual consistency on the edges between layers [16]. Lastly, we enhance the contrast following the mechanism of [39] to further preserve peripheral elemental images' visual fidelity due to its low PDD.

3.3 Latency-Quality Joint Optimization

As a view synthesis system based on sparse egocentric representation (the N spheres per network) and neural network synthesis (the N_m,N_c), neural rendering methods inevitably introduce approximation errors. The errors can be reduced by introducing additional networks (thus

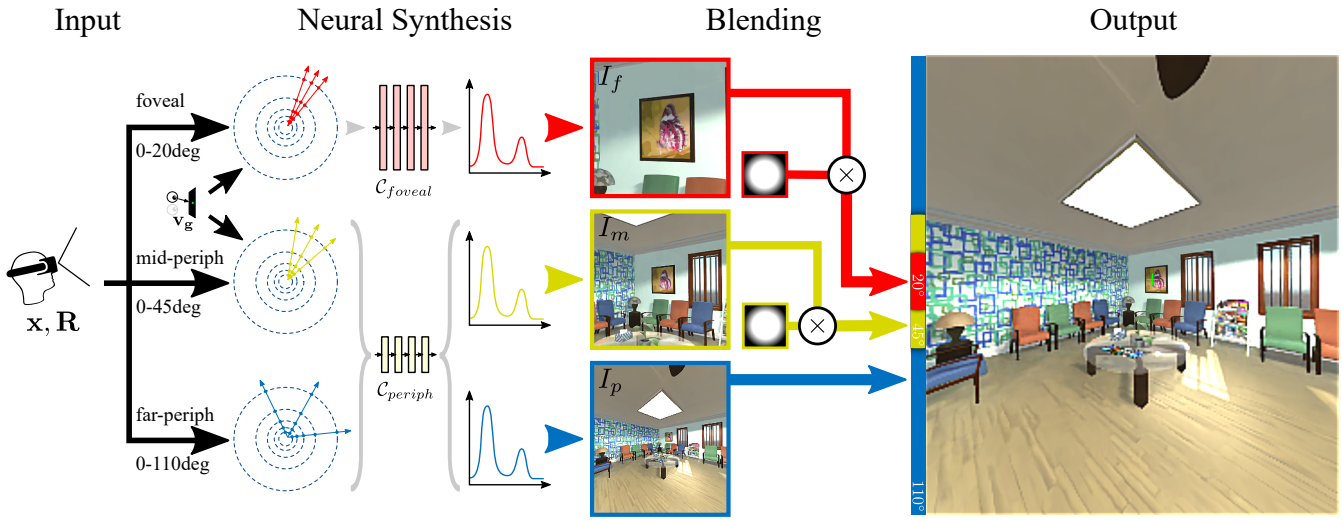
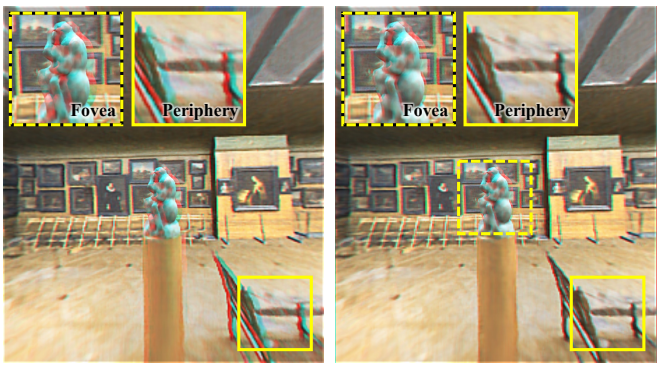


Fig. 3: *Visual acuity adaptive synthesis and rendering mechanism.* We first synthesize the elemental images from our egocentric neural representation for fovea (within 20 deg)/mid-periphery (within 45 deg)/far-periphery (within 110 deg, the capability of our VR display). Then these images are blended to the final displayed frame.



(a) w/o adaptive stereo (b) w/ adaptive stereo

Fig. 4: *Visualization of the adaptive stereo-acuity with anaglyph.* (a) shows the rendered image with 6 retinal sub-images ($I_f^{\{l,r\}}$, $I_m^{\{l,r\}}$, and $I_p^{\{l,r\}}$). (b) shows the rendered image with our adaptive and accelerated inference considering foveated stereoacuity ($I_f^{\{l,r\}}$, $I_m^{\{c\}}$, and $I_p^{\{c\}}$). Our method preserves full stereopsis in the fovea while reducing the angular resolution in the periphery for accelerated inference.

lowered N assuming a fixed number of spheres representing a scene) and increasing individual network's capability (i.e., higher N_m, N_c). However, these variables also significantly increase the online computational time that is determined by inferring function $\mathscr{C}$ and ray marching. While VR strictly demanding both quality and performance, to seek the optimal latency-quality for human viewers, we present a spatial-temporal model that analytically depicts the correlations and optimizes the variables.

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

$$\mathbf{q}'(N,\mathbf{r},\mathbf{q}) \triangleq r_k \frac{\mathbf{q}}{\|\mathbf{q}\|},\ k = \underset{j\in[1,N]}{\operatorname{argmin}} \left(\left| \|\mathbf{q}\| - r_j \right| \right). \tag{2}$$

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 $\{\mathbf{x}, \mathbf{v}\}$ is

$$p(r,\mathbf{x},\mathbf{v}) = \mathbf{x} + \left(\left((\mathbf{x}\cdot\mathbf{v})^2 - \|\mathbf{x}\|^2 + r^2 \right)^{\frac{1}{2}} - \mathbf{x}\cdot\mathbf{v} \right) \mathbf{v}, \tag{3}$$

where $\mathbf{x}$ and $\mathbf{v}$ are the ray's origin point and normalized direction, respectively.

