UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Faculdade de Ciências
Câmpus Bauru

POSTDOCTORAL SCIENTIFIC REPORT

BIONERF: BIOLOGICALLY PLAUSIBLE NEURAL RADIANCE FIELDS FOR VIEW SYNTHESIS

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - 2023/10823-6

BAURU
February, 2025

DR. LEANDRO APARECIDO PASSOS JÚNIOR

BIONERF: BIOLOGICALLY PLAUSIBLE NEURAL RADIANCE FIELDS FOR VIEW SYNTHESIS

Supervisor: Prof. Dr. João Paulo Papa

São Paulo State University (UNESP)

School of Sciences, Bauru

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - 2023/10823-6

BAURU

February, 2025

List of Figures

List of Tables

Contents

Abstract

This scientific report 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.Besides, the proposed architecture was extensively evaluated on two benchmark datasets containing synthetic and real images. Experimental results demonstrate that BioNeRF outperforms state-of-the-art methods in metrics such as PSNR, SSIM, and LPIPS, highlighting the model's robustness in generating images that more accurately reflect human perception.

1 Introduction

Neural Radiance Fields (NeRF) (**??**) 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 (**??**), accumulating those colors and densities into a 2D image.

Together with other machine learning (ML)-based approaches, NeRF contributes to a paradigm change in the archetype of modern society, performing tasks inconceivable a decade ago. Notwithstanding, such solutions come at a high computational cost and energy demand, consuming $85 - 134$ Terawatt hours (TWh) (**??**), which corresponds to 0.5% of the global energy consumption, apart from other environmental impacts related to water consumption and carbon footprint. Additionally, recent studies claim that the current data-hungry-based learning has reached its maximum capacity, and its predictive performance is about to saturate. In this context, biologically plausible algorithms are gaining notorious popularity nowadays since they provide theoretical insights towards more efficient ML applications.

Despite the brain-inspired notions employed in Perceptrons and artificial neural networks (ANN) design, the analogies are rough approximations associated with neurons that do not correspond properly to the real behavior of such cells. Biologically plausible studies aim to implement recent discoveries from neuroscience, modeling the behavior observed in brain cells, which are capable of instantly processing a huge amount of information while consuming a derisory quantity of energy, to improve learning capacities and computational efficiency while decreasing energy consumption.

Apart from such advantages, biologically plausible studies inspired by the principles of pyramidal cells (**??**), especially concerning the idea of context to steer the information flow (**????**) and integrated memory, responsible for providing additional context based on past experiences (**????**), is intrinsically correlated with some of NeRF's fundamental concepts regards encoding complex real-world 3D scene representation in the parameters of a neural

network since 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.

Contextually guided learning in biologically plausible approaches provides efficient instruments to extract latent meaning from data by integrating local features with global contextual perception, enabling more efficient pattern recognition and adaptation to complex data structures (**??**). Such a paradigm integrates conscious multisensory context-aware processing mechanisms to leverage perception and decision-making in complex neural network scenarios. The concept allows for efficient adaptation to environmental stimuli and provides meaningful insights into biologically plausible mechanisms for learning.

Recent works implemented biologically plausible variants of Neural Radiance Fields considering Spiking Neural Networks (SNNs), like the Spiking-NeRF (**??**), which converts the NeRF network to the spiking domain and Spiking NeRF (**??**), which associates the intensity flow on retina imaging process with the rendering process, which is modeled as the temporal accumulation process of spiking neurons. Guo et al. (**??**) proposed the Spike-NeRF, the first view synthesis model derived from spike data captured by neuromorphic sensors, namely spike cameras. The idea was further improved by Zhu et al. (**??**) in the SpikeNeRF, a model designed to tackle the challenges related to the assumption of optimal illumination in spike cameras in real-world scenarios.

This report introduces BioNeRF, a neural radiance field approach that explores biologically plausible concepts to obtain state-of-the-art results in the context of novel view synthesis. Further, it briefly describes the works published or submitted in which the scholarship holder contributed to the development process during the project period.

1.1 Problem

SNN-based approaches consider the energy efficacy provided by spiking behavior observed in the neural information flow to implement energy-efficient computational models. It is also very useful for executing specific tasks, like capturing the asynchronous luminance intensity from independent photons obtained from spike streams captured by spike cameras. Notwithstanding, SNNs do not consider distinct aspects observed in biological neurons, and the paradigm lacks when considering some basic principles related to contextual learning and memory formation, which are essential components of the biological learning system.

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 (**??**),

especially concerning the idea of context to steer the information flow (**??????**) and integrated memory, responsible for providing additional context based on past experiences (**????**).

1.2 Objectives

The objective of this scientific report is to present 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.

1.3 Report Structure

This report is described as follows. Academic activities are described in Chapter 2. Chapter 3 presents the related works, while Chapter 4 provides a theoretical background regarding the NeRF concepts and introduces the BioNeRF. Chapters 5 and 6 describe the methodology and present the experimental results and discussions, respectively. Finally, Chapter 7 states conclusions and future works, while Chapter 8 provides the supervisor's comments.

2 Activities

The researcher has developed several innovative works in machine learning during the project period in cooperation with other group members and worldwide collaborators. Further, the researcher also participated in the development of the work of several students, running experiments and supporting the document writing. A list of published and submitted papers is provided below.

2.1 Published Papers

2.1.1 Journals

- **ACM Computing Surveys** - *Facial Expression Analysis in Parkinson's Disease Using Machine Learning: A Review (**??**)*.

  Deep Learning algorithms have permitted a revolution in practically all fields of science and daily life activities over the previous decade. The Medicine field, in particular, gathered many benefits from such achievements since state-of-the-art models aided medical diagnosis and decision-making in many applications. Furthermore, with the advent of the telemedicine field and the proliferation of smart devices capable of high-quality video transmission, the computer-aided diagnosis became a remarkable tool to assist doctors and clinicians. Despite such accomplishments, a recent approach for image-based diagnosis, namely facial expression analysis, attracted the attention of many researchers. However, there are only a few projects concerning Deep Learning medical-based applications related to the topic. Therefore, this work proposes to review, assess and compare the current machine learning and deep learning techniques for analysis of emotional facial expressions of people with Parkinson's. Our work focuses only on approaches based on image and video input.

- **Applied Soft Computing** - *EEG microstates classification empowered with optimum-path forest using different distance measurements (**??**)*.

  Electroencephalography (EEG) is considered one of the most important tests for neurological disease diagnosis, whose signals comprise microstate information regarding the spatiotemporal characteristics of human brain activity. However, interpreting and identifying such signals denotes a complex and time-consuming activity, thus motivating the use of automated approaches like machine learning techniques in many studies.

