From Perceptual Filling-In to Stable Active Vision

Advanced Seminar in Computer Science

Efrat Friedrich

Supervised by Dr. Hadar Cohen-Duwek

The Open University of Israel

September 2025

---

1 Introduction

1.1 Introduction

Human visual perception is a remarkable feat that provides a seemingly complete, colorful, and stable view of the world. However, this subjective experience stands in apparent discrepancy with the actual retinal input, which is often incomplete, of lower resolution, and highly dynamic due to constant eye movements, known as saccades. The peripheral retina, for instance, has reduced color sensitivity and spatial acuity, transmitting primarily edge-related information rather than full surface details.To bridge this gap and computationally model how the brain reconstructs a coherent visual experience, this research leverages neuromorphic event cameras and Spiking Neural Networks(SNNs).

Event cameras are biologically inspired sensors that capture changes in luminance as asynchronous "events" rather than traditional frames. They communicate transients in luminance via the Address Event Representation (AER) protocol, transmitting spikes tagged with pixel addresses asynchronously. This event-driven approach offers several advantages, including high temporal resolution, high dynamic range, and the absence of motion blur, making them ideal for emulating the rapid intra-saccadic motion data that occurs during eye movements. By converting real-world or synthetic saccadic movements into event-based data, researchers can simulate the dynamic and sparse nature of retinal input more realistically.

Central to the modeling efforts are Spiking Neural Networks (SNNs), which are considered highly biologically plausible computational models. SNNs operate by transmitting information through discrete "spikes," mimicking the way biological neurons communicate. The Neural Engineering Framework (NEF) is a key theoretical framework utilized in this research for designing and optimizing SNNs, allowing for the neuromorphic encoding, decoding, and transformation of complex mathematical functions (mathematical constructs) into networks of spiking neurons. This framework enables the implementation of functional, large-scale neural networks that can be realized on neuromorphic hardware, offering potential energy efficiency benefits.

The collective research explores various aspects of visual perception by employing these biologically inspired systems:

1. **Perceptual Filling-In**: SNNs are designed to demonstrate how the brain might reconstruct complete visual surfaces from detected edges, supporting the isomorphic theory where activation spreads across the retinotopic map. This involves solving the Poisson equation, either through a feedforward SNN with an optimized weight matrix or an iterative recurrent SNN mimicking horizontal connections in the cortex (Cohen Duwek and Tsur, 2021)[3].

2. **Image Reconstruction from Event Data**: Event-camera data is used as input for convolutional neural networks (CNNs) that predict the image's Laplacian. This predicted Laplacian is then fed into an SNN optimized for Poisson integration, enabling efficient and compact image reconstruction with a significantly reduced number of parameters. This highlights a pathway towards fully neuromorphic image processing (Cohen-Duwek et al., 2021)[11].

3. **Perceptual Colorization of Peripheral Vision**: To address the perceived richness of color despite limited retinal input, models computationally emulate foveated color maps and intensity channels, alongside intra-saccadic motion data from event cameras. These inputs are used by adversarially optimized neural networks (often U-Net architectures) to reconstruct high-resolution and colorful images, demonstrating how peripheral color perception might be filled in based on achromatic input and learned natural image statistics (Cohen Duwek et al., 2023)[7].

4. **Reconstruction of Visually Stable Perception**: To account for the stable visual experience despite constant eye movements, models integrate successive saccadic inputs using recurrent neural networks with Long Short-Term Memory (LSTM) components. Crucially, corollary discharge (CD) that signals neural signals informing the brain about intended eye movements, are introduced to facilitate anticipatory adjustments and maintain perceptual stability. This allows the reconstruction of a visually stable scene from dynamic, saccadic retinal inputs, aligning with experimental findings on the role of CD signals in perception (Cohen Duwek et al., 2024)[8].

In essence, this research endeavors to advance our understanding of active visual perception by developing biologically plausible models that leverage the unique characteristics of neuromorphic event cameras and SNNs to address the fundamental discrepancies between retinal input and our rich, stable visual experience.

1.2 Similarity and Evaluation Metrics

To quantitatively assess the quality and perceptual accuracy of the reconstructed images generated by the models, several similarity and evaluation metrics are employed, comparing model outputs against the Ground Truth (GT) or reference images. The main evaluation metrics used in the paper include:

- SSIM (Structural Similarity Index Measure): This metric measures the structural similarity between the reconstructed image and the reference image. Higher values indicate better preservation of visual structure.
- LPIPS (Learned Perceptual Image Patch Similarity): This is a deep learning-based perceptual metric that compares the deep features of images, commonly used to evaluate Generative Adversarial Network (GAN) outputs. Lower scores signify higher perceptual similarity.
- PSNR (Peak Signal-to-Noise Ratio): PSNR quantifies the quality of a reconstructed signal by measuring the ratio of peak signal power to noise power, typically reported in dB. Higher values mean better image quality (with scores generally considered good above 30 dB).
- CIEDE2000: This is a perceptual metric for calculating color differences, which gauges perceptual color disparities by considering attributes such as lightness, chroma, and hue. Lower values indicate more accurate color reproduction.
- MSE (Mean Squared Error): This metric calculates the average squared pixel difference between the predicted and actual image. Lower values signify higher similarity, although MSE is noted as not always being perceptually accurate.

1.3 Sensitivity Analysis: Overview of the Practical Implementation

This seminar concludes with a sensitivity analysis (Chapter 6) focusing on Model 5, a lightweight Convolutional Neural Network central to the efficient image reconstruction approach outlined in Cohen-Duwek et al. (2021) [11]. This exercise systematically investigated how variations in activation functions (Mish vs. ReLU), loss-weighting schemes (e.g., emphasizing SSIM or edge loss), and architectural elements (e.g., Batch Normalization) affected the model's Laplacian prediction and subsequent image reconstruction. The findings highlighted the fragility of these ultra-compact networks (Model 5 has ~ 277 parameters) to architectural changes when training data was constrained.

2 Biologically Plausible Spiking Neural Networks for Perceptual Filling-In

2.1 Key Term Definitions

Term	Definition
Isomorphic Theory	Filling-in occurs as activation spreads across the retinotopic map from edges inward.
Neural Engineering Framework (NEF)	Framework for SNN design, enabling encoding, transforming, and compiling large-scale neural models.
Perceptual filling-in	Visual phenomenon where surfaces appear complete, reconstructed from edges. A model perceptual filling-in uses Poisson equation to reconstruct images from gradients.
Poisson equation	Partial differential equation modeling diffusion; used to reconstruct images from gradients.
Spiking Neural Networks (SNNs)	Neural networks using discrete spikes and time dynamics to mimic biological neurons.

2.2 Introduction

Visual perception starts with processing spatio-temporal edges in areas like V1, but the brain reconstructs these into complete, filled-in surfaces through a process called perceptual filling-in[29]. This phenomenon is observed in various illusions, such as the watercolor illusion and afterimage filling-in. There are two main theories for perceptual filling-in[19]:

- Symbolic or Cognitive Theory: Suggests that edge information is in low-level areas, while surface color and shape are metadata in higher areas.
- Isomorphic Theory: Proposes that filling-in occurs as an activation pattern spreads across the retinotopic map of the visual cortex, from surface edges to interiors.

The exact mechanism remains unclear, with evidence for both. Previous computational model by the authors demonstrates many visual illusions governed by perceptual filling-in using a Poisson equation-based model to reconstruct an image from its gradients[5][6]. The current work builds on this by proposing two biologically plausible spiking neural networks (SNNs) that demonstrate perceptual filling-in by resolving the Poisson equation. These SNN implementations support the isomorphic theory and show how biologically plausible neuronal models can reconstruct an image from its gradients. The two distinct SNN architectures proposed are:

- Feedforward SNN: where a weight matrix is optimized to generate a solution. The weight matrix is used to directly solve the Poisson equation (Figure 1, top).
- Recurrent SNN: which follows evidence-based feedback connections. It iteratively solves the equation through horizontal feedback connections (Figure 1, bottom).