Inversely, given a view point $\mathbf{x}$ observing $\mathbf{q}$, the ray connecting them has the direction $\mathbf{v}(\mathbf{q},\mathbf{x}) = \frac{\mathbf{q}-\mathbf{x}}{\|\mathbf{q}-\mathbf{x}\|}$. This ray may intersect with more than one sphere. Among them, the closest one to $\mathbf{q}$ is:

$$\hat{\mathbf{q}}(\mathbf{x},N,\mathbf{r},\mathbf{q}) \triangleq p(r_k,\mathbf{x},\mathbf{v}(\mathbf{q},\mathbf{x})) \mid k = \underset{j\in[1,N]}{\operatorname{argmin}} \left(\left| \|\mathbf{q}\| - r_j \right| \right). \tag{4}$$

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

$$E_{scene}(N,\mathbf{r}) = \iint \left\| \mathbf{q}'(N,\mathbf{r},\mathbf{q}) - \hat{\mathbf{q}}(\mathbf{x},N,\mathbf{r},\mathbf{q}) \right\| \mathbf{dqdx}, \quad \forall \{\mathbf{x},\mathbf{q}\} \text{ pair without occlusion in between.} \tag{5}$$

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

In comparison, a neural representation aims at predicting projected image given a $\mathbf{x}$ and $\mathbf{R}$. Thus, we further extend Equation (5) to image space to analyze the error given a set of camera's projection matrix $\mathbf{M}(\mathbf{x},\mathbf{R})$ as

$$E_{image}(N,\mathbf{r},\mathbf{x},\mathbf{R}) = \int \left\| \mathbf{M}(\mathbf{x},\mathbf{R}) \cdot (\mathbf{q} - \hat{\mathbf{q}}(\mathbf{x},N,\mathbf{r},\mathbf{q})) \right\| \mathbf{dq}. \tag{6}$$

From Figure 2 and eq. (6), we observe: given a fixed min/max range of $\mathbf{r}$, 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 $\mathbf{q}$s) also determines E_{image}.

However, for neural scene representation, infinitely increasing network capabilities may significantly raise the challenges in training precision and inference performance. Likewise, increasing representation densities (i.e., lower N) and/or network complexities (i.e.,

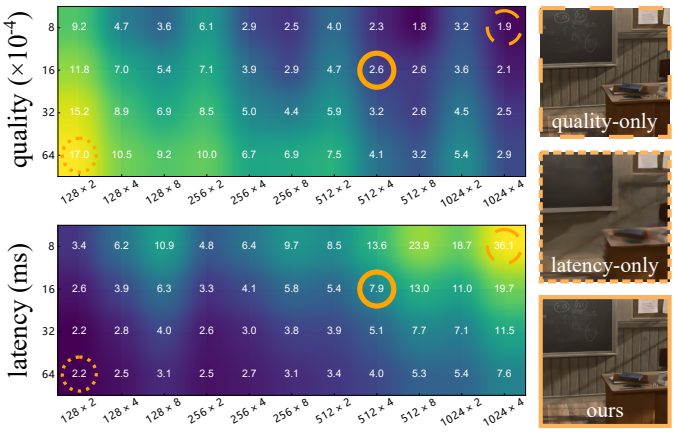


Fig. 5: *Latency-quality joint-optimization.* The two tables on the left plot E_{image} in Equation (6) and latency of an example foveal network during the optimization process (in milliseconds). The values are computed with various settings of $N_c \times N_m$ (X-axis) and N (Y-axis: given a fixed number of spheres representing a scene, lower N indicates the need of more networks thus heavier computation). The images on the right indicate corresponding foveal images under different settings. Our method balances both perceptual quality and latency.

higher N_m/N_c) naturally improves the image output quality (lower Equation (6)). 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 experiment 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 [1, 23], we perform spatial-temporal joint modeling to determine the optimal coordinate system (N) 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:

$$E(N,N_m,N_c) = \sum_t \int \mathscr{C}_{N_m,N_c}(\mathbf{q}) \times \\ \|\mathbf{M}(\mathbf{x}_t,\mathbf{R}_t)\cdot\mathbf{q} - \mathbf{M}(\mathbf{x}_{t-l},\mathbf{R}_{t-l})\cdot\mathbf{q}(r_k,\mathbf{x}_{t-l},\mathbf{v}_{t-l})\|\,\mathbf{dq}, \tag{7}$$

where $l \triangleq l(N,N_m,N_c)$ is the system latency with a given coordinate and network setting. $\mathscr{C}_{N_m,N_c}(\mathbf{q})$ is the four (r,g,b,a) output channels' L1-distance between a given network setting and the highest values $N=8, N_m=4, N_c=1024$). For simplicity, we assumed uniformly distributed $\mathbf{r}$ with a fixed range of the spherical coverage.

As suggested by Albert et al. [1], 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 ([1]), the synthesis and rendering latency shall be less than $L_0 = 24$ms. Thus, we determine the optimal $\{N,\mathbf{r}\}$ to balance latency and precision as

$$\underset{N,N_m,N_c}{\operatorname{argmin}} E(N,N_m,N_c), \quad \text{s.t. } l(N,\mathbf{r}) < L_0. \tag{8}$$

Figure 5 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 5.5.

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.

Datasets. We leveraged two groups of datasets: The synthesized datasets (i.e. *barbershop*, *classroom*, *lobby* and *stones* used in Section 5) are generated from CG scenes by rendering engine[1]. For these scenes, views are sampled uniformly in the translation and rotation box. We generate two separate datasets for every scene to train foveal and periphery networks. The foveal dataset is composed of images rendered with 40 deg field-of-view and 400×400 resolution, while the periphery dataset contains images of 400×400 with 60 deg field-of-view. Our method is also applicable for physical datasets captured by camera (as illustrated in Figure 1). For this kind of datasets, we first record a video of a physical scene using mobile phone. Then images are extracted at every second from the video and their poses are evaluated by Colmap [42]. This forms the foveal dataset. The periphery dataset is built by downsample the images in the foveal dataset by factor of 2.