Recent research has presented several techniques for detecting neurological disorders through analyzing EEG microstates. However, this area has room for advancements and improvements, using distinct methods and more in-depth scrutiny. Therefore, this paper proposes a new method for EEG signal classification through microstate analysis for diagnosing neurological diseases using a graph-based algorithm, namely the Optimum-Path Forest (OPF) classifier. Experiments were conducted over features extracted from EEG microstates obtained from the Temple University Hospital Abnormal EEG Corpus (TUAB) and the Schizophrenia EEG datasets. Such features are further processed using principal component analysis, t-distributed stochastic neighbor embedding, and uniform manifold approximation and projection dimensionality reduction techniques to improve computational efficiency and increase the classifier's performance. Furthermore, this work evaluates 44 distance measurements in OPF's graph modeling context. Finally, the experimental results show that with a reduced set of features extracted from the microstates and an appropriate distance measure, it is possible to obtain accuracy values equivalent to 100% with low processing time compared to raw data. In addition, the k-nearest neighbors and support vector machine classifiers were used to compare the experimental results.

- **Neurocomputing** - *A binary particle swarm optimization-based pruning approach for environmentally sustainable and robust CNNs (**??**)*.

  Deep Convolutional Neural Networks (CNNs), continue to demonstrate remarkable performance across various tasks. However, their computational demands and energy consumption present significant drawbacks, restricting their practical deployment and contributing to a substantial carbon footprint. This paper addresses this challenge by proposing a novel method named Binary Particle Swarm Optimization Layer Pruner (BPSO-LPruner), aimed at achieving substantial computational reduction and mitigating environmental impact during CNN inference. BPSO-LPruner utilizes a constrained Binary Particle Swarm Optimization for CNN layer pruning, integrating a masked-bit strategy and a new population initialization strategy to enhance search performance. We illustrate the effectiveness of our method in reducing model computational costs and carbon footprint emissions while improving performance across multiple models (VGG16, VGG19, DenseNet-40, ResNet18, ResNet20, ResNet34, ResNet44, ResNet56, ResNet110, ResNet50, and MobileNetv2) and diverse datasets (CIFAR-10, CIFAR-100, Tiny-ImageNet, COVID-19 X-ray dataset). Promising results underscore the performance of the proposed method. Additionally, we demonstrate that layer pruning yields benefits beyond enhanced computational performance. Our experimentation reveals that BPSO-LPruner enhances the model's reliability and robustness by effectively addressing variations in input data, inherent ambiguity in model parameters, and adversarial images.

- **Applied Soft Computing** - *A Comprehensive Study Among Distance Measures On Supervised Optimum-Path Forest Classification (**??**).*

  Supervised pattern classification relies on a labeled training set to learn decision boundaries that separate samples from different classes. Such samples can be either weakly- or reliably-labeled; in the first case, one can employ techniques specifically designed to cope with uncertainty during labeling, and in the other scenario, it relies on numerous alternatives, including metric learning. Pattern classifiers usually adopt the Euclidian distance to compare samples and assess their proximity, but this implies the feature space is embedded in a plane. However, samples are embedded in curved spaces for some applications, although not straightforward to prove. In this manuscript, we assessed the performance of the Optimum-Path Forest (OPF) classifier under different distance functions, which are used to weigh arcs among samples, for a graph encoding the feature space. This work compared 47 distance measures applied to the OPF classifier considering 22 datasets, plus Decision Trees, Logistic Regression, and Support Vector Machines. The experiments highlighted that OPF is user-friendly when handling distance measures and can obtain better accuracies in some situations than its standard (Euclidian) counterpart and the classifiers mentioned above. On the other hand, time-consuming distance calculations may affect OPF's efficiency during inference.

- **Digital Biomarkers** - *Video Assessment to Detect Amyotrophic Lateral Sclerosis (**??**).*

  Weakened facial movements are early stage symptoms of Amyotrophic Lateral Sclerosis (ALS). It is generally detected based on changes in facial expressions, but large differences between individuals can lead to subjectivity in the diagnosis. We have proposed a computerised analysis of facial expression videos. We developed a method based on the action units obtained from facial expression videos to differentiate between ALS patients and healthy people and identified the action units and facial expressions that give the best results. The Toronto NeuroFace Dataset which has nine facial expression and verbalization tasks for healthy and ALS patients, was used in this study. The best classification results area under the roc curve (AUC) was 0.94 obtained for the Spread expression. This pilot study shows the potential of using computerized facial expression analysis based on action units for the detection of ALS conditions.

- **Expert Systems** - *A review of deep learning-based approaches for deepfake content detection (**??**).*

  Recent advancements in deep learning generative models have raised concerns as they can create highly convincing counterfeit images and videos. This poses a threat to

people's integrity and can lead to social instability. To address this issue, there is a pressing need to develop new computational models that can efficiently detect forged content and alert users to potential image and video manipulations. This paper presents a comprehensive review of recent studies for deepfake content detection using deep learning-based approaches. We aim to broaden the state-of-the-art research by systematically reviewing the different categories of fake content detection. Furthermore, we report the advantages and drawbacks of the examined works, and prescribe several future directions towards the issues and shortcomings still unsolved on deepfake detection.

- **Biomedical Signal Processing and Control** - *Robust deep learning for eye fundus images: Bridging real and synthetic data for enhancing generalization (**??**)*.

  Deep learning applications for assessing medical images are limited because the datasets are often small and imbalanced. The use of synthetic data has been proposed in the literature, but neither a robust comparison of the different methods nor generalizability has been reported. Our approach integrates a retinal image quality assessment model and StyleGAN2 architecture to enhance Age-related Macular Degeneration (AMD) detection capabilities and improve generalizability. This work compares ten different Generative Adversarial Network (GAN) architectures to generate synthetic eye-fundus images with and without AMD. We combined subsets of three public databases (iChallenge-AMD, ODIR-2019, and RIADD) to form a single training and test set. We employed the STARE dataset for external validation, ensuring a comprehensive assessment of the proposed approach. The results show that StyleGAN2 reached the lowest Fréchet Inception Distance (166.17), and clinicians could not accurately differentiate between real and synthetic images. ResNet-18 architecture obtained the best performance with 85% accuracy and outperformed the two human experts (80% and 75%) in detecting AMD fundus images. The accuracy rates were 82.8% for the test set and 81.3% for the STARE dataset, demonstrating the model's generalizability. The proposed methodology for synthetic medical image generation has been validated for robustness and accuracy, with free access to its code for further research and development in this field.

- **Computer Methods and Programs in Biomedicine** - *Facial expressions to identify post-stroke: A pilot study (**??**)*.

  Background and Objective:
  Timely stroke treatment can limit brain damage and improve outcomes, which depends on early recognition of the symptoms. However, stroke cases are often missed by the first respondent paramedics. One of the earliest external symptoms of stroke is based on facial expressions.

Methods:
We propose a computerized analysis of facial expressions using action units to distinguish between Post-Stroke and healthy people. Action units enable analysis of subtle and specific facial movements and are interpretable to the facial expressions. The RGB videos from the Toronto Neuroface Dataset, which were recorded during standard orofacial examinations of 14 people with post-stroke (PS) and 11 healthy controls (HC) were used in this study. Action units were computed using XGBoost which was trained using HC, and classified using regression analysis for each of the nine facial expressions. The analysis was performed without manual intervention.

