Reconstruction of Visually Stable Perception from Saccadic Retinal Inputs Using Corollary Discharge Signals-Driven ConvLSTM Neural Networks

Yahia Showgan (yahiashowgan@gmail.com)
Hadar Cohen-Duwek (hadarco@openu.ac.il)
Elishai Ezra Tsur (elishai@nbel-lab.com)

Neuro-Biomorphic Engineering Lab (NBEL), Department of Mathematics and Computer Science
The Open University of Israel, Ra'anana, Israel

Abstract

While subjective visual experiences are remarkably stable and coherent, their underlying data is incomplete and heavily influenced by the eyes' saccadic rhythm. In this work, we show that a deep and recurrent neural network can effectively reconstruct vibrant images from restricted retinal inputs during active vision. Our method includes the creation of a dataset for synthetic retinal inputs, containing intensity, color, and event-camera-generated motion data. We demonstrate the importance of both long-short-term memory and corollary discharge signals to image stabilization and the system's sensitivity to noise, corresponding to recent experimental findings. Our study contributes to the advancement of realistic and dynamic models for image reconstruction, providing insights into the complexities of active visual perception.

Keywords: computational cognition; image reconstruction; visual perception.

Introduction

Striving to provide a stable and coherent visual experience, the human visual system confronts various challenges arising from the heterogeneous characteristics of the retina. The peripheral retina, characterized by a reduced density of cone photoreceptors, exhibits compromised color sensitivity and spatial acuity compared to the central fovea (Lee et al., 2010; Solomon et al., 2005). This reduction in visual fidelity is further compounded by the eccentricity increasing size of the retinal ganglion cells' (RGCs) receptive fields, resulting in diminished visual accuracy in the peripheral visual field (Wandell, 1995).

A significant factor contributing to limited visual representation in the peripheral retina is the RGCs' center-surround receptive fields. These fields realize spatial compression, primarily transmitting edge-related information while neglecting surface details. Consequently, neural encoding within both the fovea as well as the peripheral visual field tends to prioritize contour-based features over a comprehensive representation of the visual scene (Wandell, 1995). The visual system addresses these limitations through the mechanism of saccadic vision. By rapidly directing attention to specific salient points within the visual field, the brain can gather detailed information about important features while conserving computational resources and overcoming the limitations imposed by the retinal architecture. Saccadic eye movements function as a compensatory mechanism, mitigating the constraints imposed by the peripheral retina. This natural adaptation contributes to a more precise comprehension of visual perception in real-world scenarios. However, this continual repositioning poses an enigma: despite the constant motion of the eyes and the resulting changes in retinal input, the subjective experience of the visual world remains remarkably stable and coherent (Bridgeman et al., 1994; MacKay, 1973; Melcher, 2011).

Visual stability, the capacity to sustain a consistent perceptual experience during ocular movements, relies on a complex set of mechanisms. Trans-saccadic integration involves the integration of visual information acquired prior to and following saccades, ensuring a smooth and uninterrupted perceptual experience (Irwin, 1996; Melcher, 2011; Melcher and Colby, 2008). Recent studies have shown that trans-saccadic integration is not merely a collection of isolated snapshots taken during each fixation (Stewart and Schütz, 2018; Wolf and Schütz, 2015). Accordingly, inter-saccadic motion processing works to integrate motion information across successive eye movements, contributing to a perception of stability via gaze correction (Schweitzer and Rolfs, 2021). Saccadic suppression, a neurophysiological process, suppresses visual processing during rapid eye movements, preventing blurring of the visual scene and maintaining stability (Krekelberg, 2010). Additionally, Corollary Discharge signals (CD) play a crucial role by sending a signal concerning the anticipated changes in visual input due to upcoming eye movements. CD signals enable anticipatory adjustments that contribute to the stability of the perceptual experience (Cavanaugh et al., 2016).