Optimized network parameters. Guided by the latency-quality joint optimization described in Section 3.3 and 64 spheres depicting all the tested scenes, we implemented the foveal network with $N_m=4$, $N_c=512$ and $N=16$. For the periphery network, the values are $N_m=4$, $N_c=256$ and $N=32$. These optimized parameters achieve a proper balance between quality and latency, as shown in Figure 7 and Table 2.

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 both the foveal network and the periphery network with 200 epochs.

5 Evaluation

With various scenes as shown in Figure 7, we conduct a subjective study (Section 5.1), and calculate objective measurements (Section 5.2) to evaluate our method's perceptual quality compared with alternative solutions ([33]). Further, Section 5.4 validates the method's benefits on enabling large scale translation and view-dependent effects by comparing against prior panorama-based 6DoF synthesize literature [26]. Lastly, we analyze the intra- and inter-system performance in Section 5.5.

5.1 User Study

We conducted a psychophysical experiment to investigate how users perceive our solution (**OURS**) compared with alternative neural view synthesis ([33], **NeRF**) and full-quality ground truth images (**GT**).

Stimuli. For precise comparison and accommodating the low frame rate of alternative solutions (Section 5.5), each group of stimuli consisted of static stereo images rendered via **GT** or synthesized via **OURS/NeRF** with the same views, as shown in Figure 7. They were generated with the same and randomly defined gaze fixation across conditions. The resolution of the image per eye was 1440×1600 (full capability of the VR display). During the study, the target gaze position was indicated as a green cross on the stimuli images. We used the *classroom*, *lobby* and *barbershop* scenes for study use. Each condition from an individual scene consisted of 2 different gaze positions.

For generating **NeRF** condition, we retrained the model from [33] on our dataset with $N_m=8$, $N_c=256$, and $N=64$ for coarse network and $N=128$ for fine network. The top right insets of Figure 7 show 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. Twelve users participated in and completed the study (6 females, $M=23.25$). No participants were aware of the research, the experimental hypothesis, nor the number of conditions. All participants had a normal or corrected-to-normal vision.

[1] *barbershop* and *classroom* are rendered by Blender's Cycles rendering engine, while *lobby* and *stones* are rendered through Unity.

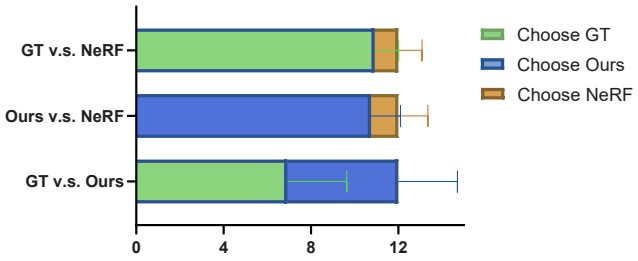


Fig. 6: *The users' preference votes from our evaluation experiment (Section 5.1).* X-axis shows stacked vote results of selecting the condition of each pair. Y-axis lists the pair accordingly.

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 / GT**) with a sampled view and gaze position. To avoid gaze motion variances among individual trials, we enforced static gaze than free-form viewing in each pair of trials. Each stimulus appeared for 500ms on the display. The duration was designed to prevent refocusing and unintentional shifting (0.3-0.5s) as studied in [8]. A forced 0.3sec break (black screen) was introduced between conditions to flush the vision. During the study, the participants were instructed to fix their gazes on a green cross rendered on the display. 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 6 visualizes the results considering all conditions 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 **GT** and **OURS** (42.4% voted for **OURS**, $SD = 0.23$). Meanwhile, a significantly higher ratio of voting **GT** over **NeRF** was observed (91.0%, voted for **GT**, binomial test showed $p < 0.005^{***}$, $SD = 0.11$). The preference applies to **OURS** vs. **NeRF** as well (89.6% voted for **OURS**, $SD = 0.09$, binomial test showed $p < 0.005^{***}$). We then calculated the effect size when sample size is 12 (in our case) and set power to 0.8. To reach 0.8 for power value, effect size is at least to be 1.69 and our effect size $Cohen'sd = 3.5$ for Ours v.s. NeRF, and $Cohen'sd = 4.5$ for GT v.s. NeRF.

Discussion. The close-to-random-guess (50% given a 2AFC task) results in **OURS** vs. **GT** revealed the statistical perceptual similarity. Meanwhile, both conditions showed significant quality preference than **NeRF**. That is, with immersive, stereo, high-FoV, and egocentric viewing settings, **OURS** synthesizes gaze-contingent retinal images with superior perceptual quality than alternative solutions ([33]).Given a fixed network representation capability, **NeRF** synthesizes the whole visual field with identical quality while **OURS** concentrates the representation based on the optimal spatial and stereoscopic acuity.

5.2 Static Visual Quality

Conditions and metrics. Complementary to the subjective measurement (Section 5.1), we further objectively validate the perceptual quality at individual eccentricities in a breakdown fashion. Specifically, with all four scenes and **GT** as the quality reference, we compare **OURS**, **NeRF**, and an additional image-space spatial foveation **F-GT** [17]. As has been studied to be perceptually identical to **GT**, the **F-GT** condition serves as a spatially adaptive quality reference.

For each eccentricity range, we compare the deep perceptual similarity (LPIPS) [58] across all scenes. LPIPS uses deep neural networks to estimate perceptual similarities between the image provided and a reference image. It widely serves as image-based metrics that quantifies humans' visual similarity judgments. Smaller values indicate higher perceptual similarity. Mildenhall et al. [33] also applied LPIPS for quality measurement. For each scene, we sample 25 views with gazes at the middle of the display, resulting in 25 data per eccentricity value, 21 eccentricity value per scene (5 deg step size), thus 525 data per scene. We used one-way repeated measures ANOVAs to compare effects across three stimuli on LPIPS value of each eccentricity range (IV: stimuli, DV: LPIPS value). 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 measurement was conducted at individual eccentricity ranges (from 0 to 110 deg, the capability of the VR HMD).