Results:
The results were evaluated using leave-one-our validation. The accuracy was 82% for Kiss and Spread, with the best sensitivity of 91% in the differentiation of PS and HC. The features corresponding to mouth muscles were most suitable.

Conclusions:
This pilot study has shown that our method can detect PS based on two simple facial expressions. However, this needs to be tested in real-world conditions, with people of different ethnicities and smartphone use. The method has the potential for a computerized assessment of the videos for use by the first respondents using a smartphone to perform screening tests, which can facilitate the timely start of the treatment.

2.1.2 Conferences

- **20th International Conference on Computer Vision Theory and Applications (VISAPP 2025)** - *Customized Atrous Spatial Pyramid Pooling with Joint Convolutions for Urban Tree Segmentation (**??**)*.

Urban trees provide several benefits to the cities, including local climatic regulation and better life quality. Assessing the tree conditions is essential to gather important insights related to its biomechanics and the possible risk of falling. The common strategy is ruled by fieldwork campaigns to collect the tree's physical measures like height, the trunk's diameter, and canopy metrics for a first-glance assessment and further prediction of the possible risk to the city's infrastructure. The canopy and trunk of the tree play an important role in the resistance analysis when exposed to severe windstorm events. However, fieldwork analysis is laborious and time-expensive because of the massive number of trees. Therefore, strategies based on computational analysis are highly demanded to promote a rapid assessment of tree conditions. This paper presents a deep learning-based

approach for semantic segmentation of the trunk and canopy of trees in images acquired from the street-view perspective. The proposed strategy combines convolutional modules, spatial pyramid pooling, and attention mechanism into a U-Net-based architecture to improve the prediction capacity. Experiments performed over two image datasets showed the proposed model attained competitive results compared to previous works employing large-sized semantic segmentation models.

- **The 27th International Conference on Pattern Recognition (ICPR) 2024** - *Graph Matching Networks Meet Optimum-Path Forest: How to Prune Ensembles Efficiently (**??**)*.

  Ensemble pruning enhances the efficiency and predictive performance of models by selecting a subset of representative classifiers, preventing redundancy and ensuring diversity in classification tasks. The Optimum-Path Forest, a stable and efficient graph-based framework, offers versatile supervised and unsupervised capabilities in a wide range of machine-learning applications. The supervised version provides remarkable results with a simple graph-based structure produced by a training process conducted over a single dataset. However, predictive capabilities can be improved towards better results by combining the ensemble diversity capabilities with a set of predictions provided by different supervised Optimum-Path Forest instances constituting an ensemble model. In addition, a promising approach involves using graph-matching networks as the underlying pruning method to identify similar graphs assembled by the training process of the supervised Optimum-Path Forest algorithm. This paper introduces an innovative ensemble pruning strategy using graph-matching networks to identify similar Optimum-Path Forest instances. The strategy selectively chooses representative instances that excel in predictive tasks from clusters generated by the unsupervised version of the Optimum-Path Forest algorithm. Results demonstrate competitive performance and comparability to state-of-the-art pruning algorithms revealed by experiments conducted over fifteen public datasets, thus encouraging further exploration of graph-matching networks in ensemble pruning research.

- **The 27th International Conference on Pattern Recognition (ICPR) 2024** - *A Quantum-inspired Approach to Estimate Optimum-Path Forest Prototypes based on the Traveling Salesman Problem (**??**)*.

  Quantum mechanics emerge as a promise for the future of computing, broadening the horizons for solutions concerning complex tasks, e.g., NP-hard problems. Alongside quantum computing, machine learning has become indispensable. This paper explores the potential integration of quantum computing principles into the Optimum-Path Forest (OPF), a graph-based framework comprised of solutions for machine learning,

optimization, and image processing. We are particularly interested in the supervised OPF approach, which elects the most representative samples for each class, aka prototypes, as the connected samples from different classes in a minimum spanning tree (MST) computed over the training set. By harnessing quantum parallelism and superposition, this paper introduces a new approach to identifying prototypes employing a quantum-based Traveler Salesman Problem (TSP) algorithm, which provides an alternative to computing MSTs and yields a hybrid version of the OPF classifier. The experiments on established datasets demonstrated the promising potential of this approach while also underscoring the necessity for further research in this field.

- **The 31st International Conference on Systems, Signals and Image Processing (IWSSIP) 2024** - *Enhancing Cyberattack Detection in IoT Environments through Advanced Resampling Techniques (**??**)*.

  As the world increasingly relies on emerging technologies like the Internet of Things, there is a growing demand for large-scale distributed software to perform various tasks, facilitate communication, and share resources between devices. However, the implementation and configuration of such softwares can create openings for intrusion attacks through vulnerabilities and weaknesses. To address this concern, we have developed a machine-learning solution that leverages Logistic Regression and Random Forest classifiers with data balancing techniques to classify intrusion attacks accurately. Our experiments demonstrated the most effective results using the Random Forest classifier and oversampling techniques.

- **The 31st International Conference on Systems, Signals and Image Processing (IWSSIP) 2024** - *Ensemble Diversity Pruning on Cybersecurity: Optimizing Intrusion Detection Systems (**??**)*.

  Recent works have shown that intrusion detection systems based on ensemble learning can reduce the misclassification of malicious traffic on computer networks. However, searching for a suitable combination of classifiers often imposes a challenging task that demands a high computational cost. This work proposes a comprehensive application of Diversity Pruning to tackle suck a task, improving the results obtained in previous works and extending the experimental analysis by introducing four new datasets to evaluate the process: KDD-CUP'99, NSL-KDD, UNSW-NB15, and ISCX- IDS-2012. The results show a significant computational cost reduction and considerable increases the attack detection rates. The proposed approach reduced the classification errors by 18.82%, 26.58%, 22.93%, and 52.34% and the training time by 98.69%, 98.97%, 98.41%, and 98.41% for the datasets mentioned above, respectively.

- **19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP) 2024** - *Convolutional Neural Networks and Image Patches for Lithological Classification of Brazilian Pre-salt Rocks (**??**)*.

  Lithological classification is a process employed to recognize and interpret distinct structures of rocks, providing essential information regarding their petrophysical, morphological, textural, and geological aspects. The process is particularly interesting regarding carbonate sedimentary rocks in the context of petroleum basins since such rocks can store large quantities of natural gas and oil. Thus, their features are intrinsically correlated with the production potential of an oil reservoir. This paper proposes an automatic pipeline for the lithological classification of carbonate rocks into seven distinct classes, comparing nine state-of-the-art deep learning architectures. As far as we know, this is the largest study in the field. Experiments were performed over a private dataset obtained from a Brazilian petroleum company, showing that MobileNetV3large is the more suitable approach for the undertaking.

- **3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC) 2024** - *AVSE-Pruner: Filter Pruning of Audio-Visual Speech Enhancement System using Multi-objective Binary Particle Swarm Optimization (**??**)*.

