Feedback-correcting ConvLSTM-driven Neural Model for Stable Saccadic Visual Perception

Hadar Cohen-Duwek (hadarco@openu.ac.il)
Yisrael Clark (yisrael.clark@gmail.com)
Elishai Ezra Tzur (elishai@nbel-lab.com)

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

Abstract

The brain utilizes corollary discharge signals to anticipate the visual consequences of saccadic eye movements and provide a coherent visual perception. However, discrepancies between a saccade's predicted and actual sensory outcomes challenge the brain's capacity to maintain visual stability. In this work, we introduce a comprehensive computational framework for visual perception incorporating a feedback corrective mechanism that dynamically adjusts predictions based on sensory discrepancies. We show that this feedback mechanism refines internal world models, and provides robust performance with an increasing number of saccades. Our results highlight the delicate balance between the benefits and vulnerabilities of predictive feedback systems supporting and extending current theories of sensory prediction and visual stability.

Keywords: visual perception; visual stability; saccadic eye movements; corollary discharge

Introduction

Visual perception adapts to the continuously changing environment by leveraging specialized retinal structures and active vision strategies. Among these structures, the retina exhibits spatial heterogeneity, with high acuity and color vision concentrated in the fovea, decreasing rapidly toward the periphery. Peripheral vision is characterized by reduced color sensitivity, increased blur due to the larger receptive fields of retinal ganglion cells, and dominance of achromatic information (Lee et al., 2010; Solomon et al., 2005). To overcome these limitations, the visual system employs active vision approaches, such as saccadic eye movements with which the focus of attention is rapidly shifted across the visual field, allowing to sample the scene with high resolution in different locations (Liversedge and Findlay, 2000).

While essential for active vision, saccadic eye movements disrupt the perception of a continuous world and thus challenge visual stability. To address this, the brain employs two complementary mechanisms: saccadic integration and Corollary Discharge (CD).

Saccadic integration combines information from different retinal regions and utilizes CD signals to transform retinal coordinates into spatiotopic, world-centered coordinates (Irwin, 1996; Melcher, 2011). This process heavily relies on working memory, which plays a crucial role in combining visual information from saccades at different locations (Cronin and Irwin, 2018). By temporarily storing and updating details of the visual scene, working memory enables the integration of information from multiple fixations to create a coherent, sharp, and color-rich representation of the visual environment. Furthermore, a predictive colorization mechanism helps infer missing details, such as peripheral colors, when sensory input is incomplete (Cohen et al., 2020; Cohen-Duwek et al., 2022).

CD complements saccadic integration by providing a predictive signal alongside motor commands, anticipating the visual consequences of saccades and correcting mismatches between expected and actual input (Cavanaugh et al., 2016; Crapse and Sommer, 2008). Together, these mechanisms maintain a stable and coherent perception of the environment.

However, when discrepancies arise between a saccade's predicted and actual sensory outcomes, errors can challenge the brain's capacity to maintain visual stability (Figure 1). Such mismatches, caused by inaccuracies in motor commands, processing delays, or unexpected changes in the visual environment, can lead to perceptual instability, making it harder to locate targets or recognize objects accurately (Bansal et al., 2018).

Predictive coding (Huang and Rao, 2011) is critical in addressing these mismatches. By comparing predictions, informed by the corollary discharge (CD) signal, with actual sensory feedback, the brain generates error signals to update its internal models. This iterative process refines future predictions and facilitates rapid error correction, ensuring perceptual stability and a coherent visual experience despite frequent saccadic disruptions (Bansal et al., 2018; Ford and Mathalon, 2019).

In addition, the brain uses a predictive colorization mechanism to infer missing peripheral colors when sensory input is incomplete. For example, using virtual reality, it was shown that most participants remained unaware when only attended areas were in color, highlighting the brain's ability to maintain the illusion of a rich, colorful world through prediction (Cohen et al., 2020; Cohen-Duwek and Tsur, 2022).

