BioNeRF: Biologically Plausible Neural Radiance Fields for View Synthesis

Abstract

This paper presents BioNeRF, a biologically plausible architecture that models scenes in a 3D representation and synthesizes new views through radiance fields. Since NeRF relies on the network weights to store the scene's 3-dimensional representation, BioNeRF implements a cognitive-inspired mechanism that fuses inputs from multiple sources into a memory-like structure, improving the storing capacity and extracting more intrinsic and correlated information. BioNeRF also mimics a behavior observed in pyramidal cells concerning contextual information, in which the memory is provided as the context and combined with the inputs of two subsequent neural models, one responsible for producing the volumetric densities and the other the colors used to render the scene. Experimental results show that BioNeRF outperforms state-of-the-art results concerning a quality measure that encodes human perception in two datasets: real-world images and synthetic data.

Keywords: View Synthesis, Neural Rendering, iologically Plausible Neural Models

1. Introduction

Neural Radiance Fields (NeRF) [1] provide a memory-efficient way to address the problem of rendering new synthetic views based on a set of input images and their respective camera poses. The model implements a Multilayer Perceptron (MLP) network whose weights store a tridimensional scene representation, producing photorealistic quality outputs from new viewpoints while preserving complex features concerning material reflectance and geometry.

Roughly speaking, NeRF comprises a fully connected network that performs regression tasks, i.e., it maps a 5D coordinate sampled from a particular camera viewpoint location relative to the scene, (x, y, z), and their corresponding 2D viewing directions (θ, ϕ), to a view-dependent RGB color, comprising a set of three real-valued numbers, and a single volume density Δ. Such 5D coordinates are transformed using a positional encoding approach to extract higher frequencies, consequently improving the quality of high-resolution representations. Further, the model computes the radiance emitted in such direction using classical volume rendering techniques [2], accumulating those colors and densities into a 2D image. Figure 1 illustrates the concept behind NeRF.

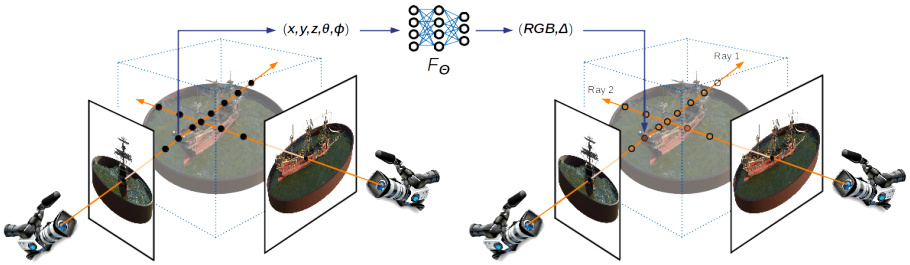

Figure 1: NeRF synthesizes images by sampling 5D coordinates (location and viewing direction) along camera rays (left) and feeding those locations into an MLP to produce color and volume density (right).

One of NeRF's fundamental concepts regards encoding complex real-world 3D scene representation in the parameters of a neural network. Besides, the model restricts the volume density prediction as a function of the camera position, allowing the RGB color to be predicted as a function of both location and viewing direction. Such concepts resemble some more biologically plausible studies inspired by neuroscience discoveries and principles of pyramidal cells [3], especially concerning the idea of context to steer the information flow [4, 5, 6] and integrated memory, responsible for providing additional context based on past experiences [7, 8].

This paper proposes BioNeRF, aka Biologically Plausible Neural Radiance Fields, that implements two parallel networks, one to forecast Δ and the other

to predict the RGB color, which interact among themselves providing a context to each other, as well as a memory mechanism to store intrinsic and correlated information in the learning flow. Such mechanisms are responsible for improving the quality of the generated views and obtaining promising results over two datasets, one composed of synthetic images and the other comprising real images. The main contributions of this work are described as follows:

- To propose BioNeRF, a novel and biologically plausible architecture for view synthesis;
- To introduce memory and context to view synthesis ambiance; and
- To obtain state-of-the-art results concerning a quality measure that encodes human perception in the context of view synthesis.

The remainder of this paper is described as follows. Section 2 presents the related works, while Section 3 provides a theoretical background regarding the NeRF concepts. Section 4 introduces the BioNeRF, while Sections 5 and 6 describe the methodology and present the experimental results and discussions, respectively. Finally, Section 7 states conclusions and future works.

2. Related Works

Before NeRF, Mildenhall et al. [9] proposed the Local Light Field Fusion (LLFF), a method for synthesizing high-quality scene views requiring a significantly smaller number of instances than traditional methods at the time. Meanwhile, Lombardi et al. [10] proposed Neural Volumes (NV), a neural volumetric representation method for modeling and rendering dynamic scenes. The technique improved the image resolution without increasing the resolution of the voxel grid. Besides, Sitzmann et al. [11] proposed the Scene Representation Networks (SRNs), which employed unsupervised learning and convolutional neural networks to learn rich representations of 3D scenes without explicit supervision. Further, Liu et al. [12] proposed the Neural Sparse Voxel Fields (NSVF), which

rendered dynamic, large-scale scenes using voxel representations and progressive training to obtain high-quality real-time renderings with a reduced time. Finally, in 2021, Mildenhall et al. [1] proposed the Neural Radiance Fields. This method uses a fully connected neural network to represent scenes as neural radiance fields, synthesizing new views of complex scenes given the camera position and direction and predicting the volumetric densities and emitted color, combined to generate new views using traditional rendering techniques.