  This paper optimizes filter pruning as a constrained multi-objective optimization problem using a new binary multi-objective particle swarm optimization with dynamic learning strategies (AVSE-Pruner). AVSE-Pruner aims to balance network performance and computational cost by incorporating dynamic learning strategies to adjust search behavior. Applying AVSE-Pruner, we pruned the baseline model for the Audio-Visual Speech Enhancement (AVSE) Challenge, which enhances speech intelligibility in noisy environments using audio and visual inputs. Our pruned model maintains high performance while significantly reducing computational burden, demonstrating its suitability for real-time embedded applications.

- **3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement (AVSEC) 2024** - *RecognAVSE: An Audio-Visual Speech Enhancement Approach using Separable 3D convolutions and Deep Complex U-Net (**??**)*.

  Audio-visual speech enhancement concerns a multimodal task that aims to provide a clean reconstruction of a speech given visual information and noisy audio signals. The assignment is particularly attractive for medical purposes since it might provide resources to aid deaf individuals, besides being useful for distinct contexts like social interactions in uproarious environments. This paper proposes RecognAVSE, an audio-visual speech

enhancement solution developed for the AVSEC-3 challenge that combines Separable 3D CNNs and Deep Complex U-Nets to create an efficient alternative to tackle the problem. The model learns correlated features between the noisy audio signal and the visual stimuli, thus inferring meaning to the speech by amplifying relevant information and suppressing noise. Experiments over the AVSEC-3 dataset show that RecognAVSE can obtain outstanding results, outperforming the baselines in quantitative and qualitative results.

2.2 Submitted Papers

2.2.1 Journals

- **The Journal of the Acoustical Society of America** - *RecognAVSE-V2: An Improved Cross-Attention-Based Audio-Visual Speech Enhancement Approach*.

  Audio-visual speech enhancement (AVSE) is a subfield of machine learning that aims to improve speech quality based on noisy audio signals while considering the context provided by visual cues to steer the information flow and leverage the learning process. In this context, the COG-MHEAR program promotes an early competition, namely the AVSE Challenge (AVSEC), to promote the development of new solutions and advances toward AVSE methods while fostering the community. This paper introduces the RecognAVSE-V2, an improved and more efficient version of RecognAVSE proposed in the AVSEC 2024, that implements a Time-Synced Cross-Attention Mechanism to exploit the temporal correlations between audio and video features. The method's architecture comprises a video encoder to extract spatiotemporal features from raw video frames, an STFT-based audio encoder to capture spectral features of the audio signal, and a time-synced cross-attention module to align audio and video features. Experimental results conducted over AVSE Challenge and CHiME3 datasets show RecognAVSE-V2 is capable of outperforming the baseline and its prior version in some cases while consisting of a model with reduced complexity, 20 times smaller, whose inference is 26% faster, and training demands a quarter of the total time.

2.2.2 Conferences

- **28th Iberoamerican Congress on Pattern Recognition (CIARP) 2025** - *Towards Efficient Intrusion Detection: Feature Selection in Multiclass CNNs with Numerical Inputs*.

  The last decade has seen exponential data storage volume and network traffic growth. This phenomenon, associated with data representation in high dimensions, contributes to

a complex scenario in which detecting malicious content becomes a harder challenge that intrusion detection systems struggle to attack. This work proposes reducing traffic data dimensionality through dimensionality reduction techniques, namely Principal Component Analysis (PCA) and Chi-Square (χ^2). Further, it considers a Convolutional Neural Network to assess shrunk data contribution in the trade-off optimization between malicious content identification accurateness and computational burden. Experimental results demonstrated the proposed approaches' classification effectiveness and computational efficiency, even considering a CPU-based execution, obtaining virtually the same results of an MLP-based baseline, i.e., 99.63% and 98.02% accuracy for χ^2 and PCA, respectively, against 99.70% obtained through the undiminished data at the cost of significantly higher training time.

- **Interspeech 2025** - *Balancing Efficiency and Reliability: Uncertainty-Aware Pruning for Audio-Visual Speech Enhancement*.

  Audio-Visual Speech Enhancement (AVSE) models improve speech intelligibility but often suffer from high computational complexity, limiting their deployment on edge devices. To overcome this challenge, we propose AWL-MOPSO-AVSE, a Multi-Objective Particle Swarm Optimization framework en7 hanced with Adaptive Weight Learning, which dynamically adjusts convergence and diversity to balance conflicting ob9 jectives. Our method jointly optimizes AVSE models by maximizing speech quality while minimizing epistemic un11 certainty, ensuring reliable enhancement. Simultaneously, we impose strict floating-point operation constraint to maintain computational efficiency. Experimental results confirm that AWL-MOPSO-AVSE effectively reduces computational over15 head without compromising performance, making it ideal for real-time, resource-limited applications.

- **IEEE Research and Technologies for Society and Industry (RTSI) 2025** - *Domain-Invariant Classification of Misplaced Medical Devices in Plain Chest X-Ray Images*.

  Chest X-ray imaging is crucial for verifying the correct placement of medical devices. However, deep learning models often fail to generalize across institutions due to domain shifts caused by variations in imaging protocols and annotation practices. This work addresses the challenge of classification of misplaced medical devices in plain chest X-ray images under domain shift by integrating domain adaptation into the training process. It proposes a framework that combines a Vision Transformer architecture with Deep Adaptation Networks, using the Maximum Mean Discrepancy loss to align the feature distributions between a large public dataset and a smaller, clinically distinct private dataset. Our results demonstrate that domain adaptation enables the detection of previously undetected misplacements in the target domain and enhances performance

on the source domain, suggesting improved feature generalizability. The baseline model achieved high accuracy but failed to identify true positives in the target domain. In contrast, despite having lower precision, the adapted model increased the number of true positives from zero to twelve.

- **4th COG-MHEAR International Audio-Visual Speech Enhancement Challenge (AVSEC-4)** - *Tackling Reverberation and Binaural Data on Audio-Visual Speech Enhancement through RecognAVSE*.

  This paper introduces RecognAVSE, an audio-visual speech enhancement (AVSE) method from the AVSEC-3 challenge, and RecognAVSE-V2, an improved version implementing a Time-Synced Cross-Attention Mechanism. We evaluate these models in the context of the 4th COGMHEAR AVSE Challenge, which introduces difficult conditions such as binaural data, multiple interferers, and reverberation. While RecognAVSE-V2 achieves a 95% reduction in parameters and a 10% decrease in GFlops compared to its predecessor, experiments on the AVSEC-4 dataset reveal that both models struggle with the new complexities. Our analysis indicates that while the simplified architecture offers substantial efficiency gains, its robustness is severely tested in a zero-shot setting against the reverberant and multi-speaker conditions of the challenging new dataset.