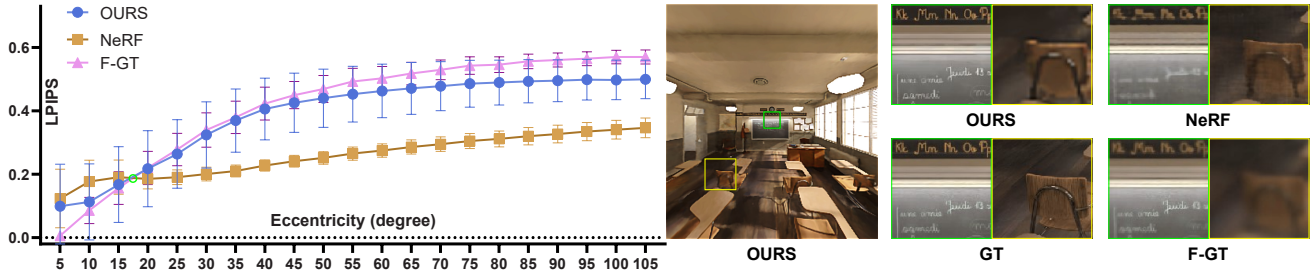
Results. Figure 7 plots LPIPS values across all scenes and eccentricity ranges (5 deg step size). From foveal to near periphery ($\leq$ 15 deg), we observed significant effects of the stimuli on LPIPS with a "large" effect size ($\eta^2 >= 0.15$). That is, **OURS** shows significantly lower LPIPS than **NeRF** ($p <= .005^{***}$), and being higher than **F-GT** in scene classroom and lobby or lower than **F-GT** in scene barbershop and stones. For example, the main effects of stimuli ($F(2,48) = 24.74, p < .001^{***}$) was significant on eccentricity $= [0,30]$ deg in scene *barbershop* (Figure 7a). **OURS** was significantly lower than **NeRF** ($t(24) = -4.26, p < .001^{***}$) and **F-GT** ($t(24) = -9.1, p < .001^{***}$) both with a "large" effect size (Cohen's $d > 0.85$). The example observation generally applies to all 4 scenes being validated. The trend extends to ($\leq$ 35 deg) for the "lobby" and "barbershop" scenes.

From near- to far- eccentricity (¿20 deg), we observed significant effects of the stimuli on LPIPS with a "large" effect size ($\omega^2 = 0.32$). **OURS** shows higher LPIPS than **NeRF** and it became significantly from eccentricity=40 deg. Whereas, comparing with **F-GT**, we observed significant lower scores ($p < .001^{***}$). For instance, the main effects of stimuli was significant on eccentricity $= 40$ deg in scene *lobby* ($F(2,48) = 87.78, p < .001^{***}$, Figure 7c). **OURS** was significantly lower than **F-GT** ($t(24) = -9.095, p < .001^{***}$), and higher than **NeRF** ($t(24) = 3.797, p < .001^{***}$) both with a "large" effect size (Cohen's $d = 0.76$). The example observation generally applies to all 4 scenes being validated.

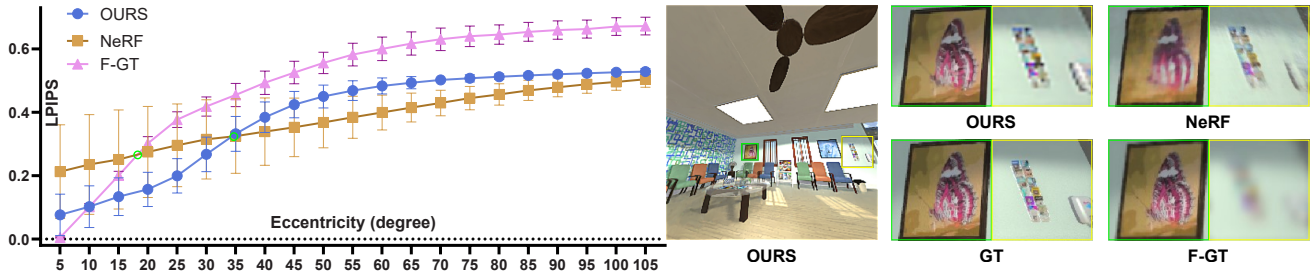
Discussion. In the foveal and near-periphery, the observation revealed our method's significantly higher visual quality than the alternative solutions. This is evidenced by the significantly stronger perceptual similarity to **GT** by comparison between **OURS** and **NeRF**. In our subjective study (Section 5.1), the statistically indistinguishable/preferred voting of **OURS** vs. **GT/NeRF** also agrees with the discovery. In the periphery with eccentricity larger than 20 deg, **OURS** showed lower LPIPS (i.e., perceptually more similar to **GT**) than **F-GT**. The latter has been shown to display identical perceptual similarity to full resolution rendering [39]. 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 above 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 5.5), our method showed superior perceptual quality than **NeRF** under first-person, high resolution, and immersive viewing.

5.3 Dynamic Viewing Quality

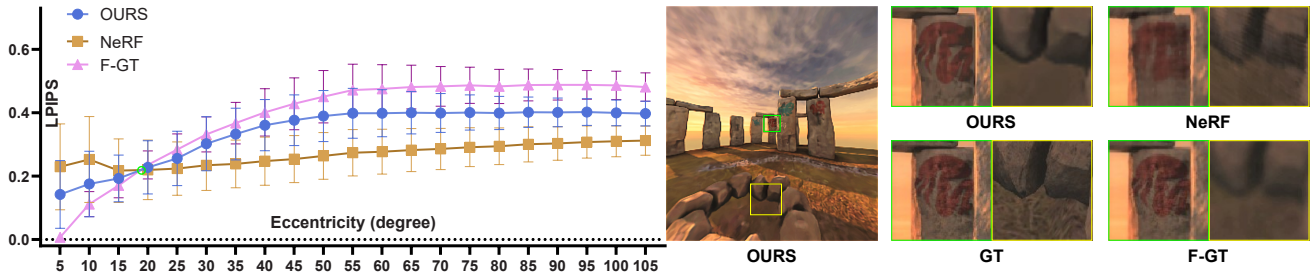
We further validated our method on dynamic scenes allowing free head and gaze motions in a VR immersive environment. Subjective 2AFC evaluations requiring free gaze motion are limited by a fundamental problem where the gaze trajectories differ between discrete 2AFC trials. This inevitably results in variations in perceived stimuli between any two methods being compared, thereby resulting in erroneous comparisons arising from potential bias. To overcome this problem, we