Both models are built using the Neural Engineering Framework (NEF), which translates mathematical functions into spiking neuron activity.

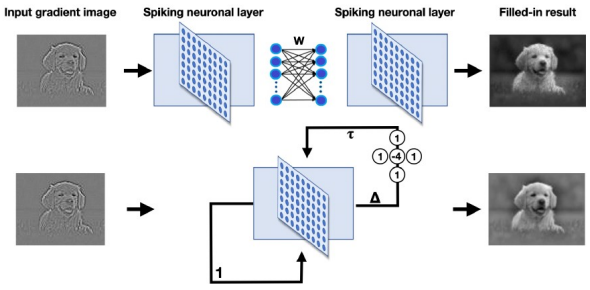

Figure 1: The two distinct SNN architectures proposed: Feedforward SNN (top) and Recurrent SNN (bottom).

2.3 Methods

2.3.1 Filling-In Model

The authors base their perceptual filling-in model on a previously developed computational approach[5][6] that reconstructs images from gradients using the diffusion equation, a form of the heat equation.

Psychophysical evidence indicates this perceptual reconstruction is rapid allowing the time-dependent term to be ignored. This reduces the diffusion equation to a steady-state Poisson equation only. This formulation mathematically captures how filled-in surfaces can emerge from edge information, aligning with the isomorphic theory of perceptual filling-in. To render both models a biologically plausible, the authors implement the Poisson equation using two distinct Spiking Neural Networks (SNNs) within the Neural Engineering Framework (NEF) which translates mathematical functions into spiking neuron activity.

2.3.2 Neural Engineering Framework (NEF)

The Neural Engineering Framework (NEF) [27] is utilized as the foundational methodology for implementing the spiking neural network (SNN) architectures by modeling how neurons encode, transform, and decode information through spiking activity. NEF enables biologically plausible encoding of mathematical functions within large-scale networks of spiking neurons, and resolves dynamic behaviors making it well-suited for modeling perceptual processes such as filling-in. NEF also supports deployment on neuromorphic hardware, thus this approach enhances the biological plausibility of the model and opens avenues for real-time neuromorphic hardware implementations of perceptual computation.

2.3.3 SNN Architectures

In this study, NEF is used to implement two distinct models of perceptual filling-in that simulate the solution of the Poisson equation: 1) a feedforward SNN that uses a weight matrix to directly solve the Poisson equation (Figure 1, top) 2) a recurrent SNN that iteratively solves the equation through horizontal feedback connections (Figure 1, bottem).

2.3.4 Feedforward SNN

The feedforward SNN solves the Poisson equation for image reconstruction through matrix manipulations by using a finite difference numerical method [28]. This model transforms the equation into a linear system using a finite difference method, where the perceived image is reconstructed from the Laplacian of the input. A weight matrix, derived as the inverse of the Laplacian matrix, encodes this transformation, allowing two layers of spiking neuron ensembles to reconstruct the image in a single pass. Despite the Laplacian matrix being sparse, its inverse is relatively dense, resulting in an all-to-one connectivity scheme.

2.3.5 Recurrent SNN

The recurrent SNN solves the Poisson equation iteratively by simulating the dynamic form of the diffusion equation. At each time step, the network updates the image reconstruction using local feedback based on neighboring pixel values, mimicking the gradual spread of activity from edges to interiors. This is implemented using horizontal (recurrent) connections within a single neural layer, where each neuron connects to its four neighbors and itself. The update rule spreads neural activity iteratively, reconstructing the perceived image over time.

2.3.6 Simulation and Pre-Processing

The SNN models were evaluated using the Nango neural compiler[1]. The inputs to both model were the Laplacian of input image. The Laplacian operator used to approximate the Difference of Gaussian (DOG) operator which is commonly used to represent the receptive fields of retinal ganglion cells[21]. Each pixel was encoded using 10–20 spiking neurons with Spiking-Rectified-Linear activation. Simulations were executed on a GPU-accelerated Azure virtual machine (6 cores, 56 GB RAM, Nvidia Tesla K80).

2.4 Results

2.4.1 Results

The data used in this study consists of four different images (a photograph of Einstein, an image of a dog, a landscape image and a black square).

2.4.2 Feedforward Method

Simulations show that image reconstruction from the Laplacian is nearly instantaneous, with rapid filling-in as the synaptic time constants determine its latency. The authors showed that neuronal activity remains stable over time, reflecting the method's non-iterative nature.

2.4.3 Recurrent Method

Unlike the Feedforward SNN, the Recurrent SNN requires multiple iterations to converge and reconstruct the image. It converges slowly, struggles with large uniform surfaces, and fails to converge when the number of neurons per pixel is reduced, except for images with long, continuous edges. This gradual convergence is reflected in the raster plots for the black box image, showing evolving neuronal activity over time.

2.5 Discussion

2.5.1 Conclusions and Innovations

The work presents two biologically plausible SNN's models that serve as potential neural mechanisms for perceptual filling-in in the brain. The authors propose two distinct neuronal implementations of a Poisson equation-based model for perceptual filling-in, realized using SNNs: a feedforward SNN and a recurrent SNN. This represents a significant innovation in demonstrating how biologically plausible neuronal models can reconstruct images from gradients. Both methods are consistent with the isomorphic hypothesis of perceptual filling-in, which posits that activation spreads across the retinotopic map from edges to interiors. The perceived reconstruction by both networks was not directly stimulated by the input image. The work suggests that both approaches might be present in the brain, with the recurrent method representing slower computation in lower visual areas and the feedforward method representing faster reconstruction in higher visual areas.

2.5.2 Key Findings

Recurrent SNN: This method iteratively solves the Poisson equation using a horizontal connectivity scheme. It aligns with experimental findings[17], where the surface's interior is filled in at a later iteration than areas near the edges, suggesting it may emulate the filling-in process in V1. However, the recurrent method requires numerous iterations to converge and demonstrates incomplete filling-in for large uniform surfaces (e.g., black squares), consistent with experimental observations of "unfilled holes" in V1 responses[30]. This implies it is insufficient to fully explain the neuronal representation of large uniform surfaces.
Feedforward SNN: This method solves the Poisson equation directly using a dense weight matrix connecting two spiking neuronal layers. It offers faster and more complete filling-in performance (e.g., the center of a black square image was filled) compared to the recurrent method. A significant drawback in terms of biological plausibility is its reliance on an all-to-one (dense) connectivity scheme, which is inconsistent with the typical receptive field organization of visual layers[23].

2.5.3 Further Work and Implications

- The current models do not comprehensively explain all visual illusions governed by filling-in, as phenomena like the watercolor illusion or Cornsweet illusion involve additional neuronal processes such as attention and lateral inhibition, which are beyond the current model's scope.
- Future research should implement neuronal lateral inhibition and feedback processes using SNNs to gain a deeper understanding of the underlying mechanisms of perceptual filling-in and to compare these approaches with experimental findings.
- For the feedforward method, future work could explore resolving its biological implausibility by separating the visual field into distinct regions that are reconstructed locally and then "stitched" together at higher visual layers. This suggests concentrating efforts on higher visual areas (V3 and V4) for further supporting evidence.
- The authors hypothesize that a fast pathway (potentially represented by the feedforward model in higher visual areas) might be involved in cognitive functions such as attention and learning, which then modulate isomorphic processes in early visual areas via recurrent signals.

3 Image Reconstruction from Neuromorphic Event Cameras using Laplacian Prediction and Poisson Integration with Spiking and Artificial Neural Networks

3.1 Key Term Definitions

Term	Definition
Address Event Representation (AER)	Protocol used by neuromorphic sensors to transmit spikes tagged with pixel addresses asynchronously.
Event Camera (Dynamic Vision Sensor DVS)	Neuromorphic sensor recording only brightness changes for fast, low-latency motion capture.
Intensity Image	Grayscale image where each pixel encodes brightness.
Laplacian Operator	Second-order operator highlighting regions of rapid change, e.g., edges in images.
Laplacian Prediction	Predicting second-order derivatives to capture edges and structure.
N-MNIST and N-Caltech101	Event-based datasets derived from MNIST and Caltech101 with synthetic saccades.
Poisson Integration / Poisson Solver	Method to reconstruct intensity images from Laplacians via Poisson equation.