3 Related Works

3.1 Neural Radiance Fields

Before NeRF, Mildenhall et al. (**??**) 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. Besides, Sitzmann et al. (**??**) proposed the Scene Representation Networks (SRNs), which employed unsupervised learning and convolutional neural networks to learn rich representations of 3D scenes without explicit supervision.Finally, in 2021, Mildenhall et al. (**??**) 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. (**??**) 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. (**??**) 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. Further, the same authors proposed Mip-NeRF 360 (**??**), an extension of Mip-NeRF that employs online distillation, non-linear scene parameterization, and distortion-based regularizer to tackle problems related to unbounded scenes present in the prior version.

Chen et al. (**??**) 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. (**??**) 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. (**??**) 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. (**??**) 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.

Chen et al. (**??**) 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. (**??**) 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. Meanwhile, Yao et al. (**??**) 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. Finally, Selvaratnam et al. (**??**) proposed the Localised-NeRF, which projects pixel points onto training images to obtain representation on HSV gradient space and further extract features to synthesize novel views. The method employs an attention-driven approach to maintain direction consistency and leverage image-based features to predict the diffuse and specular colors, exhibiting competitive performance with prior NeRF-based models.

3.2 Contextually Guided Learning in Biologically Plausible Approaches

Körding and König (**??**) explore computational learning mechanisms by addressing the primary principles of pyramidal neurons, which comprises (i) a learning rule to steer the sign and magnitude of plasticity; (ii) to multiplex the feedback and feedforward information across multiple layers using different connections; (iii) to align feedback and feedforward connections; and (iv) to attribute the credit feedback information for each neuron. The proposed model considers both dendritic and somatic levels, allowing for efficient adaptation to environmental stimuli and providing meaningful insights into biologically plausible mechanisms for learning and memory formation in artificial and biological neural systems.

A distinct approach conducted by Kay et al. (**??**) investigates local multivariate binary processors to enhance contextual information processing. The proposed model integrates local feature learning with global contextual awareness, enabling more efficient pattern recognition and adaptation to complex data structures. The study demonstrates how contextually guided mechanisms can improve learning efficiency and robustness.

In a similar study, Adeel et al. (**??**) proposed a theoretical framework for conscious multisensory integration by introducing a universal contextual field. The effectiveness of the sensory integration relies on a context-aware processing mechanism, which enhances perception and decision-making. The authors draw parallels between biological cognition and deep learning architectures, offering new perspectives on more human-like multisensory processing artificial intelligence systems.

Recently, Passos et al. (**??**) combined some elements concerning memory formation,

addressed in (**??**), with the contextual integration proposed in (**??**) and (**??**) to proposed a graph-based, biologically plausible solution for audio-visual speech enhancement. The method reinforces the suitability of biologically plausible approaches in complex machine-learning domains and inspired the mechanism implemented in BioNeRF.

4 Biologically Plausible Neural Radiance Fields

This section briefly introduces the main concepts of NeRF, the building block upon which BioNeRF was built. It then describes the report's proposed method, BioNeRF.

4.1 Neural Radiance Fields

Mildenhall et al.(**??**) 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. Figure 1 illustrates the concept.

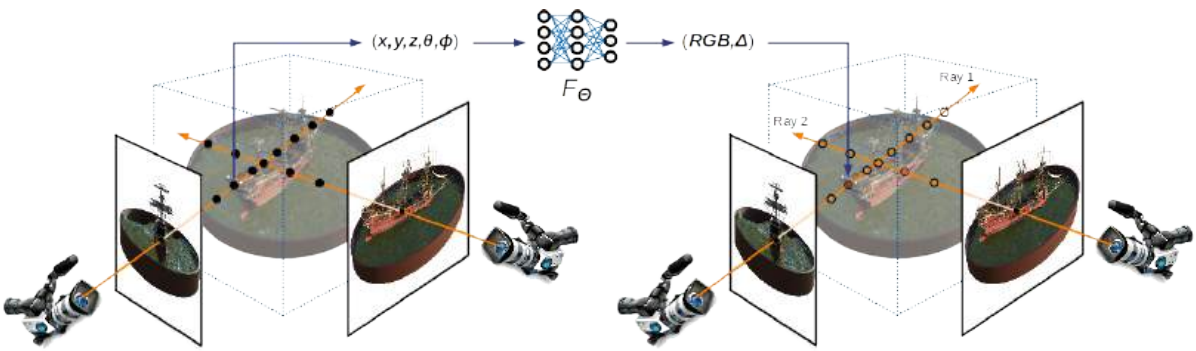

Figure 1 – NeRF synthesizes images by sampling 5D coordinates (location and viewing direction) along camera rays (left) and feeding those locations into an MLP to produce color and volume density (right).

One 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.

4.2 Proposed Approach

This paper introduces BioNeRF, which combines features extracted from both the camera position and direction into a memory mechanism inspired by cortical circuits (**????**). Further, such memory is employed as a context to leverage the information flow into deeper layers. Specifically, the proposed method spans four main steps: (i) Positional Feature Extraction, which is responsible for extracting intrinsic features from the camera position into two distinct flows, one responsible for generating the output densities and the other for the expected RGB colors; (ii) Cognitive Filtering, which computes the filters further employed to suppress irrelevant information while boosting the relevant knowledge originated from memory and camera position; (iii) Memory Updating, comprising filters and operations used to iteratively updating the memory; and (iv) Contextual Inference, responsible for filtering and applying relevant information from memory as a context to the camera position to generate the volume densities (δ) and to the camera direction to produce the output **RGB** colors. Such steps are described in detail below. Further, Figure 2 illustrates the BioNeRF architecture.

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[1,2]. 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:

[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 $\boldsymbol{h}_c$.

[2] Notice such process is inspired in recurrent networks gates' mechanism, like the LSTM (**??**).