Cohen and colleagues (2020) consider the limitations of peripheral as well as saccadic vision under real-world conditions and used a virtual reality (VR) setup with human observers to explore dynamic real-world environments where only the attended parts of the scene were presented in color (using eye-tracking system) while the visual periphery remained desaturated. Surprisingly, a significant number of observers were unaware of drastic alterations to their visual world, challenging the conventional understanding of the richness and accuracy of perceptual awareness in dynamic, active saccadic, and naturalistic viewing conditions (Cohen et al., 2020). Their study therefore demonstrates that even with partial synthetic information derived from each saccade

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

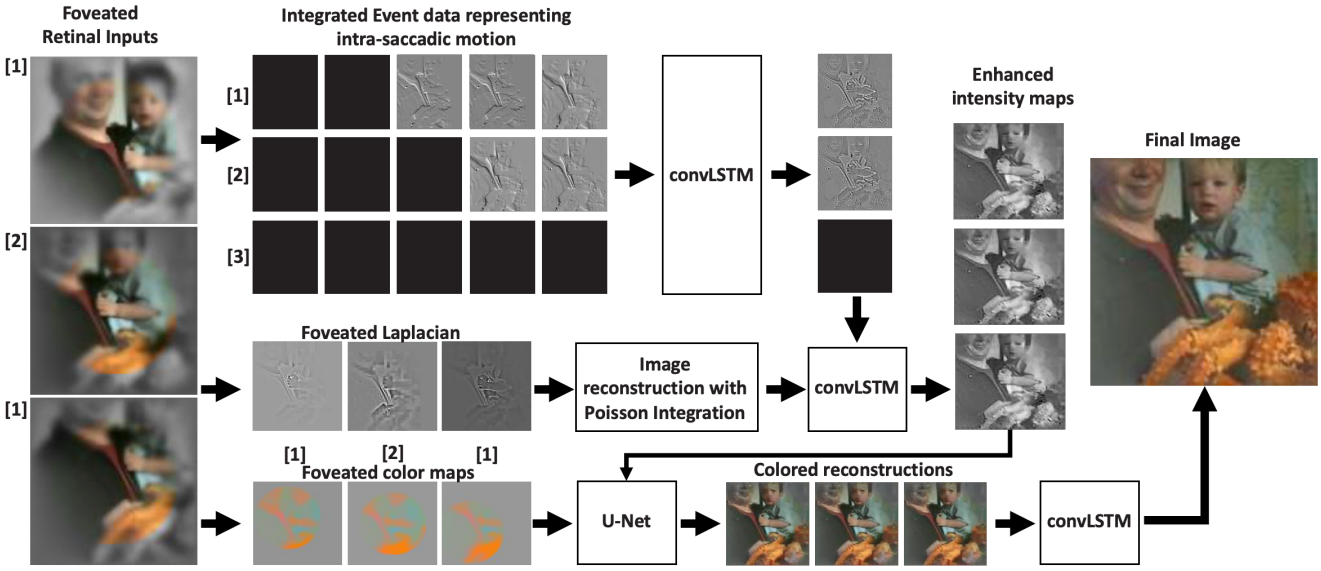

Figure 1: Our model schematic for reconstruction and colorization of stable images from 3 saccadic retinal inputs.

within the scene, visual perception maintains the impression of a rich and colorful world.

A recent study (Cohen Duwek et al., 2023) proposed a computational model that considers both the constrained visual information from retinal input and inter-saccadic motion information to reconstruct a complete and vividly colored image. The researchers utilized a straightforward dataset comprising color images and event data recorded by an event camera during three fixed-target saccades. Although they successfully demonstrated the reconstruction of a sharp and colorful image using their model, its applicability is limited to simpler datasets with fixed-target saccades, and it cannot be extended to more intricate real-world data involving non-fixed target saccades. In this work, we extended this study, allowing the reconstruction of stable images from retinal inputs involving saccadic eye movements towards points of interest, mimicking active vision. To this end, we created a new synthetic dataset comprising retinal inputs with both color and motion attributes. To facilitate image stabilization, we introduced CD signals into a Convolutional Neural Network with long-short-Term Memory (LSTM) (X. Shi et al., 2015), ensuring the creation of visually stable perception from saccadic retinal inputs.