In the same year, Wang et al. [13] proposed the Image-based Rendering Neural Radiance Field (IBRNet), a learning-based image rendering method whose architecture consists of a multilayer perceptron network and a ray transformer, capable of continuously predicting colors and spatial densities from multiple views and a per-scene fine-tuning procedure. In the meantime, Barron et al. [14] introduced the Multiscale Representation for Anti-Aliasing Neural Radiance Fields (Mip-NeRF). This method extends NeRF to represent the scene at a continuously-valued scale, improving the representation and rendering of three-dimensional scenes and overcoming the limitations of aliasing and blurring.

Chen et al. [15] proposed the Tensorial Radiance Fields (TensoRF), a technique that models and reconstructs radiance fields using 4D tensor representation. The work employs the CANDECOMP/PARAFAC decomposition to factorize tensors into compact components and vector-matrix decompositions to relax the component's constraints. Concurrently, Fridovich et al. [16] presented the Radiance fields without neural networks (Plenoxels), a system for photorealistic image synthesis that represents a scene as a sparse 3D grid with spherical harmonics. The work of Xu et al. [17] introduced the Point-Based Neural Radiance Fields (Point-NeRF), a method for reconstructing and rendering three-dimensional scenes that includes rebuilding a point directly from input images via network inference. Furthermore, Verbin et al. [18] focused on representing specular reflections in 3D scenes in the Structured View-Dependent Appearance for Neural Radiance Fields (Ref-NeRF). The model employs structured refinements to capture vision-dependent appearance, obtaining state-of-the-art results with photorealistic images generated from new points of view.

More recently, Chen et al. [19] proposed the Local-to-Global Registration for Bundle-Adjusting Neural Radiance Fields (L2G-NeRF). This method combines local pixel and global frame alignment to achieve high-fidelity reconstruction and addresses large camera pose misalignments. Meanwhile, Kulhanek et al. [20] presented the Tetra-NeRF, a method for representing neural radiance fields using tetrahedra that combines concepts from 3D geometry processing, triangle-based rendering, point cloud generation, Delaunay triangulation, barycentric interpolation, and modern neural radiance fields. Finally, Yao et al. [21] proposed the SpikingNeRF, which aligns the radiance rays with the temporal dimension of Spiking Neural Networks, producing an energy-efficient approach associated with biologically plausible inspired concepts.

3. Neural Radiance Fields

Mildenhall et al.[1] proposed NeRF for view synthesis purposes, i.e., generating 3D scenes or objects given a set of images captured at different angles. The process consists of approximating a function $F_\Theta : (\mathbf{l}, \mathbf{d}) \rightarrow (\mathbf{c}, \Delta)$ that receives the object's 3D location as input, i.e., $\mathbf{l} = (x, y, z)$, and its 2D viewing direction $\mathbf{d} = (\theta, \phi)$. The function outputs the object's emitted color $\mathbf{c} = (r, g, b)$ and its volumetric density Δ. We can represent function F_Θ using an MLP in which one wants to optimize the weights Θ.

The MLP network processes the input $\mathbf{l}$ through 8 fully connected layers using a Rectified Linear Unit (ReLU) activation function and 256 channels per layer, and the output consists of the Δ and a feature map with 256 dimensions. At this stage, Δ does not depend on the view direction, i.e., a translucent point remains translucent regardless of the direction in which it is viewed. Next, the feature vector is concatenated with $\mathbf{d}$, which represents the view direction and serves as input to a fully connected layer using ReLU and 128 channels, which outputs the color information $\mathbf{c}$.

Rays passing through the scene are processed using the principles of volume rendering, computed through differentiability, which can be optimized using the

gradient descent method. The objective is to minimize the error between the observed image and the rendered view.

One can summarize NeRF working mechanism as follows:

- To cast camera rays through the scene to sample 3D points;
- To use the 3D points $\mathbf{l}$ and their corresponding 2D viewing directions $\mathbf{d}$ as input to an MLP to produce a set of output colors $\mathbf{c}$ and densities Δ; and
- To employ volume rendering techniques to accumulate those colors and densities into a 2D image.

4. Biologically Plausible Neural Radiance Fields

This paper introduces BioNeRF, which combines features extracted from both the camera position and direction into a memory mechanism inspired by cortical circuits [22, 23]. Further, such memory is employed as a context to leverage the information flow into deeper layers. Figure 2 illustrates the BioNeRF architecture, which spans four main modules: (i) Positional Feature Extraction, (ii) Cognitive Filtering, (iii) Memory Updating, and (iv) Contextual Inference.

Positional Feature Extraction. The first step consists of feeding two neural models simultaneously, namely M_Δ and M_c, such that $\Theta_\Delta = (\boldsymbol{W}_\Delta, \boldsymbol{b}_\Delta)$ and $\Theta_c = (\boldsymbol{W}_c, \boldsymbol{b}_c)$ denote their respective set of parameters, i.e., neural weights ($\boldsymbol{W}$) and biases ($\boldsymbol{b}$). The output of these models, i.e., $\boldsymbol{h}_\Delta$ and $\boldsymbol{h}_c$, encodes positional information from the input image. Although the input is the same, the neural models do not share weights and follow a different flow in the next steps.

Cognitive Filtering. This step performs a series of operations from now on called *filters* that work on the embeddings coming from the previous step[1].

[1]We call those operations *filters*, for they output a value in the interval $[0, 1]$ for each feature coming from the embeddings $\boldsymbol{h}_\Delta$ and $\boldsymbol{h}_c$.