3.2 Introduction

Neuromorphic vision sensors, especially Dynamic Vision Sensors (DVSs), can resolved thousands of frames per seconds, excel in high-speed visual processing by capturing pixel-level changes in luminance as asynchronous spikes using the Address Event Representation (AER) protocol. These sensors are frameless, have high temporal resolution, no motion blur, and offer efficient data compression at the sensor level, allowing data transfer, storage, and processing to be optimized[2].

Recent advances in this field use Convolutional Neural Networks (CNNs) to reconstruct natural video from event data such as U-net with 10M parameters[22]. Recent work introduced smaller CNNs with residual blocks and recurrent connections (RNN), reducing computational cost with minimal performance loss.

This work proposes a neuromorphic (brain inspired) image reconstruction approach using NEF-based framework enabling executing on neuromorphic hardware which is energy efficient. The pipeline combines a CNN for Laplacian prediction with a SNN for Poisson-based image integration, enabling a low-parameter, energy-efficient solution. Furthermore, they propose non-spiking CNN version with Mish activation[12] using less than 100 parameters and achieving promising results. The models were tested on N-MNIST and N-Caltech101 datasets[14].

3.3 Related Works

Previous methods for reconstructing images from event data often relied on estimating image gradients followed by Poisson integration, or used Generative Adversarial Networks (GANs) for high-quality HDR reconstruction. Though GANs are complex and hard to train, they have provide sate-of-the-art performance. More recent efforts involve sophisticated pipelines with optical flow and super-resolution, but these typically require large models and are not neuromorphic. Some works have applied Spiking Neural Networks (SNNs) to event-based tasks like gesture recognition or object tracking, but not for full image reconstruction.

This paper addresses that gap by proposing two models enablling efficient, low-parameter image reconstruction: (1) a hybrid CNN-SNN pipeline that predicts the Laplacian and reconstructs intensity via Poisson integration, and (2) a highly compact CNN with shared-event filters and Mish activation, achieving effective reconstruction with fewer than 100 parameters. The first model CNN is entirely realized using SNNs, while the second is not.

3.4 Methods

3.4.1 Input Representation and Pre-processing

Event data from the N-MNIST and N-Caltech101 datasets were converted into event-frame tensors, representing spatial and temporal activity over multiple frames. Each frame encodes luminance changes over 50 ms intervals and is using a spatial median filter to reduce noise, before entering the neural network.

3.4.2 CNN Laplacian Prediction

The CNN Laplacian Prediction is the initial step of the model (Figure 2). The authors designed a five-layer CNN to predict the Laplacian of an image from preprocessed event-frame tensors. The network is trained using a composite loss function that includes Mean Absolute Error (MAE) for Laplacian accuracy, SSIM to enhance structural similarity of the reconstructed image, and an edge loss based on binary edge maps to preserve visual detail. ReLU activations are used throughout, and the model is implemented in Keras. Training is conducted on the N-Caltech101 dataset using the Adam optimizer, with learning rate decay and early stopping triggered by plateau in validation SSIM.

3.4.3 CNN to SNN Conversion

The trained CNN is converted into a SNN using the NEF via the NengoDL library[13]. ReLU activations are replaced with spiking rectified linear units, where neuron firing rates are proportional to positive input values, filtered through a low-pass synapse (10 ms time constant). Each image is presented for 100 ms, allowing the SNN to process temporal information. This conversion preserves the original architecture while enabling neuromorphic, event-driven processing compatible with spiking hardware, supporting efficient inference with reduced energy use.

3.4.4 Filling-in SNN

To perform Poisson integration neuromorphically, the authors implemented a feedforward SNN that reconstructs image intensity from its predicted Laplacian. This is achieved by solving a linear system derived from the discrete Poisson equation, using a fixed, precomputed weight matrix based on the Laplace operator. The network connects two neuron layers with this weight matrix, allowing efficient intensity reconstruction without training, significantly reducing the model's parameter count.

3.4.5 Direct reconstruction

The authors trained similarly sized CNNs for direct image reconstruction from event data by (passing Poisson integration) as a comparison to the Laplacian-based pipeline. The loss function was adapted to evaluate the predicted image directly against the ground truth using MAE, SSIM, and edge loss. This approach serves as a baseline to evaluate the benefits of using Laplacian prediction in the reconstruction pipeline.

3.4.6 Shared-event filters CNN

The authors introduced a novel compact CNN (<100 parameters) to further reduce parameters, that treats event data as a video-like signal across time. The input tensor is reshaped to allow temporal and spatial convolution in stages, first processing each frame independently, then combining features across time to predict the image Laplacian. They tested this design using both ReLU and Mish activations, with Mish offering smoother, more effective learning due to its non-monotonic and bounded nature.

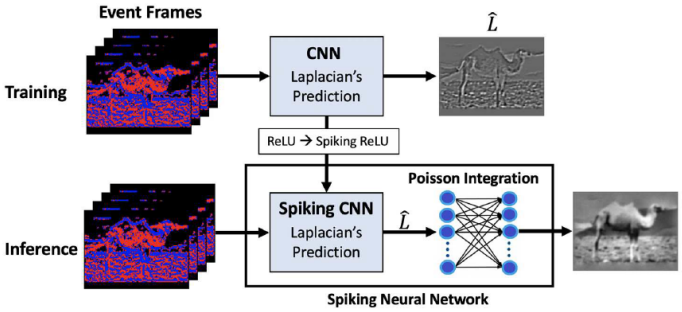

Figure 2: SNN archietecture at training time and at inference time.

3.5 Results

The authors evaluated six CNN models of varying sizes for Laplacian prediction and image reconstruction on N-MNIST and N-Caltech101 datasets. Notably, smaller networks (e.g., 277 parameters) performed comparably to larger ones, proving the method's efficiency. One compact model (with 277 parameters) was converted into an SNN, achieving reasonable performance, especially at higher firing rates (5 kHz), though with increased energy cost. They also tested direct reconstruction CNNs, which only captured image edges and underperformed compared to Laplacian-based models, even with more parameters. Finally, the shared-event filters CNN with fewer than 100 parameters showed strong results, with

Mish activation outperforming ReLU, demonstrating that high-quality reconstruction is possible with extremely lightweight architectures.

3.5.1 Reconstruction via Laplacian predication

Six CNN models with varying parameter counts were trained to predict image Laplacians from event data. In particular, smaller models, such as Model 4 (with 1,691 parameters) and Models 5 and 6 (with about only 200 parameters), achieved performance comparable to much larger ones. Model 5 (277 parameters) was selected for conversion into an SNN, enabling a fully neuromorphic image reconstruction pipeline. Performance was evaluated using three key metrics: PSNR, SSIM, and MSE. Although the SNN version of Model 5 was noisier at low firing rates (100 Hz), increasing the firing rate to 5,000 Hz substantially improved image quality, making it comparable to the non-spiking CNN though at a higher energy cost.

3.5.2 Direct Reconstruction

The authors evaluated CNNs trained to directly reconstruct images from event data, without Laplacian prediction or Poisson integration. Although the largest Model 1 (with more than 50,000 parameters) performed best among direct models, it still underperformed compared to smaller Laplacian-based networks, especially in preserving fine image details. These direct models primarily captured image edges rather than full intensity, showing that Laplacian-based reconstruction is both more effective and efficient, even with drastically fewer parameters.

3.5.3 Shared-events filters CNN

To minimize parameters further, the authors developed a shared-event filters CNN with fewer than 100 trainable parameters. By treating input as a time-series signal, early layers process individual frames, while later layers combine them to estimate the image Laplacian. The model was tested with ReLU and Mish activations[12]. Mish performed better, likely due to its smooth, non-monotonic properties. Despite its tiny size, this architecture delivered reconstruction quality comparable to much larger networks.