Methods

In this section, we will first briefly describe the generation of retinal inputs, followed by the neural network architecture we used to generate a visually stable perception. The goal here was to synthetically generate retinal inputs during three saccadic eye movements. Retinal input is comprised of an achromatic channel emulating the neural responses of On-Off center-surround retinal ganglion cells (RGCs), a chromatic channel replicating the response of color opponent RGCs, and events generated by an event-camera simulator to simulate rapid intra-saccadic motion. To generate those retinal inputs, we computationally simulated foveated color maps, an intensity channel, and intra-saccadic motion data. The neural network architecture we used to generate a visually stable perception from those retinal inputs was comprised of four phases: 1) Reconstruction of the image intensity from the event frames using convLSTM and convolution layers. convLSTMs are described in detail in (X. Shi et al., 2015); 2) Prediction of the image intensity from the input Laplacian using a Poisson solver layer. The derivation of the input's Laplacian and Poisson-driven image reconstruction is described in detail in (Cohen Duwek et al., 2021). The reproduced events-based intensity map was combined with the predicted intensity from the foveated Laplacian of each saccade through a convolutional network, resulting in an enhanced intensity representation of both outcomes; 3) A U-Net model was used to colorize the foveated color inputs along with the predicted intensity to reconstruct a fully colored image for each saccade. The U-Net model for colorful image reconstruction is described in detail in (Cohen Duwek et al., 2022); and 4) We used the convLSTM neural network to predict the final image, derived from the reconstructed images of all saccades. The framework schematic is shown in **Figure 1.** In the following few sections, we will describe those steps in greater detail.

Generating retinal inputs

As was described above, to generate retinal inputs we computationally simulated foveated color maps, an intensity channel, and intra-saccadic motion data. To this end, Retinal inputs were generated from 935 images, from the ImageNet dataset (Jia Ding et al., 2009). These images were cropped and resized to 200x200 pixels and labeled as scenes.

We generated four images representing gaze points (the initial image, followed by three saccades). Inter-saccadic motion data, which is generated during saccades, was acquired using a simulated event camera (DVS), which generates events upon pixel-level changes in luminance.

The locations for the saccadic movements were strategically chosen to simulate eye movements toward salient features within the image. Identifying these salient points involved utilizing the "Good Features to Track" method (J. Shi and Tomasi, 1994) and applying K Means with k=3 to those points, determining the three most crucial gaze points in each scene. The "Good Features to Track" algorithm is a corner detection method that identifies points in images with significant local intensity changes in multiple directions by evaluating the eigenvalues of the structure tensor at each pixel and selecting points with sufficiently large eigenvalues as distinctive features. We note that other more intricate and biological plausible techniques for selecting gaze points during motion and when prior knowledge is available were recently suggested (Thomas Parr et al., 2021). However, for the task of merely scanning a static scene and reconstructing a large portion of it, our approach suffices. We defined the visual field to be of size 128x128 and each of the four images (the initial image and the three saccades) was centered around one saccadic target location and labeled as saccades. Furthermore, the dataset contains the documented x-y coordinates (relative to the scene image) for each saccadic target location.

Chromatic Channel. To separate the chromatic and achromatic channels, we first converted the saccade images from RGB representation to opponent channels *RG*, *BY*, and the intensity *I* by using the opponent transformation:

$$\begin{pmatrix} RG \\ BY \\ I \end{pmatrix} = M_{opp} \begin{pmatrix} R \\ G \\ B \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \\ a & b & c \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \quad (1)$$

where M_{opp} is the color opponent transformation matrix in which $a = 0.2989, b = 0.587,$ and $c = 0.114$. To simulate the size of the receptive field, being smaller in the fovea and larger towards the periphery, we applied Gaussian filters with different scales on the opponent image (Perry and Geisler, 2002). To replicate the impaired color perception in the peripheral vision, we employed a circular mask positioned at the centers of each channel (RG, and BY) with a radius of 42 pixels, zeroing out all pixels outside the mask.