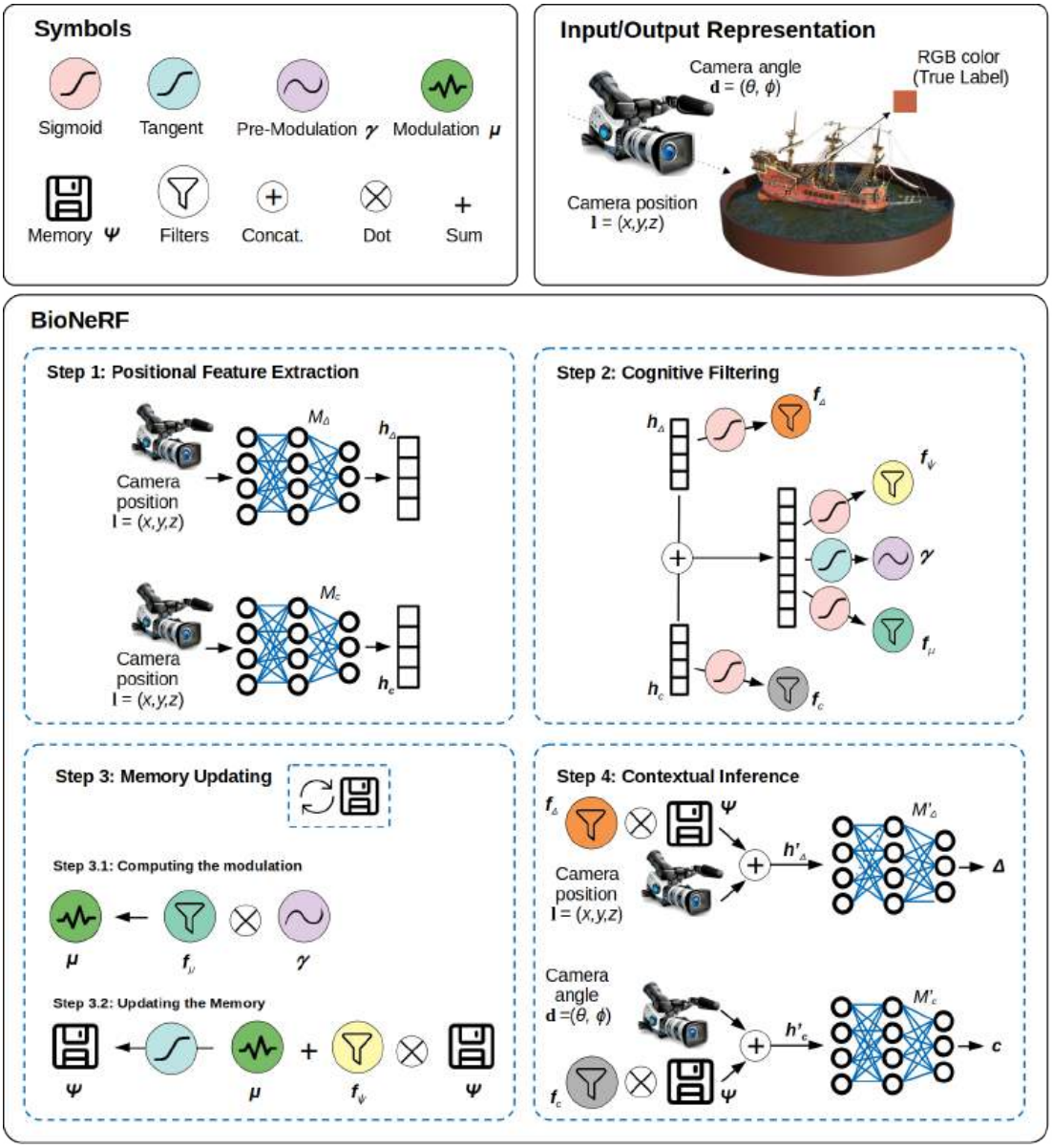

Figure 2 – Biologically Plausible Neural Radiance Fields. The top-left frame describes the symbols, while the top-right frames depict the input/output variables. The bottom block illustrates the model's overall pipeline, comprising four steps. Step 1 describes the **Positional Feature Extraction**, which 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.

$$\boldsymbol{f}_\Delta = \sigma\left(\boldsymbol{W}^f_\Delta h_\Delta + \boldsymbol{b}^f_\Delta\right), \tag{1}$$

$$\boldsymbol{f}_c = \sigma\left(\boldsymbol{W}_c^f h_c + \boldsymbol{b}_c^f\right), \tag{2}$$

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

and

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

where $\boldsymbol{W}_\Delta^f$, $\boldsymbol{W}_c^f$, $\boldsymbol{W}_\psi^f$ and $\boldsymbol{W}_\mu^f$ correspond to the weight matrices for the density, color, memory, and modulation filters, respectively. Additionally, $\boldsymbol{b}_\Delta^f$, $\boldsymbol{b}_c^f$, $\boldsymbol{b}_\psi^f$ and $\boldsymbol{b}_\mu^f$ 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 (**??**) 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 ($\boldsymbol{W} = \{\boldsymbol{W}_\Delta, \boldsymbol{W}_c, \boldsymbol{W}^f_\Delta, \boldsymbol{W}^f_c, \boldsymbol{W}^f_\psi, \boldsymbol{W}^f_\mu, \boldsymbol{W}_\gamma, \boldsymbol{W}_\Psi, \boldsymbol{W}'_\Delta, \boldsymbol{W}'_c\}$) and biases $\boldsymbol{b} = \{\boldsymbol{b}_\Delta, \boldsymbol{b}_c, \boldsymbol{b}^f_\Delta, \boldsymbol{b}^f_c, \boldsymbol{b}^f_\psi, \boldsymbol{b}^f_\mu, \boldsymbol{b}_\gamma, \boldsymbol{b}_\Psi, \boldsymbol{b}'_\Delta, \boldsymbol{b}'_c\}$.

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., Blender (**??**) and LLFF (**??**):

5.1.1 Blender

Also known as NeRF Synthetic, the Blender 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. Figure 3 displays an instance image from each scene that compose the dataset.

Figure 3 – Scene sample images extracted from Blender dataset.

5.1.2 Local Light Field Fusion

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 (**??**) implementation to compute the poses. Figure 4 provides a sample image from the scenes used in this work concerning this dataset.

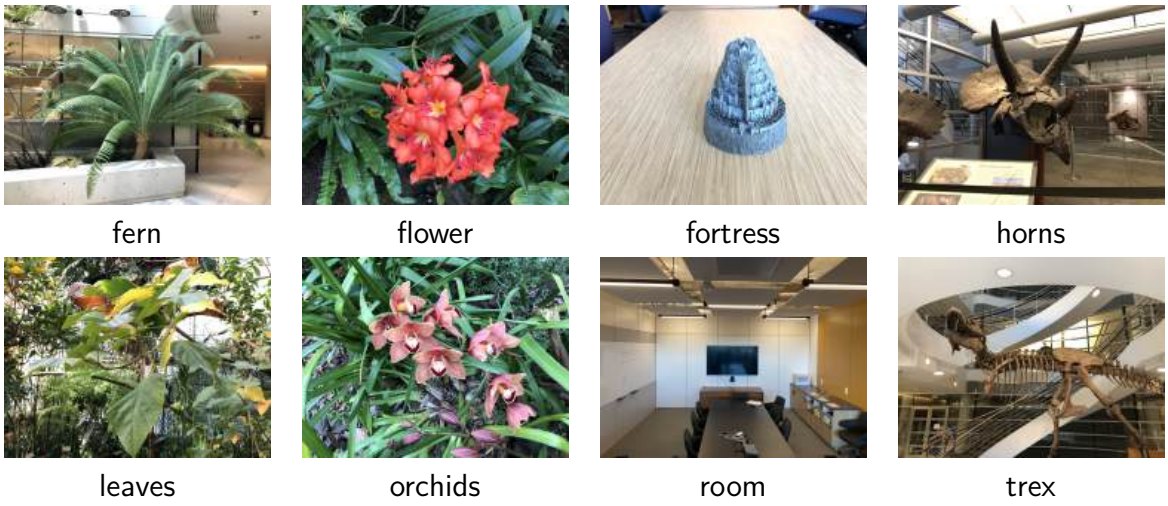

Figure 4 – Scene sample images extracted from LLFF dataset.