3.6 Discussion

3.6.1 Conclusions and Innovations

This study demonstrates that high-quality image reconstruction from event cameras can be achieved using a lightweight CNN-SNN hybrid pipeline. By predicting the image Laplacian and applying Poisson integration through an SNN, the authors drastically reduced the number of trainable parameters. They further converted the CNNs to SNNs for a complete neuromorphic design while enabling a fully neuromorphic implementation.They showed that even simple CNNs without U-Net or autoencoders can produce effective results, especially when combined with a non-trainable, efficient filling-in SNN. The shared-event filters CNN further minimized complexity, reconstructing images with less than 100 parameters. Though not yet converted to SNN, this model has potential for future neuromorphic deployment using SNN by utilizing recurrent architectures. The method currently depends on datasets created with fixed, artificial saccades, but the authors suggest it could generalize to more diverse event streams with simple architectural adjustments.

3.6.2 Key Findings and Further Work

- Compact CNNs (with as few as 277 parameters) matched or outperformed much larger models.
- SNNs with modest firing rates (e.g., 100 Hz) produced reasonable reconstructions; increasing to 5,000 Hz improved quality but raised energy costs.
- A novel shared-event filters CNN achieved strong performance with less than 100 parameters, although it was not yet adapted to spiking format.
- Future improvements could include patch-based Poisson integration or recurrent SNN architectures to better handle temporal sequences and improve memory handling.

3.6.3 Implications

This work highlights a practical path toward scalable, energy-efficient neuromorphic vision systems, showing that event-based image reconstruction can be performed effectively with minimal computational resources. Although tested fixed non-biologically plausible three saccades (datasets N-MNIST, N-Caltech101), the method shows promise for broader real-world applications, provided input variability and data acquisition strategies are addressed in future adaptations.

4 Perceptual Colorization of the Peripheral Retinotopic Visual Field Using Adversarially Optimized Neural Networks

4.1 Key Term Definitions

Term	Definition
Fovea	Small retinal center responsible for the highest acuity and sharp vision.
Foveated Visual Field / Peripheral Vision	High resolution in fovea, low resolution and weak color in retina's periphery.
GAN (Generative Adversarial Network)	Model with Generator and Discriminator trained in competition to produce realistic data.
PatchGAN Discriminator	Convolutional discriminator evaluating image patches for realistic textures.
U-Net	Encoder–decoder CNN with skip connections for image-to-image tasks like color prediction.

4.2 Introduction

The introduction highlights the discrepancy between our rich visual perception and the inherently limited retinal input. The retina has unevenly distributed photoreceptors (fovea-dense cones for color, peripheral rods for achromatic vision), and Retinal Ganglion Cells (RGCs) compress signals, leading to lower resolution and incomplete color information in the periphery. Despite this, our subjective experience is a uniformly sharp and colorful visual field, with studies showing people often don't notice peripheral color removal, suggesting the brain "fills in" color[10]. Furthermore, saccadic eye movements[15], which are rapid shifts in gaze, should cause blur, but the brain maintains perceptual stability through trans-saccadic integration of these shifting inputs. To bridge this gap, the authors propose a neural network model that reconstructs high-resolution, colorful images from retinal-like inputs, including foveated color maps, intensity channels, and intra-saccadic motion data, using a U-Net architecture and adversarial optimization.

4.3 Methods

4.3.1 Generating Retinal Input

To simulate biologically realistic retinal inputs, the authors utilized the N-Caltech101 neuromorphic dataset, combining RGB images with event-based recordings (see Figure 3 for a schematic illustration).

The RGB inputs were transformed into color-opponent channels (RG, BY) and an achromatic intensity channel (I), replicating the center-surround receptive field properties of retinal ganglion cells[20]. They use the color opponent transformation matrix and discrete Laplacian operator.

Peripheral color sensitivity was emulated by applying a foveation mask: chromatic information was preserved within a central circular region (30-pixel radius), while peripheral areas were suppressed and were zeroed outside the mask area. Spatial resolution degradation was modeled by Gaussian blurring filters with diffident scales, reflecting the increasing receptive field size toward the periphery.

Intra-saccadic motion was captured using event-based data representing pixel-level brightness changes, recorded from three simulated saccades per image via neuromorphic event camera (silicon retina). These were converted into six event-frame temporal tensors as proposed by (Cohen-Duwek et al., 2021) [11].

The final retinal input tensor combined : six motion channels (represent Intra-saccadic motion) with the three static transformed channels (RG,BY,I) (represent retinal inputs) , forming a 9-channel input. Therefore, the final input tensor to the neural network was of size batch $\times$ height $\times$ width $\times$ 9, where the spatial dimensions are (90×120).

4.3.2 Reconstruction and Colorization Neural Network

The proposed architecture reconstructs high-resolution, colorful images from simulated retinal input through a three-stage pipeline (Figure 4):

1. **Laplacian Prediction:** A 5-layer convolutional neural network (CNN) proposed by (Cohen-Duwek et al., 2021)[11] which predicts the image's Laplacian (edge map) from the event data. Here, the CNN also incorporates the foveated intensity Laplacian. They use the Mean absolute error (MAE) loss function. This stage captures fine spatial and temporal structure from intra-saccadic motion.

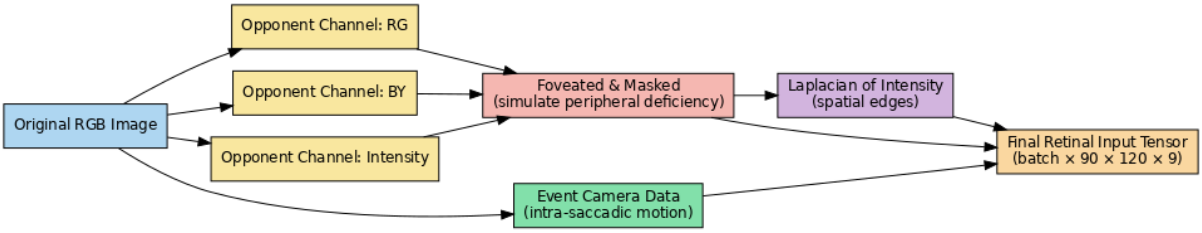

Figure 3: Retinal input generation process

2. **Poisson Solver Layer:**A differentiable Poisson integration module reconstructs the intensity image from the predicted Laplacian (edge input). Although it does not contain learnable parameters, it was incorporated into the network to back-propagate errors for end-to-end training of the network. It was based on PI algorithm[26], and was optimized using the Structural Similarity Index (SSIM) applied to the Laplacian and the predicted Laplacian of the image intensity

3. **Colorization via U-Net:** The model performs image colorization by combining the reconstructed image intensity with the opponent color channels (RG and BY) through a U-Net architecture, which features an encoder-decoder structure with skip connections[22]. The model is trained using two loss functions: (1) Mean Absolute Error (MAE) between the predicted and original opponent color channels, and (2) perceptual similarity metrics between the reconstructed and original RGB images, including SSIM and LPIPS. After predicting the opponent colors, an inverse transformation is applied to convert the image back to the RGB color space.

4.3.3 Generative Adversarial Network (GAN)

In this stage, the authors incorporated a Generative Adversarial Network (GAN) to enhance the realism of the colorized images. The model consists of two components: a Generator, which reconstructs colorful images from retinal inputs, and a Discriminator (a PatchGAN classifier[18]), which learns to distinguish between real and generated ("fake") images. Training involves an adversarial process where the generator aims to produce images that are indistinguishable from real ones, while the discriminator tries to detect the fakes.

The final objective combines adversarial loss with other loss functions from the previous stage, such as perceptual loss, color prediction error, Laplacian error, and structural similarity, resulting in end-to-end training that promotes both perceptual accuracy and realistic colorization.