Recent studies have explored computational neural models for reconstructing and colorizing images from retinal input Cohen Duwek and Ezra Tsur, 2021. For instance, (Cohen Duwek et al., 2023) proposed a U-Net-based model for image colorization using retinal input, while (Showgan et al., 2024) extended this approach with ConvLSTM layers and CD signals to integrate information across saccades. While these models effectively reconstruct images, they do not address saccadic errors, where the predicted CD target differs from

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

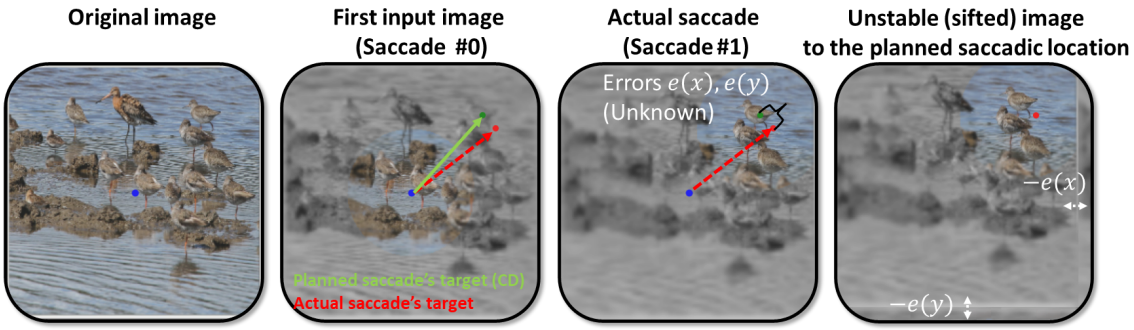

Figure 1: Illustration of the feedback corrective mechanism addressing saccadic errors by comparing visual representations before and after eye movements.

the actual target. These errors can disrupt visual stability and affect alignment between predicted and actual visual input.

To address these limitations, we introduce a feedback corrective mechanism that dynamically adjusts predictions based on sensory discrepancies. This ensures a more accurate alignment between the predicted and actual target locations. This mechanism effectively handles prediction errors, refines internal world models, and provides robust performance even with an increasing number of saccades. This work advances our understanding of visual stability by addressing mismatches caused by saccadic errors and demonstrates a generalizable approach for integrating CD signals with dynamic feedback.

Methods

We developed a computational model that simulates human-like visual processing by using foveation transformation, which maintains high detail in the central focus area while reducing detail in the periphery, combined with opponent-color-space encoding. The approach utilizes a U-Net–based architecture (Ronneberger et al., 2015) with ConvLSTM cells (Shi et al., 2015), enhanced by an adversarial loss for more realistic image prediction (Figure 2 A & B). Additionally, a neural network-based error prediction mechanism was integrated to correct saccadic errors (Figure 2 A). Below, we describe each component, detailing data preparation, the emulation of the retinal input, the neural network architecture, and the training and evaluation metrics.

Data and Preprocessing

Our primary dataset was derived from ImageNet-V2 (Recht et al., 2019). Each image was resized to 128×128 pixels and normalized to the [0,1] range. We randomly selected a subset for training, validation, and testing (1,000 images for training, 248 images for validation, and 498 images for testing) to demonstrate the model's generalization capacity.

Retinal input

To replicate the color coding employed by the visual system, each RGB frame was transformed into the Opponent (Van De Sande et al., 2009; Wandell, 1995) color space (RG, BY, L) using:

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

where M_{opp} is the color opponent transformation matrix in which $a = 0.2989$, $b = 0.587$, and $c = 0.114$.

To simulate human foveal-peripheral visual distinction, we applied a foveation transform at each time step. A circular mask with a radius of 32 pixels was generated around a current fixation point, preserving high-resolution details within that circle while zeroing out the colors and blurring intensity in the periphery (Jiang et al., 2015; Perry and Geisler, 2002). The fixation point shifted over time, producing a sequence of foveated frames.

Corollary discharge for image stabilization

Corollary discharge (CD) signals are neural signals associated with voluntary movements. In the case of saccades, CD signals inform the brain about intended motor commands (Melcher, 2011). Here, these signals were represented as translation vectors (x, y) in Cartesian coordinates for simplification. Each frame was translated to the target location relative to the initial scene using these vectors.