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 **??** 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 (**??**), 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[3].

[3] https://github.com/Leandropassosjr/BioNeRF.

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 Visual Insights

This section evaluates the quality of the views generated by the proposed method in terms of visual aspects. Figure 5 compares BioNeRF against three baselines, i.e., NeRF (**??**), Mip-NeRF 360 (**??**), and TensoRF (**??**)[4] considering a Blender dataset's Lego view. Compared to NeRF, BioNeRF presents a result closer to the ground truth, especially if the toy's grid and piston reflections are analyzed in the bottom image. Mip-NeRF 360 shows some black artifacts in the white area on the top image. Regarding TensoRF, BioNeRF shows itself more capable of extracting the reflexible components and reproducing the floor color with better precision.

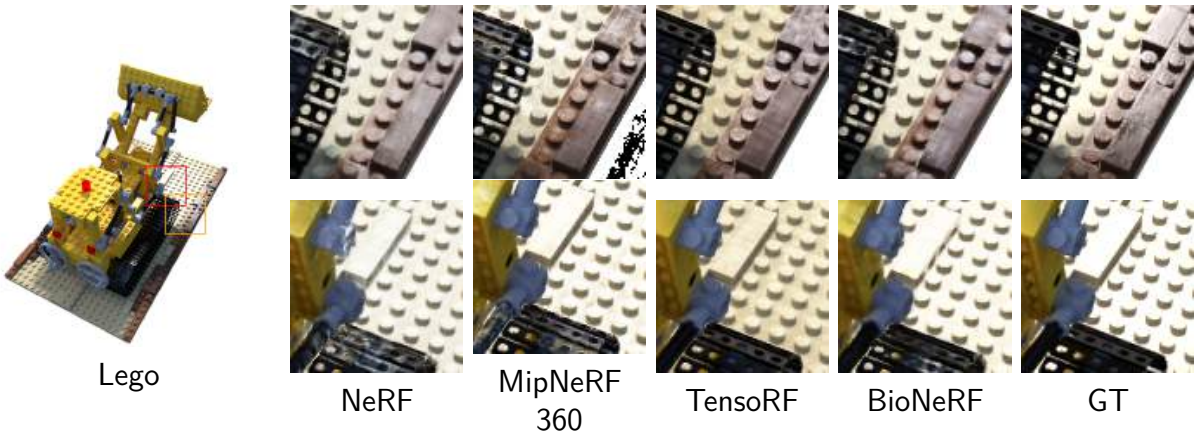

Figure 5 – Qualitative results of BioNeRF and comparison methods, namely NeRF (**??**), Mip-NeRF 360 (**??**), and TensoRF (**??**), as well as the ground truth images (GT) on two Blender dataset's scenes.

Regarding the LLFF dataset, Figure 6 compares BioNeRF against Mip-NeRF 360 (**??**) and TensoRF (**??**) considering a Fern view. In This context, BioNeRF views look cleaner than the instances generated by TensoRF, whose images present some high-frequency artifacts. Regarding Mip-NeRF 360, although the images' resolution looks a bit better than BioNeRF, they show some distorted details considering the air grid on the bottom picture.

[4] Baseline views obtained from <https://nerfbaselines.github.io/>.

Figure 6 – Qualitative results of BioNeRF and comparison methods, namely, Mip-NeRF 360 (**??**) and TensoRF (**??**), as well as the ground truth images (GT) on LLFF dataset's Fern scenes.

6.2 Per Scene Comparisson

Table 1 presents results concerning the Blender 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.

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.

BioNeRF obtained paramount results concerning the LLFF dataset, surpassing the state of the art for all measures, as presented in Table 2. 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 1 – Quantitative results considering each scene from Blender dataset.

PSNR ↑									
Method	Avg.	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship
NeRF (??) (ACM 21)	31.01	33.00	25.01	30.13	36.18	32.54	29.62	32.91	28.65
TensoRF (??) (ECCV'22)	33.14	35.76	26.01	33.99	37.41	**36.46**	30.12	34.61	30.77
Plenoxels (??) (CVPR'22)	31.71	33.98	25.35	31.83	36.43	34.10	29.14	33.26	29.62
Point-NeRF (??) (CVPR'22)	33.31	35.40	**26.06**	**36.13**	37.30	35.04	29.61	35.95	30.97
Ref-NeRF (??) (CVPR'22)	**33.99**	**35.83**	25.79	33.91	**37.72**	36.25	**35.41**	**36.76**	30.28
Mip-NeRF 360 (??) (CVPR'22)	30.34	34.01	24.36	26.66	36.44	33.20	27.91	31.50	28.66
L2G-NeRF (??) (CVPR'23)	28.62	30.99	23.75	26.11	34.56	27.71	27.60	30.91	27.31
Tetra-NeRF (??) (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 (??) (ACM 21)	0.947	0.967	0.925	0.964	0.974	0.961	0.949	0.980	0.856
TensoRF (??) (ECCV'22)	0.963	0.985	0.937	0.982	0.982	0.983	0.952	0.988	0.895
Plenoxels (??) (CVPR'22)	0.958	0.977	0.933	0.890	0.985	0.976	0.975	0.980	0.949
Point-NeRF (??) (CVPR'22)	0.978	**0.991**	**0.954**	**0.993**	**0.991**	**0.988**	0.971	**0.994**	0.942
Ref-NeRF (??) (CVPR'22)	0.966	0.984	0.937	0.983	0.984	0.981	**0.983**	0.992	0.880
Mip-NeRF 360 (??) (CVPR'22)	0.951	0.977	0.923	0.952	0.979	0.975	0.944	0.984	0.875
L2G-NeRF (??) (CVPR'23)	0.930	0.950	0.900	0.930	0.970	0.910	0.930	0.970	0.850
Tetra-NeRF (??) (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 (??) (ACM 21)	0.081	0.046	0.091	0.044	0.121	0.050	0.063	0.028	0.206
TensoRF (??) (ECCV'22)	0.027	**0.010**	0.051	**0.012**	0.013	**0.007**	0.026	0.009	0.085
Plenoxels (??) (CVPR'22)	0.049	0.031	0.067	0.026	0.037	0.028	0.057	0.015	0.134
Point-NeRF (??) (CVPR'22)	0.049	0.023	0.078	0.022	0.037	0.024	0.072	0.014	0.124
Ref-NeRF (??) (CVPR'22)	0.038	0.017	0.059	0.019	0.022	0.018	0.022	**0.007**	0.139
Mip-NeRF 360 (??) (CVPR'22)	0.060	0.032	0.083	0.048	0.039	0.028	0.067	0.021	0.164
L2G-NeRF (??) (CVPR'23)	0.070	0.050	0.100	0.060	0.030	0.060	0.060	0.050	0.130
Tetra-NeRF (??) (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**

6.3 General Evaluation

Table 3 provides the general quantitative results for view synthesis, comparing BioNeRF against state-of-the-art methods considering both synthetic and real image datasets[5]. 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 values, while the colors red, orange, and yellow denote the first, second, and third best results, respectively.