4.3.4 Implementation Details

The model was implemented in TensorFlow and trained using Google Colab. Training was conducted on the N-Caltech101 dataset using the Adam optimizer with an initial learning rate of 0.001, which was reduced by 20% upon plateauing validation loss. Loss components were weighted as follows: Laplacian prediction (100), Poisson reconstruction (25), colorization (150), perceptual similarity (100), and adversarial loss (10).

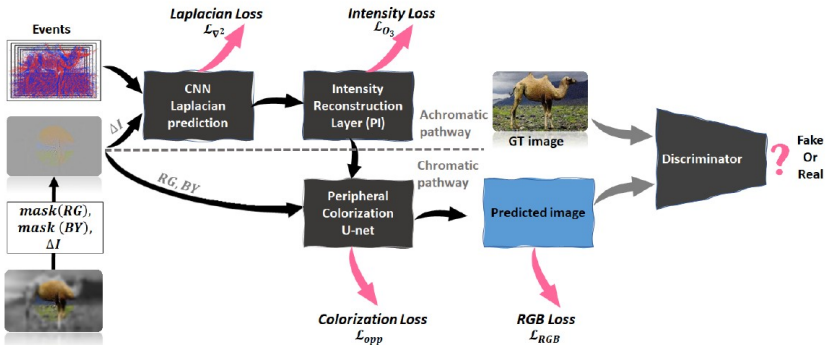

Figure 4: Retinal input generation process

4.4 Results

The model was evaluated using two training methods: (1) standard generator loss minimization, (2) adversarial training with a discriminator. Both approaches successfully reconstructed high-quality, colorful

images including peripheral areas from retinal inputs which lack peripheral information.

Quantitative metrics (SSIM and LPIPS) indicated that non-adversarial training yielded better perceptual similarity to ground truth. However, adversarial training produced more visually colorful reconstructions, especially in peripheral regions, suggesting improved perceptual plausibility.

The inclusion of event-based data (representing intra-saccadic motion) enhanced peripheral sharpness. In contrast, models trained without event data generated blurrier reconstructions, particularly in textures and peripheral details.

To conclude, comparative metrics (SSIM and LPIPS), confirm that both events and adversarial training contribute complementary improvements: events aid structural accuracy, while GANs enhance chromatic richness.

4.5 Discussion

4.5.1 Conclusions and Innovations

This study introduces a biologically inspired deep neural network that reconstructs full-color, high-resolution images from limited retinal-like inputs. By integrating achromatic and chromatic pathways that mirror human vision and incorporating intra-saccadic motion via event cameras, the model effectively compensates for missing peripheral visual information. Building on the previous work of the authors [3] [4] [9] demonstrating achromatic edge-based surface reconstruction, this study further innovates by applying adversarial training to enhance perceptual realism, particularly in color-deprived peripheral regions.

4.5.2 Key Findings

- Adversarial training produced more visually colorful reconstructions, especially in peripheral regions, suggesting improved perceptual plausibility.
- Non-adversarial training produced more perceptually accurate reconstructions (higher SSIM and lower LPIPS), while adversarial training generated more vibrant, colorful peripheral areas, possibly because some ground truth (GT) images lacked color in those regions. This phenomenon is consistent with visual illusions[9].
- Event-based data which simulate the effects of eye movements, significantly improved image sharpness in peripheral regions.
- The Adversarial model was able to infer plausible colors in achromatic regions by leveraging learned statistical regularities of natural images, consistent with perceptual color "filling in".

4.5.3 Further Work

- Utilizing a more natural and diverse dataset with randomized or attention-driven saccades to enhance peripheral colorization and reduce artifacts, addressing limitations of current unnatural inputs and fixed saccade patterns in N-Caltech101.
- Integrating a realistic peripheral vision model[16] to reflect partial peripheral color sensitivity, thereby enriching network input and moving beyond the assumption of a colorless periphery.
- Adopting recurrent architectures (e.g., convolutional LSTMs) to increase biological plausibility by simulating visual working memory, improving trans-saccadic integration, and explaining perceptual phenomena like unnoticed peripheral color removal observed in VR experiments[10].

4.5.4 Implications

Despite using "black box" models and non-biologically plausible optimization (like backpropagation), deep neural networks inspired by biological brain principles can model perceptual phenomena. This suggests that optimization-based frameworks can enhance our understanding of human vision

5 Reconstruction of visually stable perception from saccadic retinal inputs using corollary discharge signals-driven convLSTM neural networks

5.1 Key Term Definitions

Term	Definition
ConvLSTM (Convolutional Long Short-Term Memory)	RNN capturing spatial–temporal dependencies in image sequences, integrating multiple saccades for stable reconstructions.
Corollary Discharge (CD) signals	Neural signals predicting changes during saccades, keeping vision spatially aligned.
Good Features to Track	Corner detection method selecting salient points with strong local intensity changes.
Saccade vision	Rapid eye fixations across the visual field, several times per second.
Trans-saccadic Integration	Brain mechanism stabilizing perception across eye movements.

5.2 Introduction

The human visual system achieves a stable and coherent perceptual experience despite constant disruptions from saccadic eye movements and limitations of peripheral vision, such as reduced spatial and color resolution and the retina's primary encoding of edges. To compensate, the brain employs saccades to direct attention to salient areas and integrates acquired information through complex mechanisms. These include trans-saccadic integration for smooth perception across saccades, inter-saccadic motion processing for gaze correction, and saccadic suppression to prevent blurring. Crucially, Corollary Discharge (CD) signals provide anticipatory information about intended eye movements, enabling adjustments that contribute to a stable visual experience.

Prior research by Cohen et al. (2020)[10] utilized a VR setup where only attended regions were shown in color, demonstrating that participants still perceived a rich, colorful world despite limited input, challenging assumptions about perceptual awareness in active vision. Expanding on this, Cohen Duwek et al. (2023)[7] developed a model that reconstructed sharp, colorful images from constrained retinal input and inter-saccadic motion. However, this model was limited to simple datasets with fixed-target saccades.

This study extends earlier work by creating a new synthetic dataset that mimics active vision with three saccades to points of interest. The model employs Convolutional Neural Networks with Long-Short-Term-Memory (convLSTM) and incorporates CD signals to reconstruct full-color, stable images from partial and motion-based retinal inputs. This approach offers a more biologically plausible simulation of visual perception during active vision.

5.3 Methods

5.3.1 Generating Retinal Inputs

The method synthetically generates retinal data mimicking natural eye movements across three saccades directed toward points of interest, identified using the "Good Features to Track" method combined with K-Means clustering[24]. The computed retinal input process is shown in Figure 5. and consists of three main channels:

1. Chromatic Channel (Foveated Color Maps): This channel replicates the activity of color-opponent Retinal Ganglion Cells (RGCs). RGB images were transformed into opponent color space (RG, BY, and Intensity I). To simulate reduced peripheral color sensitivity, Gaussian filters were applied to mimic receptive field sizes, and a circular mask zeroed out color information beyond a 42-pixel radius.

2. Achromatic Channel: This simulates On-Off center-surround RGCs using a discrete Laplacian operator applied to the intensity channel.

3. Intra-saccadic Motion (Event Data): To capture motion during the three saccades per scene, the V2E event camera simulator was used to generate event frames based on pixel-level luminance changes from video sequences. These events were accumulated into 3–5 integrated event frames per saccade.

Crucially, Corollary Discharge (CD) signals were represented as translation vectors (x, y) to simulate the brain's anticipatory mechanism. These vectors align (translate) the event frames, achromatic channel, and chromatic channel to their correct positions relative to the original scene, facilitating image stabilization.