Achromatic Channel. To simulate the On-Off center-surround response of RGC in the achromatic channel, we applied the discrete Laplacian operator $L = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}$ on the intensity channel I as follows:

$$I_{\text{on-off}} = I * L \approx \Delta I. \quad (2)$$

Intra-saccadic Motion. We generated a video of the visual field moving across 3 gaze points starting from the center of the scene and relocating to the target saccadic locations. We used the V2E DVS emulator (Yuhuang Hu et al., 2021) to generate realistic synthetic events from the intensity frames, producing a list of events for each saccade.

V2E transforms typical video footage, captured by conventional frames-driven cameras, into data resembling the output of a dynamic vision sensor (DVS). V2E generates pixel-level events, where the intensity change surpasses a set threshold. As some saccades are shorter than others, and the event count relies on pixel alterations, the quantity of events between two intensity frames may vary. During each saccade, events occurring between two consecutive event frames were combined and summed into a single event frame. This process resulted in 15-50 event frames for each saccade, with variations based on the saccade duration influenced by the distance between the two gaze points. An event frame was constructed by initializing an empty 128x128 array and processing all events within the time frame between two consecutive timestamps of intensity frames. For each positive event, +1 (increased luminance) was added to the appropriate pixel location, and -1 for each negative event (decreased luminance). Images with videos that did not have enough frames, and therefore event frames, were discarded. We summed each 5 event frames in each saccade, into a single integrated event frame, producing 3-5 integrated event frames per saccade. Zero filled frames were added to maintain the structure of 5 frames per saccade.

Corollary discharge signals. CD signals are neural signals that accompany voluntary movements (here, the saccades), providing the brain with information about the intended motor commands (Melcher, 2011). The CD signals were represented here as translation vectors in Cartesian coordinates (x, y). With those signals, every event frame undergoes translation to the target location relative to the initial image (scene). Both the achromatic and chromatic channels were also adjusted using the identity CD vector.

Model Description

The network was adversarially trained with a discriminator to produce realistic results. Our goal was to reduce the cumulative loss functions across each stage of the Generator while enhancing the Discriminator's capability to distinguish between "fake" and "real" instances. The loss of the generator is the mean of the losses of the results of each saccade, forcing the network to also learn to reconstruct an image from each saccade, as well as utilize several saccades to improve its final reconstructed result.

Intensity prediction. For each saccade, we took the intra-saccadic event frames and used the real-world locations we saved from each video frame's center in the scene, to reallocate the event frames to their scene location in a 200x200 empty image. The image was then cropped to its middle 128x128 pixels. We summed each 5 event frames in every saccade into a single integrated event frame, producing 3-5 integrated event frames per saccade. Zero-filled frames were added to maintain the structure of 5 integrated event frames per saccade. Using the integrated event frames in real-world coordinates (by translation via the CD vectors), we reconstructed the intensity using a

ConvLSTM neural network, followed by convolution layers. Zero-filled frames were filtered out. The network was trained using Mean Absolute Error (MAE) as its loss metric. The foveated intensity Laplacian of each saccade was then used to predict the intensity of the image's center using the Reconstruction Layer (i.e., this layer performs Poisson Integration). The predicted intensity images, derived from each two successive saccades, were combined using a convolutional layer to predict an improved intensity map. The loss was calculated at this stage using MAE loss between the ground truth (GT) and the predicted intensity.

Image colorization with U-Net. Using the predicted intensity along with the foveated opponent (OPP) color channels as inputs, a U-Net architecture network was used to predict the fully colored image (Isola et al., 2017; Ronneberger et al., 2015).
Two loss functions were used on the results of the colorization:

$$\mathcal{L}_{opp} = \lambda_{opp}(MAE(O_1, \hat{O}_1) + MAE(O_2, \hat{O}_2)) \quad (3)$$

where MAE was derived from the predicted and the original image's color channels.
The second loss includes both SSIM a LPIPS metrics (described below in detail):

$$\mathcal{L}_{RGB} = \lambda_{ssim}\left(1 - SSIM(I_{RGB}, \hat{I}_{RGB})\right) + \lambda_{lpips} LPIPS(I_{RGB}, \hat{I}_{RGB}) \quad (4)$$

where $I_{RGB} = opp2rgb(O_1, O_2, O_3)$ and $\hat{I}_{RGB} = opp2rgb(\hat{O}_1, \hat{O}_2, , \hat{O}_3)$.