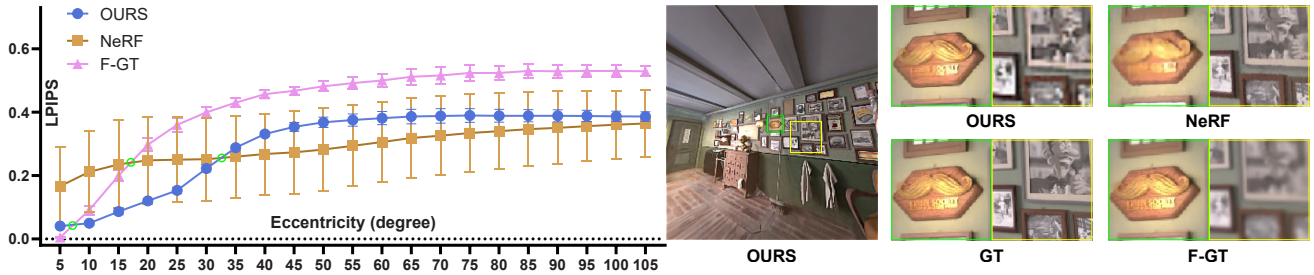
(a) classroom analysis (b) classroom example



(c) lobby analysis (d) lobby example



(e) stone analysis (f) stones example



(g) barbershop analysis (h) barbershop example

Fig. 7: *LPIPS visualization of all scenes and comparison of **OURS**, **NeRF**, and **F-GT** over the visual field.* The left column plots the analysis results from Section 5.2. X-axis indicates the eccentricity range (0 to 110deg). Y-axis shows the LPIPS loss [58] with **GT** condition as reference. Lower values mean more similar perceptual quality to the reference, i.e., better quality. The error bars indicate the standard deviation. The green circles indicate intersection points between curves. The right column visually compares foveal/periphery images with randomly sampled views. Green/yellow rectangle indicates fovea/periphery.

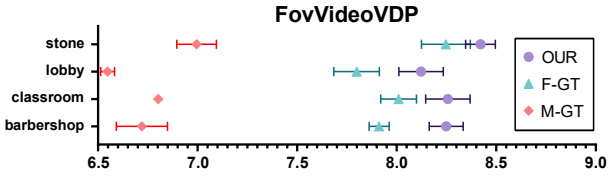


Fig. 8: *The descriptive results from fvvdp measurement.* X-axis shows the FovVideoVDP results (higher values mean perceptually closer to the reference **GT** condition). Y-axis lists four scenes being validated. Error bars indicate standard deviation.

Comparison	Mean Diff	95% CI of Diff	P Value
OURS vs F-GT	0.2706	0.1861 to 0.3551	$< .0001^{****}$
OURS vs M-GT	1.496	1.356 to 1.635	$< .0001^{****}$
F-GT vs M-GT	1.225	1.144 to 1.307	$< .0001^{****}$

Table 1: The post-hoc comparison results for FovVideoVDP ANOVA.

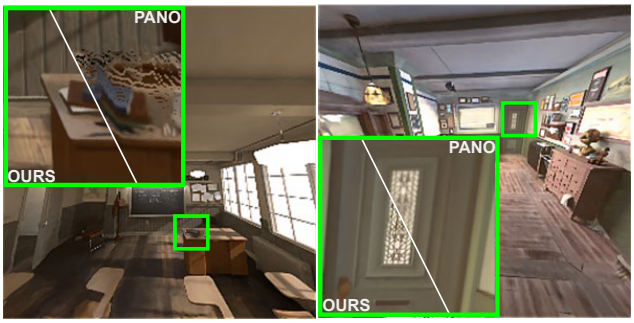
instead employ a novel combination of *subjective* viewing trajectories and recent *objective* perceptual metric for foveated video quality.

Specifically, we collect subjective gaze data wherein the users were asked to freely observe the full-resolution ground truth (**GT**) immersive VR video. We then generate foveated videos using state-of-the-art foveated rendering methods to compare against our method with the collected gaze and head data as the baseline. This way, we make sure that the gaze trajectories across all the methods are maintained same. To predict the perceived visual quality of video stimuli generated using different methods, we use the most recent FovVideoVDP metric [30] which is tailored for large field-of-view (e.g., VR) video sequences considering various perceptual factors such as spatial and temporal effects across visual eccentricities, motion and contrast. The metric effectively predicts the levels of visible visual differences between a given video stimulus and a reference.

Conditions and metrics. In this experiment, we collected 12 randomly sampled (gaze and head) trajectories from the participants in Section 5.2. Each trajectory forms one trial that is about 2 seconds long (i.e., 100 frames with **OURS**). Due to FovVideoVDP's modeling of only monocular views, we measure with the rendered frames with left eyes only, without the loss of generality. **GT**, the video sequence rendered with full-resolution, was considered as the reference in FovVideoVDP. With the target reference, we compare our method with various gaze-contingent methods, including the recently presented real-time ventral metamers [13,53] (**M**(etamer)**-GT**) and foveated real-time foveated rendering [17] (**F-GT**). Due to the significantly low temporal performance of **NeRF** condition (see Section 5.5), the trial duration (2 sec) is shorter than the needed time to generate a single frame with it (90 sec). It showed a significantly low FovVideoVDP rate (≤ 5.1). Thus, we exclude the unnecessary statistical comparison with this condition.