Saccadic errors Saccadic eye movements are rarely perfectly accurate, resulting in a discrepancy between the intended and actual landing position known as saccadic error (Collewijn et al., 1988). Figure 1 illustrates this error and its relation to CD signals. The "Actual Saccade" panel shows the intended saccade target (green dot), predicted by the CD signal, and the actual landing point (red dot). The difference between these points represents the saccadic error, quantified by horizontal $e(x)$ and vertical $e(y)$ components. The "Unstable (sifted) image" panel demonstrates the retinal image when the saccade landed on the planned target, achieved by shifting the actual image by the inverse of the error vectors. This shift, however, results in a blurred image, suggesting the

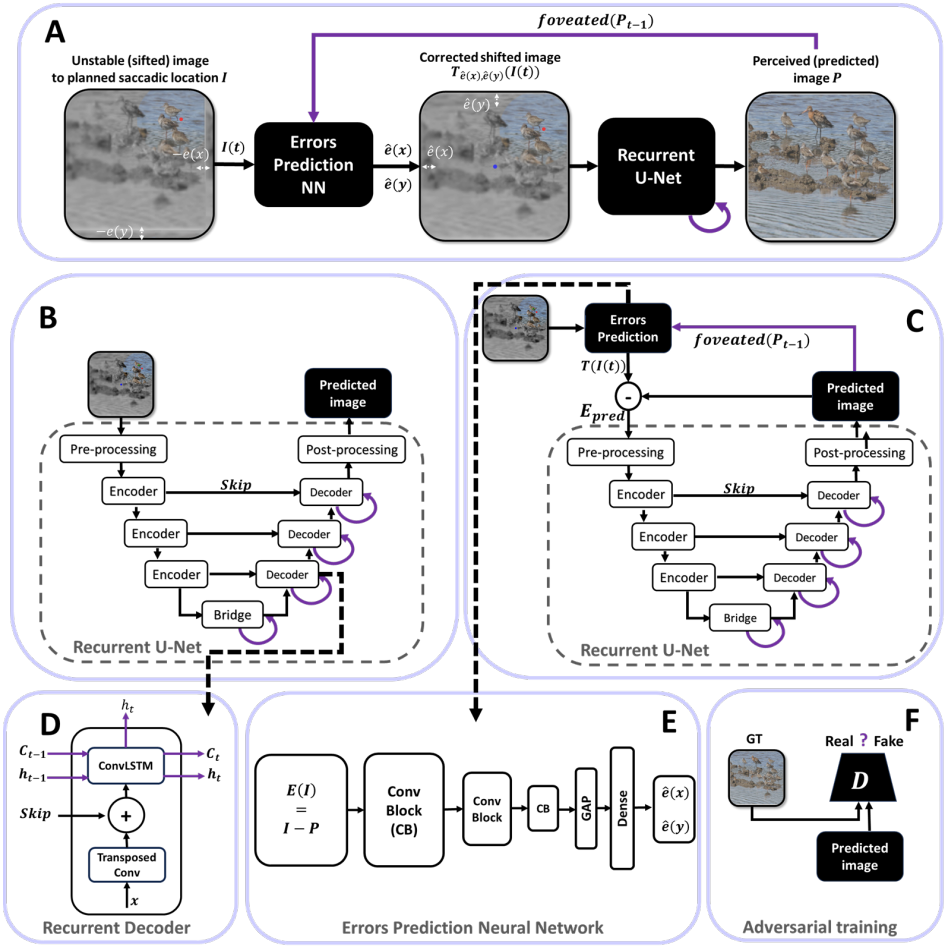

Figure 2: Illustration of the proposed feedback mechanism for dynamic error correction in saccadic predictions. (A) Feedback correction loop. (B) Recurrent U-Net architecture for *Non-Predictive-Net*. (C) Recurrent U-Net architecture for *Predictive-Net* with error prediction mechanism. (D) The architecture of the recurrent decoder. (E) The errors prediction Neural network. (F) An illustration of the adversarial training.

visual system's predictive mechanisms based on CD signals are imperfect, highlighting the challenge of maintaining visual stability despite saccadic inaccuracies.

We incorporated saccadic error into the model by sampling horizontal $e(x) = \Delta x$ and vertical $e(y) = \Delta y$ offsets from normal distributions with means of zero and standard deviations resulting in a range of variations between -10 and 10 pixels.

Saccadic integration using Recurrent U-Net

Convolutional Long Short-Term Memory (ConvLSTM) is a type of recurrent neural network that incorporates convolutional operations into its framework, enabling it to capture both spatial and temporal dependencies in sequential data (Shi et al., 2015). In our model, ConvLSTM was used to account for saccadic integration by modeling how spatial information evolves over time during eye movements. The ConvLSTM integrates spatial and temporal features by processing sequences of image frames, making it well-suited for modeling saccadic integration, where visual information from rapid eye movements is combined over time. Our main goal in this work is to add a prediction network comprised of a recurrent U-Net.