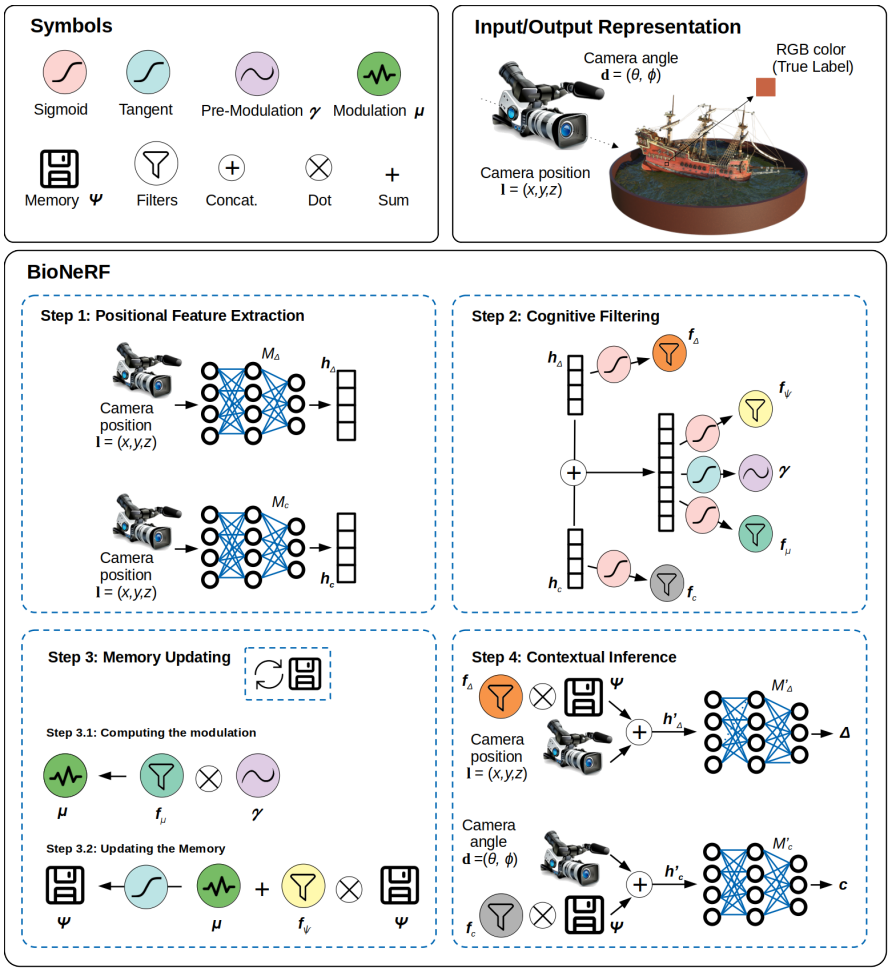

Figure 2: Biologically Plausible Neural Radiance Fields. The top-left frame describes the symbols, while the top-right frames depict the input/output variables. The bottom block illustrates the model's overall pipeline, comprising four steps. Step 1 describes the **Positional Feature Extraction**, which shall consist of two MLP blocks, namely M_Δ and M_c, responsible for extracting relevant information from the camera input to generate $\boldsymbol{h}_\Delta$ and $\boldsymbol{h}_c$. Step 2, i.e., **Cognitive Filtering**, illustrates the filters' generation process, while the **Memory Updating** in step 3 depicts the memory updating schema. Finally, Step 4, i.e., **Contextual Inference**, shows how the memory is filtered and concatenated with the camera position to feed M'_Δ to generate Δ and combined with the camera angle to feed M'_c to generate $\boldsymbol{c}$. Notice that Δ and $\boldsymbol{c}$ are used to render the output compared against the pixel color to compute the BioNeRF's loss.

There are four filters this step derives: (i) density ($\boldsymbol{f}_\Delta$), color ($\boldsymbol{f}_c$), memory ($\boldsymbol{f}_\psi$), and modulation ($\boldsymbol{f}_\mu$), computed as follows:

$$\boldsymbol{f}_\Delta = \sigma\left(\boldsymbol{h}_\Delta\right), \tag{1}$$

$$\boldsymbol{f}_c = \sigma\left(\boldsymbol{h}_c\right), \tag{2}$$

$$\boldsymbol{f}_\psi = \sigma\left(\boldsymbol{W}_\psi[\boldsymbol{h}_\Delta, \boldsymbol{h}_c] + \boldsymbol{b}_\psi\right), \tag{3}$$

and

$$\boldsymbol{f}_\mu = \sigma\left(\boldsymbol{W}_\mu[\boldsymbol{h}_\Delta, \boldsymbol{h}_c] + \boldsymbol{b}_\mu\right), \tag{4}$$

where $\boldsymbol{W}_\psi$ and $\boldsymbol{W}_\mu$ correspond to the weight matrices for the memory and modulation filters, respectively. Additionally, $\boldsymbol{b}_{\boldsymbol{\psi}}$ and $\boldsymbol{b}_{\boldsymbol{\mu}}$ stand for their respective biases. Moreover, $[\boldsymbol{h}_\Delta, \boldsymbol{h}_c]$ represents the concatenation of embeddings $\boldsymbol{h}_\Delta$ and $\boldsymbol{h}_c$, while σ denotes a sigmoid function. The pre-modulation $\boldsymbol{\gamma}$ is computed as follows:

$$\boldsymbol{\gamma} = tanh\left(\boldsymbol{W}_\gamma[\boldsymbol{h}_\Delta, \boldsymbol{h}_c] + \boldsymbol{b}_\gamma\right), \tag{5}$$

where $tanh(\cdot)$ is the hyperbolic tangent function, while $\boldsymbol{W}_\gamma$ and $\boldsymbol{b}_\gamma$ are the pre-modulation weight matrix and bias, respectively.