Results and discussion. Figure 8 and Table 1 show the results' statistical distribution. Specifically, with FovVideoVDP's visual similarity range from 0 to 10, **OURS** achieves $M = 8.26, SD = 0.14$, significantly higher than **M-GT** $M = 6.77, SD = 0.18$ and **F-GT** $M = 7.99, SD = 0.19$ with $F(11,22) = 14.15, p < .001^{***}$. We further perform a power analysis by setting sample size to 12 and power to 0.8. To reach 0.8 for power value, effect size is at least to be 1.69 and our effect size for **OURS** v.s. **F-GT** is $Cohen'sd = 2.58$, and $Cohen'sd = 14.24$ for **OURS** v.s. **M-GT**.

In addition to the previous experiment on controlled static viewing (Section 5.2), the analysis demonstrates our outperformance in spatio-temporal visual quality under dynamic viewing conditions. This is validated via comparing with existing rendering approaches that have been demonstrated to provide perceptually identical quality to a full resolution rendering (i.e., **GT** the reference of the metric). The experiment simulates real-world viewing scenarios in immersive environments where users view constantly moving with their heads and gazes.



(a) large translation w/ occlusion (b) small translation w/ flat surface

Fig. 9: *Qualitative comparison between **OURS** and panorama-based view synthesis.* Comparison between our model (**OURS**), and **PANO** [26] shows that the quality of images varies most where the scene contains complex geometries, and/or a wide scale of depth variance and occlusion (cf. *desk with books* in (a)). Meanwhile, presence of high frequency textures (cf. *door window* in (b)) show marginal difference between the methods.

5.4 Qualitative Comparison with Panorama-Based View Synthesis

Conditions and implementation. The synthesis quality of panoramic-imagery-based approaches is determined by the occlusion and translation ranges. Thus, as a case study, we visually compare our synthesis approach with [26] and varied translation ranges from the camera origin. To produce a *Layered Depth Panorama* ([26], **PANO**), we render 16 images from a virtual camera ring rig around the origin of the scene. Then as an identical input format, we passed the images to Lin et al.'s pre-trained model. Once a **PANO** is generated, we query it for images corresponding to the same camera poses used for **OURS**.

Dis-occlusion quality. Figure 9 shows sampled views of failure (large translation range of 30cm while facing occluded geometries, Figure 9a) and success (small translation range of 15cm while facing a flat surface, Figure 9b) modes of **PANO** in comparison to **OURS**. We observe that our model performs better in scenes with high occlusion and depth variance, while presenting a marginal difference for mostly flat geometries. We hypothesize that this observation is largely because our approach presents better congruence towards learning complex geometries, whereas **PANO** is biased towards modeling geometries as planes. We note that this bias is likely conducive to **PANO**'s ability to retain high frequency textures if the variance in depth is minimal.

View-dependent effects. Another limitation faced by single view-port based panorama image view synthesis is the limited view-dependent effects. This becomes more visible in highly glossy and reflective scenes. Figure 10 provides an example that visually compares **OURS**' advancement in producing realistic view-dependent effects.

5.5 Performance

Virtual and augmented reality demands high frame rates along with high quality to ensure an immersive and comfortable experience. Our neural synthesis method achieves real-time performance through our egocentric neural representation (Section 3.1) as well as the spatial and angular (stereoscopic) foveation (Section 3.2). Further, our latency-quality joint optimization for the representation and network also balances quality and performance (Section 3.3). Here, we evaluate the performance of each component 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 2. While our egocentric neural representation (which costs 562ms) significantly speedup the synthesis compared to **NeRF** and **PANO** (the bottom rows of Table 2), our spatial foveation further

(a) **OURS**

(b) **PANO**

Fig. 10: *Comparing view-dependent effects between **OURS** and **PANO**.* Note the view-dependent effect appeared on the glossy wall-mounted clock. With an up-down rotating and translating camera, **OURS** (a) shows stronger view-dependent effects than **PANO** (b) condition.

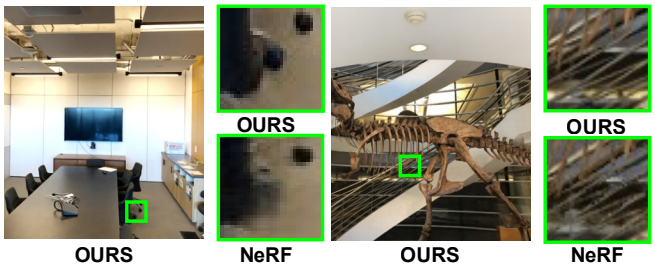
improves the system performance from offline computation to interactive speed (about 30FPS). Finally, our full method with both spatial and stereoscopic foveation achieves a high performance of 50 FPS, contributing to a temporally continuous viewing experience without loss of perceived resolution and quality.

foveal infer (per eye)	8
mid- & far-periphery infer (per eye)	5.4
blending & contrast enhancement	0.1
OUR method without foveation	562
OUR method without stereo foveation	27
OUR full method	**21.5**
NeRF [33]	9.0×10^4
PANO [26]	1.0×10^4

Table 2: *Time consumption breakdown and comparison.* The numbers show the average time consumption (in ms) of each component and the overall system per frame. All units are in millisecond.