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,

[5] Values marked with – describe results not provided by the authors.

Table 2 – Quantitative results considering each scene from LLFF Real dataset.

	PSNR ↑								
Method	Avg.	Fern	Flower	Fortress	Horns	Leaves	Orchids	Room	T-Rex
NeRF (??) (ACM 21)	26.50	25.17	27.40	31.16	27.45	20.92	20.36	32.70	26.80
TensoRF (??) (ECCV'22)	26.73	25.27	**28.60**	31.36	28.14	21.30	19.87	32.35	26.97
Plenoxels (??) (CVPR'22)	26.29	25.46	27.83	31.09	27.58	21.41	20.24	30.22	26.48
Mip-NeRF 360 (??) (CVPR'22)	–	24.59	27.56	31.34	**28.51**	19.84	19.51	**33.49**	**27.86**
L2G-NeRF (??) (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 (??) (ACM 21)	0.811	0.792	0.827	0.881	0.828	0.690	0.641	0.948	0.880
TensoRF (??) (ECCV'22)	0.839	0.814	0.871	0.897	0.877	0.752	0.649	0.952	0.900
Plenoxels (??) (CVPR'22)	0.839	0.832	0.862	0.885	0.857	0.760	0.687	0.937	0.890
Mip-NeRF 360 (??) (CVPR'22)	–	0.820	0.867	0.900	**0.909**	0.721	0.660	**0.965**	**0.927**
L2G-NeRF (??) (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 (??) (ACM 21)	0.250	0.280	0.219	0.171	0.268	0.316	0.321	0.178	0.249
TensoRF (??) (ECCV'22)	0.124	0.155	0.106	0.075	0.123	0.153	0.201	0.082	0.099
Plenoxels (??) (CVPR'22)	0.210	0.224	0.179	0.180	0.231	0.198	0.242	0.192	0.238
Mip-NeRF 360 (??) (CVPR'22)	–	0.210	0.140	0.117	0.124	0.231	0.246	0.115	0.158
L2G-NeRF (??) (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**

Table 3 – General results considering both synthetic and real image datasets.

	Blender(??)			LLFF (Real)(??)		
Method	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
SRN (??) (NeurIPS'19)	22.26	0.846	0.170	22.84	0.668	0.378
LLFF (??) (ACM ToG 2019)	24.88	0.911	0.114	24.13	0.798	0.212
NeRF (??) (ACM 21)	31.01	0.947	0.081	26.50	0.811	0.250
IBRNet (??) (CVPR'21)	28.14	0.942	0.072	26.73	0.851	0.175
Mip-NeRF (??) (ICCV'21)	33.09	0.961	0.043	–	–	–
Mip-NeRF 360 (??) (CVPR'22)	30.34	0.951	0.060	26.59	0.846	0.168
TensoRF (??) (ECCV'22)	33.14	0.963	0.027	26, 73	0.839	0.124
Plenoxels (??) (CVPR'22)	31.71	0.958	0.049	26, 29	0.839	0.210
Point-NeRF (??) (CVPR'22)	33.31	0.978	0.027	–	–	–
Ref-NeRF (??) (CVPR'22)	**33.99**	0.966	0.038	–	–	–
L2G-NeRF (??) (CVPR'23)	28.62	0.930	0.070	24.54	0.750	0.200
Tetra-NeRF (??) (ICCV'23)	32.53	**0.982**	0.041	–	–	–
SpikingNeRF (??) (arXiv'23)	32.45	0.956	–	–	–	–
Localised-NeRF (??) (CVPR'24)	33.25	0.969	–	–	–	–
BioNeRF (Ours)	31.45	0.953	**0.026**	**27.01**	**0.861**	**0.068**

outperforming all state-of-the-art results in this context. Such results are highly positive since

LPIPS was designed to measure the "perceptual distance," which identifies how similar two images are in a way that coincides with human judgment. The method systematically evaluates deep features across different architectures to discover the perceptual similarities, a property shared across deep visual representations that correspond to human perceptions far better than the commonly used metrics like PSNR and SSIM (**??**).

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.

6.4 Ablation study

This section provides an evaluation of the BioNeRF effectiveness over the Blender dataset considering three implemented versions, based on the standard NeRF (**??**), NeRFacto[6], and TensoRF (**??**), which will be referred here as BioNeRF, BioNeRFacto, and BioTensoRF, respectively. The biologically plausible module is implemented in a substructure of NeRF's pipeline called field. In this context, BioNeRF changes are implemented in the Coarse and Fine blocks of NeRF's pipeline[7]. Considering BioNeRFacto and BioTensoRF, such changes are implemented in NeRFacto's nerfacto field and TensoRF's field, respectively.

Table 4 provides the average results achieved after training the models for $50k$ iteration. Considering the standard base models against the biologically plausible versions, BioNeRFacto outperformed NeRFacto over the PSNR metric while failing over SSIM and LPIPS. Such results show that the small architecture of NeRFacto is probably not robust enough to capture the required information to perform Steps 1 (Positional Feature Extraction) and 4 (Contextual Inference) in the proposed method. TensoRF outperformed BioTensoRF considering all three metrics, showing that TensoRF's image representation as a 3D voxel grid with per-voxel multi-channel features is not compatible with the proposed approach. Finally, BioNeRF outperformed NeRF over all three metrics, which allows the conclusion that the proposed module offers a relevant, biologically plausible solution to foster the scene view synthesis field.

Comparing the models with the biologically plausible module, BioNeRF and BioTensoRF outperformed BioNeRFacto results, considering all the evaluation metrics. Considering both methods, BioNeRF outperformed BioTensoRF by 0.001 considering the SSIM and a far advantage considering the LPIPS metric, while BioTensoRF obtained a slightly better PSNR result. Running the training during more iterations for some dataset scenes, we could observe

[6] <https://docs.nerf.studio/nerfology/methods/nerfacto.html>
[7] <https://docs.nerf.studio/nerfology/methods/nerf.html>

that BioNeRF sometimes surpasses BioTensoRF PSNR values.

Table 4 – Average results obtained considering three distinct implementations of BioNeRF over Blender datasets.

Method	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓
NeRF	27.78	0.916	0.089
NeRFacto	18.45	0.826	0.180
TensoRF	30.90	0.948	0.042
BioNeRFacto	19.01	0.792	0.216
BioTensoRF	29.15	0.928	0.063
BioNeRF	28.81	0.929	0.031

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 scientific report 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 (**??**). Additionally, we strive to extend our model to infer in real time and learn from a few shots.

8 Supervisor's Comments

The report is well written and presents the results obtained by the researcher during his project, which he worked on with great commitment. All the project steps were completed optimally, exceeding expectations and meeting the deadline. Moreover, the researcher was actively involved throughout his project, attending research group meetings and collaborating on several publications with laboratory colleagues, which consist of 8 high-impact journals and 8 conference papers, as well as 4 other papers submitted during the period.