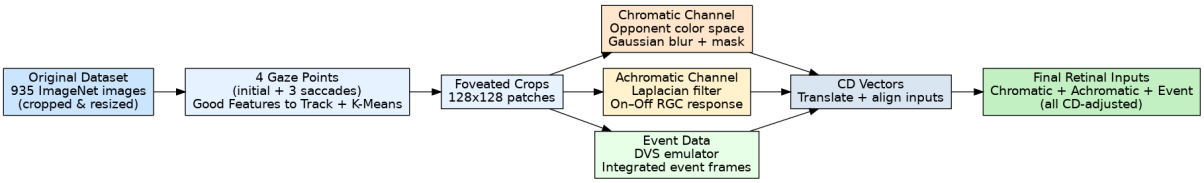

Figure 5: Retinal input generation process

5.3.2 Model Architecture and Training

The reconstruction model is built on a Generative Adversarial Network (GAN) framework, consisting of a Generator (G) with a convolutional PatchGAN classifier[18] and a Discriminator (D), trained to produce visually stable and realistic results. The Generator's reconstruction process is broken down into four sequential phases:

1. **Intensity Reconstruction from Events:** For each saccade, the integrated event frames were aligned to their real-world positions by CD vectors, placed at their corresponding scene locations, and then cropped to 128×128 pixel. These frames were grouped into 3–5 integrated event frames per saccade (with padding if needed), and then processed by a Convolutional Neural Network with Long-Short-Term-Memory (ConvLSTM) [25] followed by convolutional layers using Mean Absolute Error (MAE) as loss function to generate an initial intensity map.

2. **Enhanced Intensity Prediction:** A parallel processing step uses a Poisson solver layer (Poisson Integration) to predict intensity from the foveated Laplacian input as described in Cohen Duwek et al. (2021) [3]. The intensity maps from both the event reconstruction and the Laplacian prediction are combined via a convolutional layer using MAE loss function to yield enhanced intensity maps for each saccade.

3. **Colorization (U-Net):** A U-Net architecture uses the enhanced intensity maps along with the foveated opponent color channels (RG and BY) to reconstruct a fully colored image for each saccade. The model was trained with two loss functions: MAE for pixel-level color accuracy, and perceptual loss (SSIM + LPIPS) for structural and perceptual quality, as detailed in Cohen Duwek et al. (2022) [9].

4. **Saccadic Integration:** The four reconstructed colored images (the original and one per gaze point) are finally integrated using additional ConvLSTM layers to refine the result and produce the final stable image output. This integration merges visual information across successive saccades, allowing the network to refine and enhance the final output.

The optimization goal minimizes a combined loss function for the Generator, including Adversarial loss ($\mathcal{L}_{GAN}$), Reconstruction losses (MAE), and Perceptual similarity losses ($\mathcal{L}_{PERP}$), which combines SSIM and LPIPS metrics. Two models were trained: One with integrated event frames and one without event frames. The model's performance was evaluated using four key metrics to assess visual fidelity and color realism: SSIM (structural similarity), LPIPS (perceptual similarity), PSNR (signal quality), and CIEDE2000 (perceptual color differences).

The architecture is designed to combine spatial, temporal, and motion data for robust image reconstruction under dynamic, eye-movement conditions. The framework schematic is shown in Figure 6.

5.3.3 Implementation details

The dataset was created from 935 ImageNet images. The model was implemented in TensorFlow and trained on an NVIDIA A100 GPU (80GB RAM). The dataset of 935 ImageNet images was split into training, validation, and testing sets. Training was conducted for 200 epochs, with a batch size of 8 and an initial learning rate of 0.001. Specific weights were assigned to different loss components to guide optimization.

5.4 Results

The model was evaluated under various conditions: with and without event data, with 1 to 3 saccades, and with and without noise in the CD signals. The key findings, shown in the results table and demonstrated in the attached example, are as follows:

- More saccades enhanced reconstruction quality: From 1 to 3 saccades, reconstructions became sharper and more colorful, with the best SSIM and LPIPS scores achieved using three saccades. Quantitatively, the highest similarity scores (SSIM and LPIPS) were achieved when input from three saccades was used.

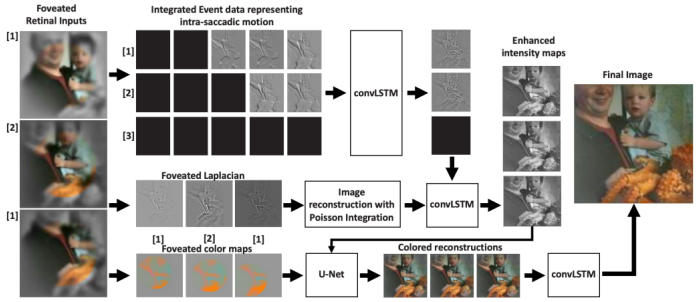

Figure 6: The architecture of the proposed reconstruction and colorization model

- Using event data led to better perceptual quality: Event data (representing inter-saccadic motion) improved perceptual similarity metrics (SSIM and LPIPS), although pixel-level accuracy (PSNR) and color difference (CIEDE2000) were sometimes better without events.
- CD noise sensitivity: Adding Gaussian noise to the CD vectors caused a drop across all evaluation scores and introduced visual blurriness in the reconstructed images.

5.5 Discussion

5.5.1 Conclusions and Innovations

This study demonstrates that a deep recurrent neural network can effectively reconstruct detailed images from restricted retinal inputs during saccadic eye movements, specifically involving saccadic eye movements towards points of interest. Key innovations include:

- The creation of a synthetic dataset incorporating retinal inputs with intensity, color, and motion information.
- The crucial incorporation of Long-Short-Term Memory (LSTM) and Corollary Discharge (CD) signals for enhancing image stabilization through eye movements.
- Unlike previous models limited to fixed non-saccadic images or simpler datasets[7], this work extends the capability to reconstruct stable images from retinal inputs involving saccadic eye movements mimicking active vision.
- Convolutional LSTMs are highlighted as serving as a functional visual memory for trans-saccadic integration, crucial for high visual accuracy in the peripheral region and vision stabilization across saccades.

5.5.2 Key Findings

- The model reproduces findings from Cohen and colleagues (2020) [10], showing that observers can perceive a colorful scene even when peripheral colors are eliminated in each saccade.
- Using event data led to better perceptual quality and indicating a trade-off between perceptual similarity and strict pixel accuracy. This highlights the importance of events for perceptual similarity.
- CD noise significantly degrades output quality, aligning with biological evidence of its role in perceptual stability. This underscores the critical role of accurate CD signals in maintaining stable visual perception.

5.5.3 Further work

Further work could involve expanding the training dataset beyond three saccades to include varied numbers, which is anticipated to enhance the model's adaptability and generalization to diverse real-world visual processing scenarios.

5.5.4 Implications

This research significantly advances the understanding and modeling of active vision and image reconstruction. By integrating saccadic integration, CD signals, and LSTM, it provides more realistic models for stable visual perception despite continuous eye movements and limited retinal input.

6 Implementations

6.1 Introduction

This exercise builds on Model 5 from Cohen-Duwek et al. (2021) [11], a lightweight CNN designed for Laplacian prediction in event-based reconstruction. For clarity, I refer to Model 5 as ConvNet5 throughout this section and in the figures. ConvNet5, featuring only ~ 277 trainable parameters, was selected because prior work demonstrated that such compact models could achieve reconstruction performance comparable to networks featuring significantly more parameters. The goal of this implementation was not to develop a new state-of-the-art method, but rather to demonstrate understanding and implementation by reproducing the baseline and running a small sensitivity analysis of architectural and loss variations.

6.2 Methods

Experiments used the Caltech101 event dataset , preprocessed into event-frame tensors [14], with Laplacian ground truth, and framed the task as predicting image Laplacians followed by Poisson integration for intensity reconstruction. The baseline architecture was ConvNet5, a compact CNN (~ 277 params). Reconstruction quality assessed with PSNR, SSIM, and MSE, emphasizing how small design changes affect these metrics under limited training.