Figure 2 (B and C) illustrates the architecture of a Recurrent U-Net for image prediction (reconstruction), designed to integrate spatial and temporal information. The network begins with pre-processing, where the input is split into intensity and opponent color channels. The intensity is processed through convolutional layers to extract spatial features, and the intensity and color channels are fused and refined to form the input for the encoder. It then passes through a series of hierarchical encoders that downsample and extract increasingly abstract spatial features. At the bridge, a ConvLSTM layer incorporates recurrent connections, enabling the integration of temporal context and spatial features. In the decoding stages, the network progressively upsamples features, with each decoder using recurrent connections via ConvLSTM to refine outputs by leveraging temporal dependencies, Figure 2D. Skip connections merge spatial details from cor-

responding encoder layers, ensuring accurate feature reconstruction. The network ends with post-processing, where the reconstructed intensity is refined with a skip connection to retain details, and the opponent color channels are activated with a *tanh* function to produce the final predicted image. The use of recurrent connection simulates saccadic integration, combining spatial and temporal information to enhance prediction.

Saccadic errors prediction and correction

The shift prediction neural network (NN) model estimates positional shifts $(\Delta x, \Delta y)$by processing error signals between predicted and observed visual inputs. The model begins with convolutional layers to extract hierarchical spatial features, followed by a global pooling layer to condense this information. These features are then passed through fully connected layers, ending in a linear output layer that predicts the 2D shift, Figure 2E. This design effectively models spatial discrepancies, enabling tasks such as visual alignment and error correction in cognitive systems.

Figure 2A illustrates how our system iteratively refines image predictions using a combination of errors correction NN and a recurrent U-Net. The difference between the model's prediction and the current foveated input provides an error signal, which is fed into an errors-prediction network (errors prediction NN) to estimate the saccadic errors $(\hat{e}(x), \hat{y}(x))$. The error signal is calculated as the difference between the current retinal input $I(t)$ and a foveated transformation of the predicted image generated from the previous time step's input. This foveation transformation, $foveated(P_{t-1})$, is applied to the predicted image using the provided CD signal (excluding errors) to facilitate a direct comparison between the expected and actual visual input. The image is then adjusted by translating it in the direction opposite to the errors $I_T = T_{\hat{e}(x),\hat{e}(y)}(I(t)) = \mathrm{T} \cdot I$ where:

$$\mathrm{T} = \begin{bmatrix} 1 & 0 & -\hat{e}(x) \\ 0 & 1 & -\hat{e}(y) \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

The newly shifted image I_T is then passed to a recurrent U-Net, which updates its hidden state and produces a more accurate reconstruction.

Generative adversarial training

We implemented a PatchGAN-like (Isola et al., 2017) discriminator D to provide adversarial feedback, leading to realistic output in RGB space (Figure 2F). We jointly trained the recurrent U-Net model G, the error-prediction network E, and the PatchGAN discriminator. To ensure accurate and perceptually consistent predictions, the training process optimized a combination of loss functions:

Notation. Let $\mathrm{RGB}_{\mathrm{GT}}$ denote the ground truth (GT) frames in the RGB domain, while $\mathrm{RGB}_{\mathrm{pred}}$ are the frames predicted by our generator. We similarly denote L_{GT} and $\mathrm{Opp}_{\mathrm{GT}}$ as the intensity and opponent color channels of the GT frames, respectively, and L_{pred} and $\mathrm{Opp}_{\mathrm{pred}}$ as the corresponding predicted channels. The vectors $\hat{e}_x, \hat{e}_y$ and $e_x^{\mathrm{gt}}, e_y^{\mathrm{gt}}$ refer to the predicted and ground truth of the saccadic error vectors, respectively. We define temporal weights w_t for each frame $t \in \{1, \ldots, T\}$ in a sequence where $w_t = \frac{2t}{T(T+1)}$

(1) Error vector loss. Enforces accurate saccadic error predictions by minimizing the mean absolute error (MAE) between predicted and ground-truth motion vectors:

$$\mathcal{L}_{\mathrm{error\_vector}} = \mathrm{MAE}(\hat{e}_x, e_x^{\mathrm{gt}}) + \mathrm{MAE}(\hat{e}_y, e_y^{\mathrm{gt}}). \tag{3}$$