Memory Updating. Updating the memory requires the implementation of a mechanism capable of obliterating trivial information, which is performed using the memory filter $\boldsymbol{f}_\psi$ (Step 3.1 in Figure 2). Fist, one needs to compute the modulation $\boldsymbol{\mu}$, where $\otimes$ represents the dot product:

$$\boldsymbol{\mu} = \boldsymbol{f}_\mu \otimes \boldsymbol{\gamma}. \tag{6}$$

New experiences are introduced in the memory $\boldsymbol{\Psi}$ through the modulating variable $\boldsymbol{\mu}$ using a *tanh* function (Step 3.2 in Figure 2):

$$\boldsymbol{\Psi} = tanh\left(\boldsymbol{W}_{\Psi}\left(\boldsymbol{\mu} + (\boldsymbol{f}_{\psi} \otimes \Psi)\right) + \boldsymbol{b}_{\Psi}\right), \tag{7}$$

where $\boldsymbol{W}_{\Psi}$ and $\boldsymbol{b}_{\Psi}$ are the memory weight matrix and bias, respectively.

Contextual Inference. This step is responsible for adding contextual information to BioNeRF. We generate two new embeddings $\boldsymbol{h}'_{\Delta}$ and $\boldsymbol{h}'_{c}$ based on filters $\boldsymbol{f}_{\Delta}$ and $\boldsymbol{f}_{c}$, respectively (Step 4 in Figure 2), which further feed two neural models, i.e., $M'_{\Delta} = (\boldsymbol{W}'_{\Delta}, \boldsymbol{b}'_{\Delta})$ and $M'_{c} = (\boldsymbol{W}'_{c}, \boldsymbol{W}'_{c})$, accordingly:

$$\boldsymbol{h}'_{\Delta} = [\boldsymbol{\Psi} \otimes \boldsymbol{f}_{\Delta}, \mathbf{l}], \tag{8}$$

and

$$\boldsymbol{h}'_{c} = [\boldsymbol{\Psi} \otimes \boldsymbol{f}_{c}, \mathbf{d}]. \tag{9}$$

Subsequently, M'_{Δ} outputs the volume density Δ, while color information $\boldsymbol{c}$ is predicted by M'_{c}, ending up in the final predicted pixel information $(\Delta, \boldsymbol{c})$, further used to compute the loss function.

Loss Function. Let $r : \Re \times \Re^3 \rightarrow \Re^3$ be a volume rendering technique [2] that computes the pixel color given the volume density and the color. The BioNeRF loss function is defined as follows:

$$\mathcal{L} = MSE(r(\Delta, \boldsymbol{c}), \boldsymbol{g}), \tag{10}$$

where $MSE(\cdot)$ is the mean squared error function and $\boldsymbol{g}$ corresponds to the ground truth pixel color.

The error is then back-propagated to the model's previous layers/steps to update its set of weights ($\boldsymbol{W} = \{\boldsymbol{W}_{\Delta}, \boldsymbol{W}_{c}, \boldsymbol{W}_{\psi}, \boldsymbol{W}_{\mu}, \boldsymbol{W}_{\gamma}, \boldsymbol{W}_{\Psi}, \boldsymbol{W}'_{\Delta}, \boldsymbol{W}'_{c}\}$) and biases $\boldsymbol{b} = \{\boldsymbol{b}_{\Delta}, \boldsymbol{b}_{c}, \boldsymbol{b}_{\psi}, \boldsymbol{b}_{\mu}, \boldsymbol{b}_{\gamma}, \boldsymbol{b}_{\Psi}, \boldsymbol{b}'_{\Delta}, \boldsymbol{b}'_{c}\}$.

5. Methodology

This section provides information regarding the datasets employed in this work and the configuration adopted in the experimental setup.

5.1. Datasets

We conducted experiments over two well-known datasets concerning view synthesis, i.e., Realistic Synthetic 360° [1] and LLFF [9]:

5.1.1. Realistic Synthetic 360°.

Also known as Blender or NeRF Synthetic, the Realistic Synthetic 360° comprises eight object scenes with intricate geometry and realistic non-Lambertian materials. Six of these objects are rendered from viewpoints tested on the upper hemisphere, while the remaining two come from viewpoints sampled on a complete sphere, all at a resolution of 800×800 pixels. Figure 3 (top) displays an instance image from each scene that compose the dataset.

5.1.2. Local Light Field Fusion (Real).

This paper considers 8 scenes from the LLFF Real dataset, which comprises 24 scenes captured from handheld cellphones with $20 - 30$ images each. The authors used a COLMAP structure from motion [24] implementation to compute the poses. Figure 3 (bottom) provides a sample image from the scenes used in this work concerning this dataset.

Figure 3: Scene sample images extracted from Realistic Synthetic 360° (top) and LLFF (bottom) datasets.

5.2. Experimental Setup

The experiments conducted in this work aim to evaluate the behavior of BioNeRF in the context of scene-view synthesis against state-of-the-art methods. In BioNeRF, the camera 3D coordinates feed two neural models, i.e.,

blocks of dense layers with ReLU activations, each comprising 3 layers with $h = 256$ hidden neurons. Notice that the number of layers and hidden neurons were empirically selected based on values adopted in the original NeRF model. Using separate blocks resembles the biological brain's efficiency in fusing information from multiple sources. Further, the mechanism described in Section 4 is performed to generate the filters and update the memory $\Psi \in \mathbb{R}^{z \times h}$, where $z = 8,192$ is the number of directional rays processed in parallel.