6 Conclusion

We present a gaze-contingent neural scene representation and view synthesis method tailored for egocentric and stereoscopic VR viewing. To unlock the practical deployment of neural radiance fields in VR, we overcome their challenges such as high latency, low resolution, and low fidelity that are exacerbated with stereoscopic and immersive displays. This is achieved by encoding not only the scene content but also *how* human vision perceives it, i.e., the gaze-contingent visual acuity and stereopsis. Our network individually synthesizes foveal, mid-, and far-periphery retinal images, and then blended them to form a wide field-of-view image. We also derive an analytical model depicting quality-latency balance and optimizes these two essential factors based on psychophysical study data. Orthogonal to traditional rendering pipeline, our method takes advantage of NeRF's versatile capability of synthesising not only virtual content but also physically captured scenes, as demonstrated in Figure 11. Compared with NeRF [33], our method creates significantly faster and higher perceptual fidelity for high resolution and high FoV immersive viewing. Furthermore, with the support of view-dependent effects (Figure 10), our method is



(a) room scene (b) trex scene

Fig. 11: *Neural rendering with physically captured scenes.* With the data from [33], we compare the visual quality between **OURS** and **NeRF**. Due to the limited data field-of-view coverage, we trained the model with fovea-only network parameters. The zoom-in images further visualize **OURS**' enhanced quality in high depth disparity and high frequency areas.

robust to occluded scenes with complex geometry, a challenge faced by panorama-based synthesis approaches (e.g., [26]).

Limitations and future work. While our method achieves superior performance compared to existing approaches, its broader deployment requires combating several constraints. Our multi-spherical representation renders optimal quality when the virtual camera is within the innermost sphere. Although our method unlocks large translation scale than panorama + neural dis-occlusion approaches, the optimal translation range is constrained by the radius of this sphere. However, infinitely enlarge it will decrease synthesized image quality. Therefore, devising an adaptive and dynamic system that automatically optimizes the coordinates may shed lights on allowing traversal-scale translation range without decreasing the perceptual quality or performance. Similarly, multiscale coordinates that consider various level-of-details of the 3D space have shown their effectiveness of interpolating geometries [55]. Developing a corresponding network that synthesizes the imagery from global to local level-of-details would be an interesting direction for future work that extends the applicability to large-scale scenes with strong occlusions.

Aliasing in the periphery exists because of the low sampling rate. Although we have not apply any anti-aliasing methods yet, the result of our user study shown that the aliasing doesn't cause significant degradation of perceptual quality. While traditional anti-aliasing methods (such as MSAA used in Guenter et al. [16]) may have an effect, these methods will introduce a significant performance penalty when combined with neural synthesis methods. We think that a candidate solution may be applying multi-scale-encoding, as proposed by Barron et al. [5, 6].

We sample the scene fully based on eccentricity, considering acuity and stereopsis. However, we envision that fine-grained visual sensitivity analysis, such as luminance [52] or depth [50], would provide more insights on achieving higher quality and/or faster performance.

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 $\mathbf{r}$. This is due to the potentially significantly higher 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., $\mathbf{r}$) of the spheres would further improve the synthesis quality and performance. The perceptually-based method is orthogonal yet compatible with other remarkable neural radiance field inference acceleration solutions, e.g., [27, 56]. Combining varies perspectives may unlock the future instant and high quality immersive viewing experience such as teleportation.

Acknowledgments

This work was partially supported by the National Key Research and Development Program of China (2018YFB1004902).

REFERENCES

[1] R. Albert, A. Patney, D. Luebke, and J. Kim. Latency requirements for foveated rendering in virtual reality. *ACM Transactions on Applied Perception (TAP)*, 14(4):1–13, 2017.

[2] B. Attal, S. Ling, A. Gokaslan, C. Richardt, and J. Tompkin. MatryODShka: Real-time 6DoF video view synthesis using multi-sphere images. In *European Conference on Computer Vision (ECCV)*, Aug. 2020.

[3] B. Attal, S. Ling, A. Gokaslan, C. Richardt, and J. Tompkin. Matryodshka: Real-time 6dof video view synthesis using multi-sphere images. In *Computer Vision – ECCV 2020*, pp. 441–459. Springer International Publishing, Cham, 2020.

[4] M. Atzmon and Y. Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2565–2574, 2020.

[5] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5855–5864, October 2021.

[6] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022.

[7] M. Broxton, J. Flynn, R. Overbeck, D. Erickson, P. Hedman, M. Duvall, J. Dourgarian, J. Busch, M. Whalen, and P. Debevec. Immersive light field video with a layered mesh representation. *ACM Trans. Graph.*, 39(4), July 2020. doi: 10.1145/3386569.3392485

[8] F. Campbell and G. Westheimer. Dynamics of accommodation responses of the human eye. *The Journal of physiology*, 151(2):285–295, 1960.

[9] R. Chabra, J. E. Lenssen, E. Ilg, T. Schmidt, J. Straub, S. Lovegrove, and R. Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *European Conference on Computer Vision*, pp. 608–625. Springer, 2020.

[10] P. Chakravarthula, Z. Zhang, O. Tursun, P. Didyk, Q. Sun, and H. Fuchs. Gaze-contingent retinal speckle suppression for perceptually-matched foveated holographic displays. *IEEE Transactions on Visualization and Computer Graphics*, 27(11):4194–4203, 2021.

[11] S. Chen, B. Duinkharjav, X. Sun, L.-Y. Wei, S. Petrangeli, J. Echevarria, C. Silva, and Q. Sun. Instant reality: Gaze-contingent perceptual optimization for 3d virtual reality streaming. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2157–2167, 2022.

[12] I. Choi, O. Gallo, A. Troccoli, M. H. Kim, and J. Kautz. Extreme view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7781–7790, 2019.

[13] J. Freeman and E. P. Simoncelli. Metamers of the ventral stream. *Nature neuroscience*, 14(9):1195–1201, 2011.

[14] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 43–54, 1996.

[15] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020.

[16] B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder. Foveated 3d graphics. *ACM Trans. Graph.*, 31(6), Nov. 2012. doi: 10.1145/2366145.2366183

[17] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1072–1080, 2015.

[18] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2016)*, 35(6), 2016.

[19] A. S. Kaplanyan, A. Sochenov, T. Leimkühler, M. Okunev, T. Goodall, and G. Rufo. Deepfovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Trans. Graph.*, 38(6), Nov. 2019. doi: 10.1145/3355089.3356557

[20] R. Konrad, A. Angelopoulos, and G. Wetzstein. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Trans. Graph.*, 39, 2020.