(2) Intensity and opponent color losses. Encourage fidelity in intensity (L) and color opponent (Opp) channels by penalizing deviations from ground truth:

$$\mathcal{L}_L = \sum_{t=1}^{T} w_t \cdot \mathrm{Mean}\big(|L_{\mathrm{pred},t} - L_{\mathrm{GT},t}|\big), \tag{4}$$

$$\mathcal{L}_{\mathrm{Opp}} = \sum_{t=1}^{T} w_t \cdot \mathrm{Mean}\big(|\mathrm{Opp}_{\mathrm{pred},t} - \mathrm{Opp}_{\mathrm{GT},t}|\big). \tag{5}$$

(3) Reconstruction (SSIM) and perceptual (LPIPS) losses. Preserve structural and perceptual similarity between the predicted and ground truth frames:

$$\mathcal{L}_{\mathrm{reconstruction}} = 1 - \mathrm{SSIM}\big(\mathrm{RGB}_{\mathrm{GT}}, \mathrm{RGB}_{\mathrm{pred}}\big), \tag{6}$$

$$\mathcal{L}_{\mathrm{SSIM},L} = 1 - \sum_{t=1}^{T} w_t \cdot \mathrm{SSIM}\big(L_{\mathrm{pred},t}, L_{\mathrm{GT},t}\big), \tag{7}$$

$$\mathcal{L}_{\mathrm{LPIPS}} = \sum_{t=1}^{T} w_t \cdot \mathrm{LPIPS}\big(\mathrm{RGB}_{\mathrm{GT},t}, \mathrm{RGB}_{\mathrm{pred},t}\big). \tag{8}$$

(4) Adversarial loss. Encourages realism through a minimax game between generator G and discriminator D:

$$\begin{aligned} \mathcal{L}_{\mathrm{GAN}}(G,D) = {} & \mathbb{E}_{(x,y)}[\log D(x,y)] \\ & + \lambda_g \, \mathbb{E}_{(x,G(x))}[\log(1 - D(x, G(x)))]. \end{aligned} \tag{9}$$

(5) Generator objective. Combines all relevant losses (with α_i as weighting factors):

$$\begin{aligned} \mathcal{L}_{\mathrm{Generator}} = {} & \mathcal{L}_{\mathrm{GAN}} + \alpha_1 \mathcal{L}_L + \alpha_2 \mathcal{L}_{\mathrm{Opp}} + \alpha_3 \mathcal{L}_{\mathrm{SSIM},L} \\ & + \alpha_4 \mathcal{L}_{\mathrm{reconstruction}} + \alpha_5 \mathcal{L}_{\mathrm{LPIPS}}. \end{aligned} \tag{10}$$

where $\alpha_1 = 100$, $\alpha_2 = 2000$, $\alpha_3 = 25$, $\alpha_4 = 25$ and $\alpha_5 = 100$.

(6) Final training objective. The complete optimization unifies all losses to ensure perceptually consistent and accurate predictions:

$$G^* = \arg\min_G \max_D \min_E \Big(\mathcal{L}_{\mathrm{Generator}}(G,D,E) + \mathcal{L}_{\mathrm{error\_vector}}(E) \Big) \tag{11}$$

This framework ensures saccadic error estimation, color and brightness consistency, structural correctness, and adversarial realism in the predicted frames.