In the sequence, the memory is used as a context to the consecutive blocks, being concatenated with the camera 3D coordinates to feed the second neural model M'_{Δ}, which comprises two dense layers with 256 neurons and an output layer with a single neuron to predict the volume density Δ. Similarly, the memory is concatenated with the 2D viewing direction (θ, ϕ) to feed the second model M'_{c}, which is composed of a dense layer with 128 neurons and an output layer with three neurons to predict the directional emitted color. The network parameters were selected empirically based on the methodology employed in [1], consisting of the Adam optimizer with a learning rate of $5e-4$ during $400k$ updates.

Additionally, three quality measures were considered to evaluate the models: (i) the Peak Signal Noise Ratio (PSNR), (ii) the Structural Similarity Index Measure (SSIM), and (iii) the Learned Perceptual Image Patch Similarity (LPIPS). PSNR defines the ratio between the maximum power of a signal and the noise that affects the signal representation and is used to quantify a signal's reconstruction. SSIM predicts the perceived quality of digital images, considering its degradation as a change in the structural information. Lastly, LPIPS computes the similarity between the activations of two image patches for some predefined network, presenting itself as an adequate metric to match human perception.

The experiments were conducted on an Ubuntu 18 system running on a 2x Intel® Xeon Bronze 3104 processor with 62GB of memory and an NVIDIA® Tesla T4 GPU. The code was implemented using Python and the PyTorch

framework and is available on GitHub[2].

6. Experiments

This section evaluates BioNeRF concerning view synthesis against state-of-the-art procedures over two datasets comprising real and synthetic images.

6.1. General Evaluation

Table 1 provides the general quantitative results for view synthesis, comparing BioNeRF against state-of-the-art methods considering both synthetic and real image datasets. The analysis is conducted based on three image reconstruction metrics: PSNR and SSIM, whose objective is to be maximized, and LPIPS, whose lower values denote better results. Notice that bolded numbers represent the most accurate results.

The results confirm the effectiveness of BioNeRF since it could obtain the best outcomes for all three metrics considering the LLFF dataset, which comprises 8 scenes from a real environment. BioNeRF also obtained the lowest LPIPS value over the synthetic dataset, outperforming all state-of-the-art results in this context. Such results are highly positive since LPIPS correspond to human perceptual judgments far better than the commonly used metrics like PSNR and SSIM [25].

The win-win results of BioNeRF are primarily due to its memory and context components. Since NeRF relies on the network weights to store the 3D representation of a scene, it makes sense to introduce a memory mechanism that is updated by keeping more relevant information and discarding what is unessential. Further, using such memory as a context forces the model to associate correlated features and produce coherent view-dependent output colors and volume densities, thus improving its performance.

[2]Hidden due to blind review policy

[2]Values marked with − describe results not provided by the authors.

Table 1: Quantitative results considering both synthetic and real image datasets.

	Realistic Synthetic 360°[1]			**LLFF (Real)[9]**		
Method	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
SRN [11] (NeurIPS'19)	22.26	0.846	0.170	22.84	0.668	0.378
NV [10] (ACM ToG 19)	26.05	0.893	0.160	–	–	–
LLFF [9] (ACM ToG 2019)	24.88	0.911	0.114	24.13	0.798	0.212
NSVF [12] (NeurIPS'20)	31.75	0.964	0.047	–	–	–
NeRF [1] (ACM 21)	31.01	0.947	0.081	26.50	0.811	0.250
IBRNet [13] (CVPR'21)	28.14	0.942	0.072	26.73	0.851	0.175
Mip-NeRF [14] (ICCV'21)	33.09	0.961	0.043	–	–	–
TensoRF [15] (ECCV'22)	33.14	0.963	0.027	26, 73	0.839	0.124
Plenoxels [16] (CVPR'22)	31.71	0.958	0.049	26, 29	0.839	0.210
Point-NeRF [17] (CVPR'22)	33.31	0.978	0.027	–	–	–
Ref-NeRF [18] (CVPR'22)	**33.99**	0.966	0.038	–	–	–
L2G-NeRF [19] (CVPR'23)	28.62	0.930	0.070	24.54	0.750	0.200
Tetra-NeRF [20] (ICCV'23)	32.53	**0.982**	0.041	–	–	–
SpikingNeRF [21] (arXiv'23)	32.45	0.956	–	–	–	–
BioNeRF (Ours)	31.45	0.953	**0.026**	**27.01**	**0.861**	**0.068**

6.2. Comparisons on Realistic Synthetic 360° dataset.

Table 2 presents results concerning the Realistic Synthetic 360° dataset. Regarding the PSNR measure, one can observe that Ref-NeRF obtained the most accurate average results, achieving higher PSNR values over half of the scenes, while Point-NeRF stood out in two of them, and TensoRF and Tetra-NeRF did well in Lego and Ship, respectively. Concerning the SSIM metric, even though Point-NeRF obtained the best results on six out of eight scenes, Tetra-NeRF achieved the best average accuracy overall due to an excellent performance over the Ship scene.

Table 2: Quantitative results considering each scene from Realistic Synthetic 360° dataset.