Saccadic integration. At this point, we possess four reconstructed color images, each predicted through a distinct saccade. By utilizing ConvLSTM layers in conjunction with additional convolution layers, we generated our final result. With each saccade, predictions were integrated, and used to enhance the result accuracy.

Generative adversarial network (GAN). We trained our reconstruction and colorization convolutional neural networks using an adversarially trained generator and a discriminator, intending to optimize the loss functions previously mentioned. We trained two different models, one using all the inputs mentioned (Foveated OPP, and integrated event frames), and the second model trained without the integrated event frames (i.e., no events). Both models were trained using input from three successive saccades. We employed an adversarially trained discriminator, denoted as D, utilizing a convolutional PatchGAN classifier (Isola et al., 2017). The primary objective of *D* is to discern the "fake" images generated by the trained Generator, denoted as *G*. The Generator produces reconstructed outputs, denoted as y, which are intentionally crafted to be indistinguishable from "real" images denoted as x. The adversarial interplay involves *D* striving to maximize its ability to identify fake images, while *G* endeavors to minimize this detectability. The GAN loss was computed using:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[logD(x, y)] + \lambda_D \mathbb{E}_{x,G(x)} \log[(1 - D(x, G(x))] \quad (5)$$

Here, x represents the retinal input, y is the ground truth image transformed into the opponent color space, $\mathbb{E}$ denotes the expected value, and λ_D serves as a gain parameter. The first term of Equation (5) involves presenting GT examples to the discriminator, while the second term introduces fake examples generated by the Generator.
In conclusion, the ultimate end-to-end minimization objective is expressed as:

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \mathcal{L}_{RGB} + \mathcal{L}_{opp} + \mathcal{L}_{O_3} + \mathcal{L}_{\nabla^2} \quad (6)$$

Implementation details

The implementation of the model utilized TensorFlow and was trained on an NVIDIA A100 GPU featuring 80GB of RAM. The dataset, which was created from 935 ImageNet images, was partitioned into three sets: 70% for training, 15% for validation, and 15% for testing. We set $\lambda_{\nabla^2} = 100$, $\lambda_{PI} = 25$, $\lambda_{opp} = 150$, $\lambda_{ssim} = 100$, and $\lambda_D = 10$ over the whole experiment, for the loss functions. Each model tested was trained for 200 Epochs, with a batch size of 8, and an initial learning rate of 0.001.

Similarity metrics

In this work, we used four metrics to assess our model. SSIM serves as a metric for image similarity, with a higher score denoting increased likeness to the reference image. LPIPS (Zhang et al., 2018) functions as a perceptual metric for assessing image quality, wherein a lower score signifies a more favorably perceived image quality. Peak Signal-to-Noise Ratio (PSNR) quantifies the quality of a reconstructed signal by comparing it to the original, measuring the ratio of peak signal power to noise power. Additionally, CIEDE2000 (Luo et al., 2001) is a color difference metric that gauges perceptual color disparities between samples, considering attributes such as lightness, chroma, and hue.

Results

Figure 2 illustrates the original ground truth (GT) images alongside the reconstructed images produced by two training methods: with and without event data. We observe a clear improvement in the output quality with each saccade, going from a somewhat blurry image to a more detailed one over time. Notable improvements are especially visible in the colors. For example, in the bird picture, the top right corner lacks color in the first saccade but becomes gradually greener in subsequent saccades. Similarly, in the second

image, the bottom right corner starts with missing color but becomes more vibrant with each saccade. The facial features of the child also become sharper and more detailed in each saccade, evolving from an initially blurry appearance. Table 1 shows the scores of the SSIM, LPIPS, PSNR, and CIEDE2000 similarity metrics. The scores for the input (Input column, Table 1) were computed by transforming the masked color channels (RG, BY) and the reconstructed intensity channel (*I*, after performing Poisson integration on this channel) into the RGB color space. Table 1 indicates that the outcomes for input involving three saccades yield the highest scores, showing improved SSIM and LPIPS scores with events. Beginning with an input of the first saccade, we observe enhancements in the reconstruction quality with each additional saccade. Interestingly, better results were achieved without events in the PSNR and CIEDE2000 metrics, indicating the importance of events for perceptual similarity.