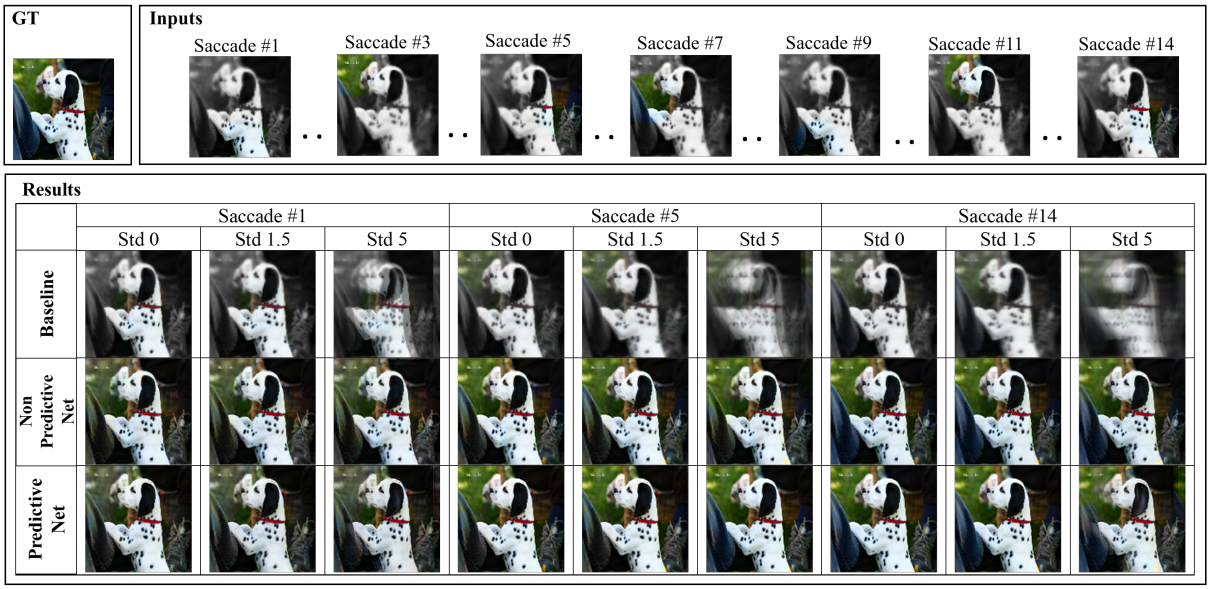

Figure 3: Predictive model performance under saccadic errors: comparison of *Predictive-Net*, *Non-Predictive-Net*, and *Baseline*

Evaluation

We compare the predictive feedback model (Figure 2C), which incorporates feedback saccadic error correction (i.e., *Predictive-Net*), to a model that relies solely on retinal input and employs a recurrent U-Net to integrate saccades and reconstruct a stable image (i.e., *Non-Predictive-Net*) (Figure 2B). Additionally, we evaluated both models using a baseline method (i.e., *Baseline*). The baseline processes each saccadic input by averaging the inputs over time.

The baseline computes an averaged image for each saccade to provide saccadic integration representation over time. For the first retinal input $(t = 0)$ the averaged image is simply the first frame of the input sequence, as no prior frames are available. For subsequent saccades $(t > 0)$ the averaged image is calculated as the mean of all frames from the start of the sequence up to the current saccade.

Both the *Predictive-Net* and the *Non-Predictive-Net* were trained on sequences of 5 saccades (each sequence comprising 5 frames per image in the batch) over 200 epochs under high-noise conditions (standard deviation = 5). We evaluated our two models as well as *Baseline* using two metrics: Structural Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018). SSIM measures image similarity by comparing local patterns of pixel intensities between two images, considering luminance, contrast, and structure, with higher scores indicating greater resemblance to the reference image. LPIPS, a perceptual metric, assesses image quality by comparing deep feature representations learned by a neural network, with lower scores indicating better-perceived quality as judged by human observers.

Results

Figure 3 demonstrates the performance of a predictive model in reconstructing visual information under conditions influenced by saccadic errors, with (*Predictive-Net*) and without (*Non-Predictive-Net*) feedback mechanisms, compared to the *Baseline*. The first row represents the ground truth image (GT) and the retinal images captured from different saccade targets. Subsequent rows present reconstructions generated by the predictive model at various noise levels, representing saccadic errors. Each row illustrates the impact of the feedback and noise on the model's output, with columns comparing the baseline, feedback-enabled, and feedback-disabled models. It can be seen that the baseline reconstructions exhibit lower visual quality compared to the predictive model outputs (Figure 3). Baseline reconstructions appear more degraded, especially as noise levels increase, highlighting the model's inability to handle saccadic errors. In contrast, the feedback-enabled model produces the most accurate reconstructions, closely resembling the ground truth, while the model without feedback performs better than the baseline but lacks the clarity of the feedback-enabled results.