PSNR ↑

Method	Avg.	—	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship
NeRF [1] (ACM 21)	31.01	—	33.00	25.01	30.13	36.18	32.54	29.62	32.91	28.65
TensoRF [15] (ECCV'22)	33.14	—	35.76	26.01	33.99	37.41	**36.46**	30.12	34.61	30.77
Plenoxels [16] (CVPR'22)	31.71	—	33.98	25.35	31.83	36.43	34.10	29.14	33.26	29.62
Point-NeRF [17] (CVPR'22)	33.31	—	35.40	**26.06**	**36.13**	37.30	35.04	29.61	35.95	30.97
Ref-NeRF [18] (CVPR'22)	**33.99**	—	**35.83**	25.79	33.91	**37.72**	36.25	**35.41**	**36.76**	30.28
L2G-NeRF [19] (CVPR'23)	28.62	—	30.99	23.75	26.11	34.56	27.71	27.60	30.91	27.31
Tetra-NeRF [20] (ICCV'23)	32.53	—	35.05	25.01	33.31	36.16	34.75	29.30	35.49	**31.13**
BioNeRF (Ours)	31.45	—	34.63	25.66	29.56	37.23	31.82	29.74	33.38	29.57

SSIM ↑

Method	Avg.	—	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship
NeRF [1] (ACM 21)	0.947	—	0.967	0.925	0.964	0.974	0.961	0.949	0.980	0.856
TensoRF [15] (ECCV'22)	0.963	—	0.985	0.937	0.982	0.982	0.983	0.952	0.988	0.895
Plenoxels [16] (CVPR'22)	0.958	—	0.977	0.933	0.890	0.985	0.976	**0.975**	0.980	0.949
Point-NeRF [17] (CVPR'22)	0.978	—	**0.991**	**0.954**	**0.993**	**0.991**	**0.988**	0.971	**0.994**	0.942
Ref-NeRF [18] (CVPR'22)	0.966	—	0.984	0.937	0.983	0.984	0.981	0.983	0.992	0.880
L2G-NeRF [19] (CVPR'23)	0.930	—	0.950	0.900	0.930	0.970	0.910	0.930	0.970	0.850
Tetra-NeRF [20] (ICCV'23)	**0.982**	—	0.990	0.947	0.989	0.989	0.987	0.968	0.993	**0.994**
BioNeRF (Ours)	0.953	—	0.977	0.927	0.965	0.980	0.963	0.957	0.978	0.874

LPIPS ↓

Method	Avg.	—	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship
NeRF [1] (ACM 21)	0.081	—	0.046	0.091	0.044	0.121	0.050	0.063	0.028	0.206
TensoRF [15] (ECCV'22)	0.027	—	**0.010**	0.051	**0.012**	0.013	**0.007**	0.026	**0.009**	0.085
Plenoxels [16] (CVPR'22)	0.049	—	0.031	0.067	0.026	0.037	0.028	0.057	0.015	0.134
Point-NeRF [17] (CVPR'22)	0.049	—	0.023	0.078	0.022	0.037	0.024	0.072	0.014	0.124
Ref-NeRF [18] (CVPR'22)	0.038	—	0.017	0.059	0.019	0.022	0.018	0.022	0.007	0.139
L2G-NeRF [19] (CVPR'23)	0.070	—	0.050	0.100	0.060	0.030	0.060	0.060	0.050	0.130
Tetra-NeRF [20] (ICCV'23)	0.041	—	0.016	0.073	0.023	0.027	0.022	0.056	0.011	0.103
BioNeRF (Ours)	**0.026**	—	0.011	**0.047**	0.017	**0.010**	0.016	**0.018**	0.018	**0.068**

Apart from such results, BioNeRF asserted the average best results considering the LPIPS measure, obtaining the minimum value over four out of eight scenes, i.e., Drums, Hotdog, Materials, and Ship. Further, it virtually broke even with the best results considering the Chair scene, reaching an LPIPS of 0.011 against 0.010 from TensoRF. Such results are very positive since LPIPS is the measure that best corresponds to human perceptual judgments, thus implying better visual appearance. Figure 4 depicts some examples from BioNeRF view synthesis.

Figure 4: Ground-truth (top) and synthetic view (bottom) images generated by BioNeRF regarding four Realistic Synthetic 360° dataset's scenes.

6.3. Comparisons on LLFF Real dataset.

BioNeRF obtained paramount results concerning the LLFF dataset, surpassing the state of the art for all measures, as presented in Table 3. The method received the highest average PSNR value, reaching the best results in five out of eight scenes, while TensoRF performed better in Flower and Horn's scenes, and NeRF obtained the best results in the Room scenes.

Concerning the SSIM and LPIPS measures, BioNeRF achieved the best results overall, outperforming other techniques in every scene. Such results confirm the robustness of BioNeRF and the relevance of the context to steer the

information flow and extract coherent features that lead to more adequate outputs, as well as the memory mechanism responsible for filtering the relevant knowledge and focusing on proper representations.

Table 3: Quantitative results considering each scene from LLFF Real dataset.