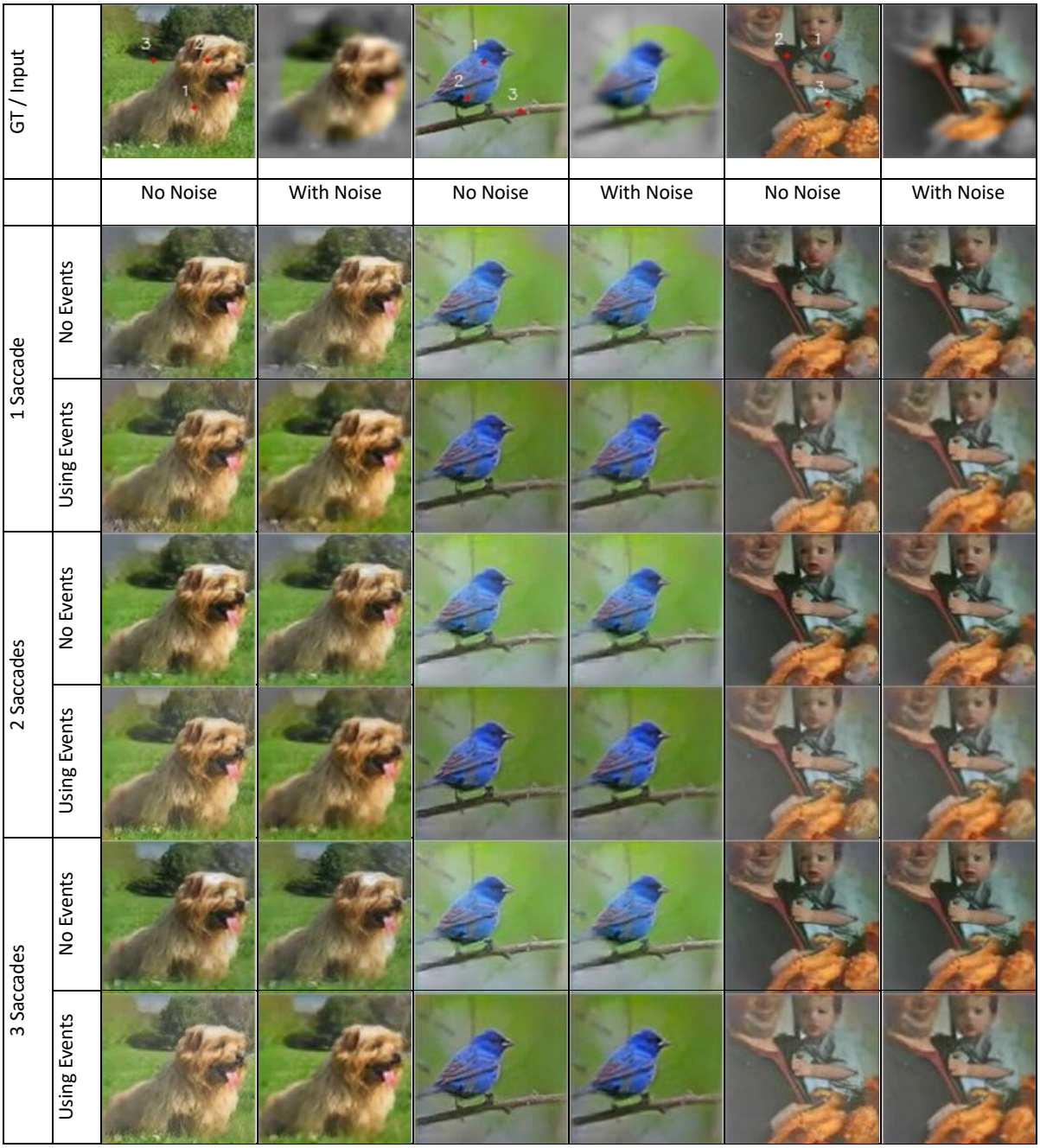

Figure 2: Chosen image reconstructions with and without events, with and without noise and when 1-3 saccades were used for reconstruction.