Figure 4A shows the SSIM (first column) and the LPIPS (second column) as a function of the number of saccades (x-axis) for four different noise levels (std = 0, 1.5, 2.5, 5). Each row corresponds to one noise level, and we include the three models: *Predictive-Net*, *Non-Predictive-Net* and a *Baseline*. These plots reveal that, under lower noise (e.g., std = 0 or 1.5), *Predictive-Net*'s SSIM steadily increases with more saccades, eventually surpassing the other two methods, while its LPIPS values accordingly decrease, indicating improved perceptual fidelity. Interestingly, under higher noise (std = 5), *Non-Predictive-Net* maintains better SSIM and LPIPS than the feedback-based approach, presumably because iterative

corrections can compound the misalignment errors. In all conditions, *Baseline* remains the weakest performer, showing minimal improvement regardless of saccade count.

Figure 4B provides a complementary aggregate view, showing mean SSIM and LPIPS across all saccades at each noise level. The x-axis is std itself, while the y-axes represent SSIM (left) and LPIPS (right). It can be seen that *Predictive-Net* is the best (i.e., highest SSIM, lowest LPIPS) at low noise (std = 0), whereas *Non-Predictive-Net* outperforms *Predictive-Net* at high noise levels (e.g., std = 5). *Baseline* trails both learned models in every condition, reinforcing that simple averaging over saccades is suboptimal.

Discussion

This study investigated the impact of a feedback mechanism on image colorization and stabilization during saccadic eye movements. Human visual stability depends on perceptual continuity, not pixel-level accuracy, using object correspondence and spatial priors. Our model abstracts this principle by testing whether feedback correction of CD errors can support this stability across saccades. The feedback mechanism utilized the perceived image from the previous fixation to predict and correct errors in the internal model of eye movements (corollary discharge). This predictive correction aimed to compensate for inaccuracies in the estimated eye movement trajectory, which can arise from noise in the corollary discharge signal. Representing CD errors as 2D shifts simplifies the model while isolating feedback's role. Our results demonstrated that corrective feedback can substantially enhance image colorization and stabilization when the corollary discharge (alignment) error is relatively small. With each additional saccade, *Predictive-Net* refines its prediction, leading to higher structural similarity and lower perceptual error. However, when alignment noise becomes large, the feedback loop that would otherwise confer an advantage appears to accumulate errors, allowing the simpler *Non-Predictive-Net* approach to yield better final performance.

This finding supports current theories of sensory prediction and visual stability. Particularly,Cavanaugh et al., 2016 shows that corollary discharge is crucial for integrating retinal inputs into a stable visual scene during saccades. Similarly, Bansal et al., 2018 highlights how prediction failures under noisy signals lead to perceptual instability and degraded performance. Our results extend these insights by showing that while predictive mechanisms excel under low-noise conditions, their dependence on accurate corollary discharge can render them susceptible to error accumulation when noise increases. This aligns with cognitive theories indicating that predictive processes perform well under reliable sensory signals but may fail in conditions of high uncertainty (Hohwy, 2013).

These results highlight the delicate balance between the benefits and vulnerabilities of predictive feedback systems, suggesting that optimizing feedback mechanisms for varying levels of sensory uncertainty could pave the way for more adaptive and resilient computational models of visual perception. Though based on deep learning, our model serves as a computational hypothesis for predictive correction in vision. Future work could incorporate object-based or non-linear CD mechanisms for greater biological relevance.