[21] B. Krajancich, P. Kellnhofer, and G. Wetzstein. Optimizing depth perception in virtual and augmented reality through gaze-contingent stereo rendering. *ACM Trans. Graph.*, 39, 2020.

[22] M. Levoy and P. Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 31–42, 1996.

[23] M. Li, Y.-X. Wang, and D. Ramanan. Towards streaming perception. In *Computer Vision – ECCV 2020*, pp. 473–488. Springer International Publishing, Cham, 2020.

[24] Q. Li and N. Khademi Kalantari. Synthesizing light field from a single image with variable mpi and two network fusion. *ACM Transactions on Graphics*, 39(6), 12 2020. doi: 10.1145/3414685.3417785

[25] C.-H. Lin, C. Wang, and S. Lucey. Sdf-srn: Learning signed distance 3d object reconstruction from static images. *arXiv preprint arXiv:2010.10505*, 2020.

[26] K.-E. Lin, Z. Xu, B. Mildenhall, P. P. Srinivasan, Y. Hold-Geoffroy, S. DiVerdi, Q. Sun, K. Sunkavalli, and R. Ramamoorthi. Deep multi depth panoramas for view synthesis. In *Computer Vision – ECCV 2020*, pp. 328–344. Springer International Publishing, Cham, 2020.

[27] D. B. Lindell, J. N. P. Martel, and G. Wetzstein. Autoint: Automatic integration for fast neural volume rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[28] L. Liu, J. Gu, K. Z. Lin, T.-S. Chua, and C. Theobalt. Neural sparse voxel fields. *arXiv preprint arXiv:2007.11571*, 2020.

[29] C. Lu, P. Chakravarthula, Y. Tao, S. Chen, and H. Fuchs. Improved vergence and accommodation via purkinje image tracking with multiple cameras for ar glasses. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 320–331. IEEE, 2020.

[30] R. K. Mantiuk, G. Denes, A. Chapiro, A. Kaplanyan, G. Rufo, R. Bachy, T. Lian, and A. Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19, 2021.

[31] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4460–4470, 2019.

[32] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019.

[33] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[34] H. Mochizuki, N. Shoji, E. Ando, M. Otsuka, K. Takahashi, T. Handa, et al. The magnitude of stereopsis in peripheral visual fields. *Kitasato Med J*, 41:1–5, 2012.

[35] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5379–5389, 2019.

[36] M. Oechsle, L. Mescheder, M. Niemeyer, T. Strauss, and A. Geiger. Texture fields: Learning texture representations in function space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4531–4540, 2019.

[37] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 3500–3509, 2017.

[38] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 165–174, 2019.

[39] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Trans. Graph.*, 35(6), Nov. 2016. doi: 10.1145/2980179.2980246

[40] A. P. Pozo, M. Toksvig, T. F. Schrager, J. Hsu, U. Mathur, A. Sorkine-Hornung, R. Szeliski, and B. Cabral. An integrated 6dof video camera and system design. *ACM Trans. Graph.*, 38(6), Nov. 2019. doi: 10.1145/3355089.3356555

[41] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2304–2314, 2019.

[42] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[43] A. Serrano, I. Kim, Z. Chen, S. DiVerdi, D. Gutierrez, A. Hertzmann, and B. Masia. Motion parallax for 360° rgbd video. *IEEE Transactions on Visualization and Computer Graphics*, 2019.

[44] H. Shum and S. B. Kang. Review of image-based rendering techniques. In

Visual Communications and Image Processing 2000, vol. 4067, pp. 2–13. International Society for Optics and Photonics, 2000.

[45] V. Sitzmann, E. R. Chan, R. Tucker, N. Snavely, and G. Wetzstein. Metasdf: Meta-learning signed distance functions. In *arXiv*, 2020.

[46] V. Sitzmann, J. N. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*, 2020.

[47] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2019.

[48] V. Sitzmann, M. Zollhöfer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019.

[49] Q. Sun, F.-C. Huang, J. Kim, L.-Y. Wei, D. Luebke, and A. Kaufman. Perceptually-guided foveation for light field displays. *ACM Trans. Graph.*, 36(6), Nov. 2017. doi: 10.1145/3130800.3130807

[50] Q. Sun, F.-C. Huang, L.-Y. Wei, D. Luebke, A. Kaufman, and J. Kim. Eccentricity effects on blur and depth perception. *Opt. Express*, 28(5):6734–6739, Mar. 2020. doi: 10.1364/OE.28.006734

[51] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhöfer, C. Stoll, and C. Theobalt. Patchnets: Patch-based generalizable deep implicit 3d shape representations. In *European Conference on Computer Vision*, pp. 293–309. Springer, 2020.

[52] O. T. Tursun, E. Arabadzhiyska-Koleva, M. Wernikowski, R. Mantiuk, H.-P. Seidel, K. Myszkowski, and P. Didyk. Luminance-contrast-aware foveated rendering. *ACM Trans. Graph.*, 38(4), July 2019. doi: 10.1145/3306346.3322985

[53] D. R. Walton, R. K. D. Anjos, S. Friston, D. Swapp, K. Akşit, A. Steed, and T. Ritschel. Beyond blur: real-time ventral metamers for foveated rendering. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[54] A. B. Watson. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of vision*, 14(7):15–15, 2014.

[55] T. Winkler, J. Drieseberg, M. Alexa, and K. Hormann. Multi-scale geometry interpolation. *Computer Graphics Forum*, 29(2):309–318, 2010. doi: 10.1111/j.1467-8659.2009.01600.x

[56] S. Wizadwongsa, P. Phongthawee, J. Yenphraphai, and S. Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[57] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, B. Ronen, and Y. Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020.

[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[59] Z. Zhang, Z. Yang, C. Ma, L. Luo, A. Huth, E. Vouga, and Q. Huang. Deep generative modeling for scene synthesis via hybrid representations. *ACM Transactions on Graphics (TOG)*, 39(2):1–21, 2020.