		1 Saccade				2 Saccades				3 Saccades			
		No Events		With Events		No Events		With Events		No Events		With Events	
	Input	No Noise	With Noise	No Noise	With Noise	No Noise	With Noise	No Noise	With Noise	No Noise	With Noise	No Noise	With Noise
SSIM (%)	59.42	75.62	71.65	77.86	73.75	78.39	73.17	81.05	75.08	80.63	73.68	**82.78**	75.88
LPIPS (%)	53.54	30.32	31.69	27.35	31.21	26.22	28.01	24.13	28.52	22.92	25.27	**21.73**	26.67
PSNR (dB)	17.01	23.22	22.74	23.57	23.24	23.72	23.08	24.04	23.5	**25.19**	24.15 dB	25.17	24.39 dB
CIEDE2000	13.72	6.51	6.92	6.52	6.7	6.26	6.59	6.10	6.38	**5.3**	5.75	5.31	5.66

Table 1: Reconstruction image evaluation metrics. Higher values of SSIM and PSNR, and lower values of LPIPS and CIEDE2000 indicate higher similarity.

Sensitivity to Noise

We further tested the sensitivity of the CD signals to noise by introducing Gaussian noise (standard deviation (σ) of 1 and a mean of 0, ranging from -4 to 4 pixels) to the CD vectors, affecting both the relocation process of the integrated event frames, as well as the foveated saccade images. The noise introduced a random displacement of up to 4-pixel in each axis for each gaze point. The randomness in the noise distribution influences the precise positioning of the inputs within the scene. Table 1 demonstrates a decreased quality across all evaluation scores, suggesting that predictions are heavily influenced by CD noise. Furthermore, Figure 2 demonstrates that images reconstructed with noisy CD vectors exhibit a slight blurriness.

Discussion

In this study, we demonstrate that a deep recurrent neural network can effectively reconstruct vibrant images from restricted retinal inputs during active vision, involving saccadic eye movements directed towards points of interest in the scene. Our method includes the creation of a synthetic dataset that incorporates retinal inputs containing information on intensity, color, and motion. Significantly, the incorporation of both Long-Short-Term-Memory (LSTM) and CD signals stands out as a crucial element in the model, contributing to the enhancement of image stabilization through eye movements. The model can reproduce the results observed by Cohen and colleagues (Cohen et al., 2020), indicating that observers can perceive a colorful scene even when colors are eliminated from their peripheral field of view in each saccade. A more recent study (Cohen Duwek et al., 2023) demonstrates that peripheral color can be perceived (predicted) based on achromatic input in the peripheral visual field. However, this model was designed for fixed non-saccadic images. In contrast, our demonstration illustrates that deep neural networks can achieve both colorization of the peripheral visual field and stable perception. This is achieved by incorporating saccadic integration, implemented as a Long Short-Term Memory (LSTM) component, and utilizing the Corollary Discharge (CD) signal for stabilization. In this scenario, convolutional LSTMs (X. Shi et al., 2015) serve the role of a functional visual memory (Stewart and Schütz, 2018) in the context of trans-saccadic integration. This function potentially plays a crucial role in achieving both high visual accuracy in the peripheral region and the stabilization of vision across saccades.