6.3 Experimental Setup and Technical Details

The models were implemented in TenserFlow and trained using Google Colab (CPU). All experimental runs utilized the ConvNet5 baseline architecture (with ~ 277 parameters) as the starting point and followed the optimization settings of the original study. To maintain a controlled and quick probe, all models were trained using only 150 training samples. Importantly, the Mish activation and loss-weighting variants (SSIM Emph., Edge Emph.) were initialized using the original ConvNet5 pretrained weights. In contrast, the normalization variants (ReLU+BN, ReLU+Dil.+BN) were trained from scratch to avoid weight-mismatch effects. The training framework developed for this exercise is highly flexible, allowing different configurations of activations, normalization layers, dilation, and loss weighting to be combined and tested systematically. Furthermore, while the implementation supported various normalization strategies (BatchNorm, LayerNorm, GroupNorm), only Batch Normalization (BN) was tested, as the other options were anticipated to introduce too many parameters compared to the lightweight baseline design. The full configuration of each variant including activation type, loss weights, hyperparameters, and intended purpose, is summarized in Table 1.

Variant	**Model**	**Activation**	$\lambda = (\lambda_1, \lambda_2, \lambda_3)$	**Hyperparameters**	**Purpose**
Baseline	ConvNet5	ReLU	1.0,0.25,0.25	lr=1e-6,ep=10	Provide baseline reference as in the paper.
Mish	ConvNet5	Mish	1.0,0.25,0.25	lr=5e-4,ep=30	Test whether Mish improves quality over ReLU.
SSIM Emph.	ConvNet5	ReLU	1.0,0.35,0.15	lr=5e-4,ep=50	Examine effect of emphasizing SSIM on perceptual quality.
Edge Emph.	ConvNet5	ReLU	1.0,0.15,0.35	lr=5e-4,ep=50	Test whether edge emphasis sharpens reconstructions.
ReLU+BN	ConvNet5+BN	ReLU	1.0,0.25,0.25	lr=1e-3,ep=50	Improve stability and reduce oscillations
ReLU+Dil.+BN	ConvNet5+BN+dil=2	ReLU	1.0,0.25,0.25	lr=1e-3,ep=50	Test whether dilation improves spatial context capture.

Table 1: Summary of model variants. *lr* = learning rate; *ep* = epochs; Batch size was set to 16 across all experiments.

6.4 Results

The baseline ConvNet5 model (277 parameters) served as the reference point for all comparisons. As expected, its role was not to achieve the best possible metrics but to provide a benchmark for assessing the sensitivity of the different variants. Although the absolute scores were lower than those reported in the original paper, the baseline consistently produced recognizable reconstructions and therefore offered a stable point of comparison. See Table 2 for quantitative results across variants and Figure 7 for visual reconstruction comparison.

Experiment	Params	PSNR ↑	SSIM ↑	MSE ↓	PSNR CV	SSIM CV	MSE CV
Baseline	277	19.599	0.761	0.013885	0.144	0.129	0.881818
Mish	277	15.380	0.670	0.041912	0.233	0.184	1.081851
SSIM Emph.	277	19.073	0.748	0.015453	0.146	0.134	0.839801
Edge Emph.	277	17.301	0.710	0.022825	0.161	0.147	0.703932
ReLU+BN	325	10.135	0.360	0.116998	0.257	0.391	0.683272
ReLU+Dil.+BN	325	10.235	0.381	0.115278	0.262	0.377	0.687837

Table 2: Quantitative results across variants. *CV* denotes the coefficient of variation (std/mean).

When replacing ReLU with Mish activation, the theoretical expectation was a slight gain in perceptual quality, particularly in SSIM, since Mish tends to produce smoother gradients and improved stability. In practice, the qualitative tendency was only partially realized: SSIM stayed relatively close to the baseline, whereas PSNR and MSE diverged more clearly, both in their means and dispersion. Similarly, Adjusting the loss toward SSIM produced the expected smoother appearance. All three metrics remained very close to the baseline, with SSIM the closest of the three, matching the intended perceptual emphasis. However, boosting the edge term sharpened contours as intended, accompanied by a moderate reduction in global similarity. SSIM lay closer to the baseline than PSNR and MSE, reflecting a perceptual and pixelwise trade-off in this setting. Nevertheless, the phenomenon's were modest, and in some cases performance dropped relative to baseline. It is plausible that had these three variants (Mish, SSIM-emphasis, Edge-emphasis) been trained from scratch rather than initialized from pretrained baseline weights, the outcomes might have aligned more closely with the theoretical expectations.

In contrast, the two normalization-based variants diverged more dramatically. Adding Batch Normalization to the baseline, with or without dilated convolutions, was expected to stabilize training and potentially enhance reconstruction by expanding the receptive field. Instead, both variants collapsed, producing very poor reconstructions. The additional parameters and complexity appear to have overwhelmed the extremely lightweight architecture under the limited-sample training regime.

In general, and as illustrated in Table 2 and Figure 7, only the baseline variants and the SSIM / Edge emphasized maintained acceptable reconstruction quality. The other modifications underperformed, highlighting the fragility of ultra-compact networks to architectural changes when training data and time are constrained.

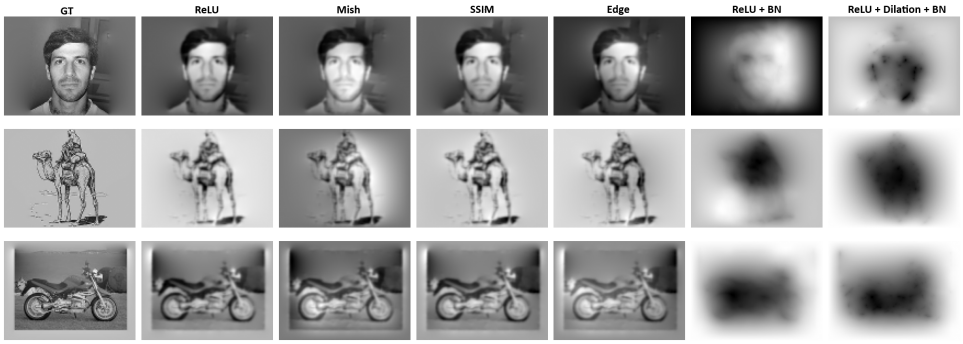

Figure 7: Visual reconstruction comparison: Ground truth vs. ConvNet5 variants.

6.5 Discussion

The experiments underscored how delicate very small CNNs are when applied to event-based image reconstruction. The baseline ReLU network served as a consistent point of reference, yet most alternative configurations failed to outperform it. Variants with Mish activation, SSIM-focused loss, or edge-focused loss followed the expected qualitative trends, but their quantitative results remained modest likely because they were initialized from pretrained baseline weights rather than trained independently from scratch. In contrast, adding Batch Normalization or dilation, which in theory could provide stability and a wider receptive field, led to a collapse in performance. These results suggest that in lightweight models, even small architectural changes can undermine training under limited data and compute conditions. Taken together, the findings indicate that initialization, training design, and data scale are critical factors, with the baseline model proving the most reliable within the scope of this study.

7 Summary

The collective research computationally models how the human visual system achieves a coherent, colorful, and stable perception despite receiving incomplete, low-resolution, and highly dynamic retinal input due to constant eye movements (saccades).

7.1 Overview Topics

This body of work focuses on four fundamental aspects of active visual perception, utilizing neuromorphic approaches:

1. Perceptual Filling-In: Modeling the reconstruction of complete visual surfaces from detected edges. This process supports the isomorphic theory, where activation spreads across the retinotopic map.
2. Efficient Image Reconstruction: Developing compact and efficient neural networks to convert sparse, dynamic event camera data into intensity images.
3. Perceptual Colorization of Peripheral Vision: Reconstructing high-resolution, colorful images from inputs lacking peripheral color information (foveated color maps and achromatic input).
4. Reconstruction of Visually Stable Perception: Integrating successive, dynamic saccadic inputs to generate a stable visual scene, overcoming the instability caused by constant eye movements.