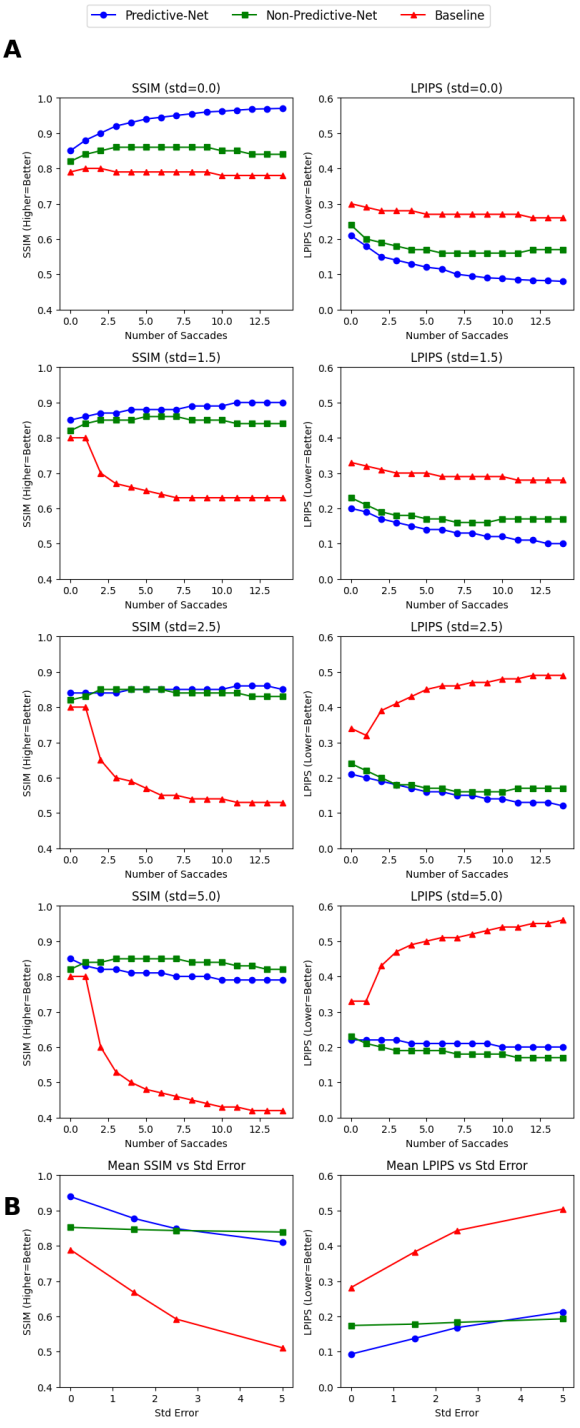

Figure 4: Performance evaluation of predictive and non-predictive Models under varying Saccades and error variability.

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

Bansal, S., Ford, J. M., & Spering, M. (2018). The function and failure of sensory predictions. *Annals of the New York Academy of Sciences*, *1426*(1), 199–220.

Cavanaugh, J., Berman, R. A., Joiner, W. M., & Wurtz, R. H. (2016). Saccadic corollary discharge underlies stable visual perception. *Journal of Neuroscience*, *36*(1), 31–42.

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

Cohen Duwek, H., & Ezra Tsur, E. (2021). Biologically plausible spiking neural networks for perceptual filling-in. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Cohen Duwek, H., Showgan, Y., & 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., Slovin, H., & Ezra Tsur, E. (2022). Computational modeling of color perception with biologically plausible spiking neural networks. *PLoS Computational Biology*, *18*(10), e1010648.

Cohen-Duwek, H., & Tsur, E. E. (2022). Biologically plausible illusionary contrast perception with spiking neural networks. *2022 IEEE International Conference on Image Processing (ICIP)*, 1586–1590.

Collewijn, H., Erkelens, C. J., & Steinman, R. M. (1988). Binocular co-ordination of human vertical saccadic eye movements. *The Journal of physiology*, *404*(1), 183–197.

Crapse, T. B., & Sommer, M. A. (2008). Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, *9*(8), 587–600.

Cronin, D. A., & Irwin, D. E. (2018). Visual working memory supports perceptual stability across saccadic eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(11), 1739.

Ford, J. M., & Mathalon, D. H. (2019). Efference copy, corollary discharge, predictive coding, and psychosis. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *4*(9), 764–767.

Hohwy, J. (2013). *The predictive mind*. OUP Oxford.

Huang, Y., & Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580–593.

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., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). Salicon: Saliency in context. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1072–1080.

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

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in cognitive sciences*, *4*(1), 6–14.

Melcher, D. (2011). Visual stability.

Perry, J. S., & Geisler, W. S. (2002). Gaze-contingent real-time simulation of arbitrary visual fields. *Human vision and electronic imaging VII*, *4662*, 57–69.

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? *International conference on machine learning*, 5389–5400.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, *28*.

Showgan, Y., Cohen Duwek, H., & Ezra Tsur, 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*.

Solomon, S. G., Lee, B. B., White, A. J., Rüttiger, L., & Martin, P. R. (2005). Chromatic organization of ganglion cell receptive fields in the peripheral retina. *Journal of Neuroscience*, *25*(18), 4527–4539.

Van De Sande, K., Gevers, T., & Snoek, C. (2009). Evaluating color descriptors for object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, *32*(9), 1582–1596.

Wandell, B. (1995). Foundations of vision.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & 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.