					PSNR ↑					
Method	Avg.	—	Fern	Flower	Fortress	Horns	Leaves	Orchids	Room	T-Rex
NeRF [1] (ACM 21)	26.50	—	25.17	27.40	31.16	27.45	20.92	20.36	**32.70**	26.80
TensoRF [15] (ECCV'22)	26.73	—	25.27	**28.60**	31.36	**28.14**	21.30	19.87	32.35	26.97
Plenoxels [16] (CVPR'22)	26.29	—	25.46	27.83	31.09	27.58	21.41	20.24	30.22	26.48
L2G-NeRF [19] (CVPR'23)	24.54	—	24.57	24.90	29.27	23.12	19.02	19.71	32.25	23.49
BioNeRF (Ours)	**27.01**	—	**26.51**	27.89	**32.34**	27.99	**22.23**	**20.80**	30.75	**27.56**
					SSIM ↑					
Method	Avg.	—	Fern	Flower	Fortress	Horns	Leaves	Orchids	Room	T-Rex
NeRF [1] (ACM 21)	0.811	—	0.792	0.827	0.881	0.828	0.690	0.641	0.948	0.880
TensoRF [15] (ECCV'22)	0.839	—	0.814	0.871	0.897	0.877	0.752	0.649	0.952	0.900
Plenoxels [16] (CVPR'22)	0.839	—	0.832	0.862	0.885	0.857	0.760	0.687	0.937	0.890
L2G-NeRF [19] (CVPR'23)	0.750	—	0.750	0.740	0.840	0.740	0.560	0.610	0.950	0.800
BioNeRF (Ours)	**0.861**	—	**0.837**	**0.873**	**0.914**	**0.882**	**0.796**	**0.714**	**0.956**	**0.911**
					LPIPS ↓					
Method	Avg.	—	Fern	Flower	Fortress	Horns	Leaves	Orchids	Room	T-Rex
NeRF [1] (ACM 21)	0.250	—	0.280	0.219	0.171	0.268	0.316	0.321	0.178	0.249
TensoRF [15] (ECCV'22)	0.124	—	0.155	0.106	0.075	0.123	0.153	0.201	0.082	0.099
Plenoxels [16] (CVPR'22)	0.210	—	0.224	0.179	0.180	0.231	0.198	0.242	0.192	0.238
L2G-NeRF [19] (CVPR'23)	0.200	—	0.260	0.170	0.110	0.260	0.330	0.250	0.080	0.160
BioNeRF (Ours)	**0.068**	—	**0.093**	**0.055**	**0.025**	**0.070**	**0.103**	**0.122**	**0.029**	**0.044**

Figure 5 depicts some high-quality views generated from LLFF dataset scenes. Such samples confirm BioNeRF's robustness in generating pleasant images for human perceptual judgments.

Figure 5: Ground-truth (top) and synthetic view (bottom) images generated by BioNeRF regarding four LLFF dataset's scenes.

6.4. Neuronal Activity Analysis

This section briefly analyzes the neuronal activity concerning M_Δ, M_c, M'_Δ, and M'_c. Figure 6 depicts the average rate of neuronal activation per model and their respective standard deviation for both datasets considering the first $10,000$ iterations. In both cases, models M_Δ and M_c present a higher neuronal activity in initial iterations, decreasing as time passes. Such a behavior can be explained by learning more generic features at the beginning of the process and refining them with more intrinsic and subtle features as training converges.

Regarding the M'_Δ and M'_c, there is an opposite behavior on both datasets, i.e., higher activity concerning M'_Δ and lower activity concerning M'_c over Realistic Synthetic 360°, and an inversed conduct considering LLFF. Such a behavior sheds light on the complexity of each dataset: Realistic Synthetic 360° comprises single, artificially modeled objects, consecutively demanding less effort to learn the density of the scenes but requiring more resources to estimate the reflected colors. On the other hand, LLFF comprises scenes composed of several layers of objects and backgrounds, imposing more efforts to learn the scenes densities while alleviating the color estimation burden since the images were captured in a smaller range of angles, resulting in a lighter spectrum of reflected lights.

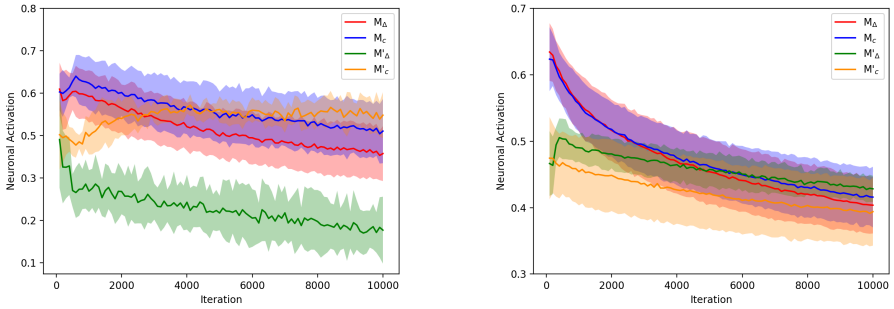

Figure 6: Average neuronal activity and standard deviation for neural models M_Δ, M_c, M'_Δ, and M'_c over the (left) Realistic Synthetic 360° and (right) LLFF datasets.

7. Discussions, Conclusions, and Future Works

Discussions. The BioNeRF mechanism reinforces learning through contextual insights to steer the information flow and extract coherent features. Such context is distilled from a memory state updated iteratively in a process that aims to mimic the biological behavior of forgetting by filtering irrelevant information and learning by introducing novel and more relevant discoveries.

Experiments conducted over two benchmarking datasets for view synthesis, one comprising synthetic and the other real scenes, demonstrate that such a mechanism provided BioNeRF with an outstanding scene representing power, capable of outperforming state-the-art architectures. Regarding such results, BioNeRF achieved the most accuracy overall considering the real scenes dataset over three evaluation metrics for image reconstruction, i.e., PSNR, SSIM, and LPIPS. Regarding the synthetic dataset, BioNeRF obtained the lowest values of LPIPS, outperforming state-of-the-art approaches in a quality measure that best matches human perceptual judgments, thus representing views with better visual quality.