7.2 Innovation Keys

The novelty of this research lies in leveraging biologically inspired systems to solve computational challenges:

- Neuromorphic Modeling: Utilizing Spiking Neural Networks (SNNs), designed through the Neural Engineering Framework (NEF), for high biological plausibility and energy efficiency, allowing networks to be realized on neuromorphic hardware.
- Event-Driven Input: Employing Event Cameras (DVS) to generate high-temporal resolution, sparse data that realistically emulates the rapid intra-saccadic motion captured by the retina, which is crucial for high-quality reconstruction.
- Laplacian-Poisson Pipeline: Implementing a two-phase network for efficient image reconstruction (CNN for Laplacian Prediction + SNN for Poisson Integration), which drastically reduces the number of trainable parameters compared to traditional methods.
- Corollary Discharge (CD) Integration: Introducing CD signals into recurrent ConvLSTM architectures to facilitate the necessary anticipatory adjustments required for maintaining visual stability across saccades.

7.3 Implementation and Results

The models were tested on various datasets (N-MNIST, N-Caltech101, and a synthetic ImageNet-based dataset for stability), employing similarity metrics like SSIM, LPIPS, PSNR, MSE, and CIEDE2000. An overview of these settings and findings:

Topic	**Key Implementation Details**	**Core Results**
Perceptual Filling-In	SNNs solve the Poisson equation by reconstructing images from Laplacians. Two architectures: Feedforward SNN (dense, rapid) and Recurrent SNN (iterative).	Recurrent SNN matched experimental findings on delayed activation spread in V1. Feedforward SNN performed better on large uniform surfaces.
Efficient Image Reconstruction	Compact CNN predicts image Laplacian from event frames, fed into non-trainable Filling-in SNN (Poisson solver). Efficiency from shared-event filters and Mish activation.	Achieved reconstruction with < 100 parameters. A 277-parameter model outperformed a 50k-parameter baseline. Full spiking version (SNN5 at 5 kHz) comparable to non-spiking.
Perceptual Colorization	U-Net trained with GAN + PatchGAN discriminator, combining foveated color channels, Laplacian, and intra-saccadic event data.	Adversarial training improved peripheral colorization; non-adversarial achieved better perceptual scores (SSIM/LPIPS). Events improved acuity and resolution in periphery.
Visually Stable Perception	RNN with ConvLSTM integrates three successive saccades. CD signals relocate event frames to maintain scene stability.	Reconstruction improved with each saccade. Model sensitive to CD noise, consistent with experimental findings on CD importance.

7.4 Applications

The research contributes to:

- Computational Neuroscience: Providing biologically plausible models for reconstructing complex perceptual phenomena (filling-in, colorization, visual stability).
- Neuromorphic Engineering: Developing highly compact, energy-efficient networks for vision processing that can be implemented on neuromorphic hardware (e.g., Loihi, TrueNorth).
- Active Vision Systems: Advancing dynamic models for image reconstruction from event cameras under conditions simulating real-world active viewing.

7.5 Further Research

Potential future directions derived from this work include:

- SNN Enhancement: Converting the most compact CNN models (Shared-event filters) into a fully spiking implementation, possibly using recurrent topology (integrators) for memory.
- Cognitive Integration: Implementing cognitive functions like attention and lateral inhibition in SNNs to explain visual illusions governed by filling-in, beyond basic Poisson integration.
- Dataset Realism: Expanding datasets to include more realistic, varied, or random saccadic motion (instead of fixed three-saccade patterns) to improve model generalization.
- Recurrent Architectures: Incorporating recurrent components like ConvLSTM into the colorization and image reconstruction models to better handle random saccades and model visual working memory/trans-saccadic integration.

8 References

References

[1] Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T. C., Rasmussen, D., and Eliasmith, C. (2014). Nengo: a Python tool for building large-scale functional brain models. *Frontiers in Neuroinformatics*, 7, 48.

[2] Posch, C., Serrano-Gotarredona, T., Linares-Barranco, B., and Delbruck, T. (2014). Retinomorphic event-based vision sensors: bioinspired cameras with spiking output. *Proceedings of the IEEE*, 102(10), 1470–1484.

[3] Cohen Duwek, H., and Tsur, E. E. (2021). Biologically Plausible Spiking Neural Networks for Perceptual Filling-In. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43.

[4] Cohen Duwek, H., and Tsur, E. E. (2022). Biologically Plausible Illusionary Contrast Perception with Spiking Neural Networks. In *IEEE International Conference on Image Processing (ICIP)*.

[5] Cohen Duwek, H., and Spitzer, H. (2018). A Model for a Filling-in Process Triggered by Edges Predicts "Conflicting" Afterimage Effects. *Frontiers in Neuroscience*, 12, 559.

[6] Cohen Duwek, H., and Spitzer, H. (2019). A compound computational model for Filling-in processes triggered by edges: watercolor illusions. *Frontiers in Neuroscience*, 13, 225.

[7] Cohen Duwek, H., Showgan, Y., and Tsur, E. E. (2023). Perceptual colorization of the peripheral retinotopic visual field using adversarially-optimized neural networks. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45(45).

[8] Cohen Duwek, H., Showgan, Y., and Tsur, E. E. (2024). Reconstruction of visually stable perception from saccadic retinal inputs using corollary discharge signals-driven convLSTM neural networks. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0).

[9] Cohen Duwek, H., Slovin, H., and Tsur, E. E. (2022). Computational Modeling of Color Perception with Biologically Plausible Spiking Neural Networks. *PLoS Computational Biology*, 18(10), e1010648.

[10] Cohen, M. A., Botch, T. L., and Robertson, C. E. (2020). The limits of color awareness during active, real-world vision. *Proceedings of the National Academy of Sciences of the United States of America*, 117(24), 13821–13827.

[11] Cohen-Duwek, H., Shalumov, A., and Tsur, E. E. (2021). Image Reconstruction From Neuromorphic Event Cameras Using Laplacian-Prediction and Poisson Integration With Spiking and Artificial Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1333–1341.

[12] Misra, D. (2019). Mish: A self regularized non-monotonic neural activation function. *arXiv preprint*, arXiv:1908.08681.

[13] Rasmussen, D. (2019). NengoDL: Combining deep learning and neuromorphic modelling methods. *Neuroinformatics*, 17(4), 611–628.

[14] Orchard, G., Jayawant, A., Cohen, G. K., and Thakor, N. (2015). Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9, 437.

[15] Gilchrist, I. (2011). *Saccades*. In The Oxford Handbook of Eye Movements.

[16] Haun, A. M. (2021). What is visible across the visual field? *Neuroscience of Consciousness*, 2021(1).

[17] Huang, X., and Paradiso, M. A. (2008). V1 Response Timing and Surface Filling-In. *Journal of Neurophysiology*, 100(1), 539–547.

[18] Isola, P., Zhu, J. Y., Zhou, T., and Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967–5976.

[19] Komatsu, H. (2006). The neural mechanisms of perceptual filling-in. *Nature Reviews Neuroscience*, 7(3), 220–231.

[20] Kuffler, S., Nicholls, J., and Martin, A. (1984). *From Neuron to Brain* (2nd eds.).

[21] Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman and Company.

[22] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science*, 9351, 234–241.

[23] Salin, P. A., and Bullier, J. (1995). Corticocortical connections in the visual system: Structure and function. *Physiological Reviews*, 75(1), 107–154.

[24] Shi, J., and Tomasi, C. (1994). Good features to track. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 593–600.

[25] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-C., and Kong Observatory, H. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems*, 28.

[26] Simchony, T., Chellappa, R., and Shao, M. (1990). Direct Analytical Methods for Solving Poisson Equations in Computer Vision Problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5), 435–446.

[27] Stewart, T. C., and Eliasmith, C. (2014). Large-scale synthesis of functional spiking neural circuits. *Proceedings of the IEEE*, 102(5), 881–898.

[28] Volpert, V. (2014). *Elliptic Partial Differential Equations* (Vol. 104). Springer Basel.

[29] Von Der Heydt, R., Friedman, H. S., and Zhou, H. (2009). Searching for the Neural Mechanism of Color Filling-In. In *Filling-In: From Perceptual Completion to Cortical Reorganization*. Oxford University Press.

[30] Zweig, S., Zurawel, G., Shapley, R., and Slovin, H. (2015). Representation of Color Surfaces in V1: Edge Enhancement and Unfilled Holes.