We assessed two configurations of the model— one with event data (representing intersaccadic motion) and one without event data— using four distinct similarity matrices (SSIM, LPIPS, PSNR, CIEDE2000). Interestingly, the model configuration with events outperformed the model without events in terms of SSIM and LPIPS. However, it exhibited lower performance in both PSNR and CIEDE2000, indicating the importance of events' contribution to perceptual similarity. Additionally, we introduced noise to the CD vectors, demonstrating the model's sensitivity to lesser controlled conditions. Recognizing the influence of noise on the precision of relocation is crucial for interpreting the model's performance in the presence of more unpredictable inputs. Our findings uncover the importance of accurate Corollary Discharge (CD) signals. This outcome aligns with an experimental study conducted by Cavanaugh and colleagues (Cavanaugh et al., 2016). In their research, they eliminated the CD signal in old-world monkeys and illustrated the crucial role of this signal in the monkeys' perception during saccadic movements. Our model, currently trained exclusively on datasets with three saccades, could benefit from an expansion of the dataset to include varied numbers of saccades. This adjustment will probably further enhance the model's adaptability to diverse scenarios, allowing it to handle a broader range of input variations and better simulate the complexities of real-world visual processing. While our model showed significant capabilities in simulating retinal input based on three saccades, introducing variability in the training set could uncover patterns and features that contribute to improved generalization across various eye movement scenarios.

This research represents a significant step forward in understanding and replicating the dynamics of active vision within the context of image reconstruction. By integrating saccadic integration, CD signals, LSTM, and colorization, our study contributes to the advancement of realistic and dynamic models for image reconstruction, providing insights into the complexities of visual perception, particularly in scenarios where active vision processes are essential.

Acknowledgments

The authors would like to thank the members of the Neuro-Biomorphic Engineering Lab at the Open University of Israel for the insightful discussions.

References

Cavanaugh, J., Berman, R. A., Joiner, W. M., and Wurtz, R. H. (2016). Saccadic Corollary Discharge Underlies Stable Visual Perception. *Journal of Neuroscience*, *36*(1), 31–42.

Cohen Duwek, H., Showgan, Y., and Ezra Tsur, 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).

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

Cohen Duwek, H., Slovin, H, and Ezra Tsur, E. (2022) Computational Modeling of Color Perception with Biologically Plausible Spiking Neural Networks. P*LoS Computational Biology*, 18(10): e1010648

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.

Irwin, D. E. (1996). Integrating information across saccadic eye movements. *Current Directions in Psychological Science*, *5*(3), 94–100.

Isola, P., Zhu, J. Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-January*, 5967–5976.

Krekelberg, B. (2010). Saccadic suppression. *Current Biology*, *20*(5), 228–229.

Lee, B. B., Martin, P. R., and Grünert, U. (2010). Retinal connectivity and primate vision. *Progress in Retinal and Eye Research*, *29*(6), 622–639.

Luo, M. R., Cui, G., and Rigg, B. (2001). The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application*, *26*(5), 340–350.

Melcher, D. (2011). Visual stability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1564), 468–475.

Melcher, D., and Colby, C. L. (2008). Trans-saccadic perception. *Trends in Cognitive Sciences*, *12*(12), 466–473.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9351*, 234–241.

Schweitzer, R., and Rolfs, M. (2021). Intrasaccadic motion streaks jump-start gaze correction. *Science Advances*, *7*(30).

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

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*.

Solomon, S. G., Lee, B. B., White, A. J. R., Rüttiger, L., and Martin, P. R. (2005). Chromatic Organization of Ganglion Cell Receptive Fields in the Peripheral Retina. *Journal of Neuroscience*, *25*(18), 4527–4539.

Stewart, E. E. M., and Schütz, A. C. (2018). Optimal trans-saccadic integration relies on visual working memory. *Vision Research*, *153*, 70–81.

Thomas Parr, Noor Sajid, Lancelot Da Costa, M. Berk Mirza, Karl J. Friston. (2021) Generative Models for Active Vision. Front. Neurorobot., 13 April 2021, Volume 15 - 2021

Wandell, B. A. (1995). *Foundations of vision*. Sinauer Associates.

Wolf, C., and Schütz, A. C. (2015). Trans-saccadic integration of peripheral and foveal feature information is close to optimal. *Journal of Vision*, *15*(16), 1–1.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.