Regarding the model limitations, the primary concern regards the increased number of parameters due to the memory updating system. The problem imposed a more significant challenge due to the limited memory available on the Tesla T4 GPU, i.e., 8GB. In this context, the experiments were conducted using

a batch of directional rays with half the size of the value employed by NeRF to overcome the issue. Such a solution emphasizes the model's robustness since it could obtain state-of-the-art results, even considering such a reduction in instances per iteration.

Conclusions. The paper introduces BioNeRF for view synthesis. This method implements a mechanism inspired by recent discoveries in neuroscience regarding the behavior of pyramidal cells to model memory and a contextual-based information flow. Such a mechanism allowed BioNeRF to outperform state-of-the-art results over two datasets, considering three metrics for image reconstruction and producing images that best match human perceptual judgments.

Future Works. Regarding future works, we aim to improve the model by introducing more elements from a biologically plausible perspective, like spiking neurons, implemented in SpikingNeRF [21]. Additionally, we strive to extend our model to infer in real time and learn from a few shots.

Acknowledgments

The authors are grateful to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil grants #429003/2018-8, #307066/2017-7 and #427968/2018-6, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil grants #2021/05516-1, #2013/07375-0, #2014/12236-1, #2023/10823-6, and #2019/07665-4.

References

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng, NeRF: Representing scenes as neural radiance fields for view synthesis, Communications of the ACM 65 (1) (2021) 99–106.

[2] J. T. Kajiya, B. P. Von Herzen, Ray tracing volume densities, ACM SIGGRAPH Computer Graphics 18 (3) (1984) 165–174.

[3] K. P. Körding, P. König, Supervised and unsupervised learning with two sites of synaptic integration, Journal of computational neuroscience 11 (2001) 207–215.

[4] A. Adeel, Conscious multisensory integration: introducing a universal contextual field in biological and deep artificial neural networks, Frontiers in Computational Neuroscience 14 (2020) 15.

[5] J. Kay, D. Floreano, W. A. Phillips, Contextually guided unsupervised learning using local multivariate binary processors, Neural Networks 11 (1) (1998) 117–140.

[6] L. A. Passos, J. P. Papa, J. Del Ser, A. Hussain, A. Adeel, Multimodal audio-visual information fusion using canonical-correlated graph neural network for energy-efficient speech enhancement, Information Fusion 90 (2023) 1–11.

[7] A. Adeel, M. Franco, M. Raza, K. Ahmed, Context-sensitive neocortical neurons transform the effectiveness and efficiency of neural information processing, arXiv preprint arXiv:2207.07338.

[8] L. A. Passos, J. P. Papa, A. Hussain, A. Adeel, Canonical cortical graph neural networks and its application for speech enhancement in audio-visual hearing aids, Neurocomputing 527 (2023) 196–203.

[9] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, A. Kar, Local light field fusion: Practical view synthesis with prescriptive sampling guidelines, ACM Transactions on Graphics 38 (4) (2019) 1–14.

[10] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, Y. Sheikh, Neural volumes: learning dynamic renderable volumes from images, ACM Transactions on Graphics 38 (4) (2019) 1–14.

[11] V. Sitzmann, M. Zollhöfer, G. Wetzstein, Scene representation networks: Continuous 3d-structure-aware neural scene representations, Advances in Neural Information Processing Systems 32.

[12] L. Liu, J. Gu, K. Zaw Lin, T.-S. Chua, C. Theobalt, Neural sparse voxel fields, Advances in Neural Information Processing Systems 33 (2020) 15651–15663.

[13] Q. Wang, Z. Wang, K. Genova, P. P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, T. Funkhouser, Ibrnet: Learning multi-view image-based rendering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4690–4699.

[14] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, 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, 2021, pp. 5855–5864.

[15] A. Chen, Z. Xu, A. Geiger, J. Yu, H. Su, TensoRF: Tensorial radiance fields, in: European Conference on Computer Vision, Springer, 2022, pp. 333–350.

[16] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, A. Kanazawa, Plenoxels: Radiance fields without neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5501–5510.

[17] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, U. Neumann, Point-nerf: Point-based neural radiance fields, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5438–5448.

[18] D. Verbin, P. Hedman, B. Mildenhall, T. Zickler, J. T. Barron, P. P. Srinivasan, Ref-nerf: Structured view-dependent appearance for neural radiance

fields. in 2022 ieee, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5481–5490.

[19] Y. Chen, X. Chen, X. Wang, Q. Zhang, Y. Guo, Y. Shan, F. Wang, Local-to-global registration for bundle-adjusting neural radiance fields, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8264–8273.

[20] J. Kulhanek, T. Sattler, Tetra-NeRF: Representing neural radiance fields using tetrahedra, arXiv preprint arXiv:2304.09987.

[21] X. Yao, Q. Hu, T. Liu, Z. Mo, Z. Zhu, Z. Zhuge, J. Cheng, Spiking nerf: Making bio-inspired neural networks see through the real world, arXiv preprint arXiv:2309.10987.

[22] S. Grossberg, A canonical laminar neocortical circuit whose bottom-up, horizontal, and top-down pathways control attention, learning, and prediction, Frontiers in Systems Neuroscience 15.

[23] F. Capone, M. Paolucci, F. Assenza, N. Brunelli, L. Ricci, L. Florio, V. Di Lazzaro, Canonical cortical circuits: current evidence and theoretical implications, Neuroscience and Neuroeconomics 5 (2016) 1–8.

[24] J. L. Schonberger, J.-M. Frahm, Structure-from-motion revisited, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 4104–4113.

[25] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.