# A Saccading Model for Temporal Illusions

Sahil Arora*, Seung Je Jung*, Taylor Del Matto*, Joel Ye*

## Abstract

*Temporal illusions are illusions which are perceived over time as a result of our continuously changing percept of the target. To understand these illusions, we must move beyond feedforward visual networks and model vision as a temporal process. We add recurrence, sparsity, self-supervision, and a saccading mechanism to a standard convolutional autoencoder architecture to better model this temporality. We specifically focus on training our model to reproduce phenomena from three fixation illusions: the uniformity illusion, rotating snakes, and Troxler's fading. While there were challenges reproducing complex stimuli, we were able to reproduce select phenomena from these illusions.*

## 1 Introduction

While current deep convolutional neural networks appear to reflect static aspects of visual perception such as hierarchical processing, visual perception is known to be a temporal process [5, 14]. Thus to deepen our understanding of biological vision, it is important to be able to model this temporality. Current works that model temporality study how recurrence is involved in object recognition, though this tangles the study of hierarchical processing and long-range recurrent connections [14].

We instead consider the simpler case of visual "temporal" illusions. This subclass of visual illusions are illusions that arise over time or mimic some passage of time even though the image itself is completely static. These temporal illusions provide a good litmus test for any hypothesized temporal models. We thus aim to evaluate how well a simple temporal model of the visual system can recreate temporal illusions, specifically the uniformity illusion [11], rotating snakes [9], and Troxler's fading [8].

It should be noted that all of these temporal illusions are also related to fixation. The uniformity illusion and Troxler's fading both take time to onset, as they emerge only after a period of fixation, and subsequently break when the viewer's eyes saccade around the image. Rotating snakes differs from the rest in that it occurs immediately, and breaks when the viewer focuses on one part of the image. The specific illusions and their respective phenomena are described as follows.

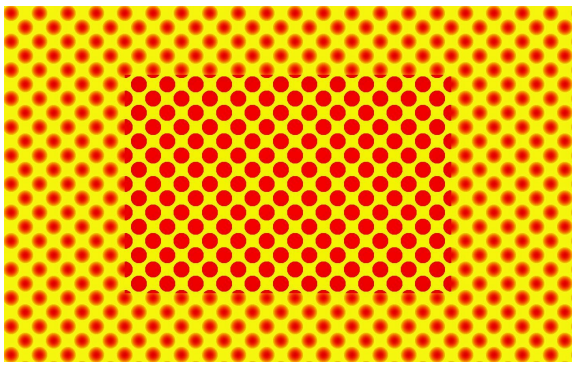

**Figure 1.** Uniform Blurriness Illusion

The uniformity illusion is a class of optical illusions in which the center of the image differs from the periphery of the image, but appears to be completely uniform when fixated on at the center [11]. This illusion is theorised to occur due to limitations of resolution and color perception at the visual periphery, which the brain fills in, creating the uniformity [11]. We will investigate how our model fills in the uniformity illusion over different saccades after given a uniform prior.

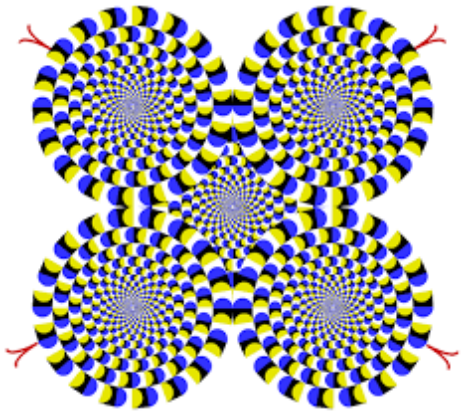

**Figure 2.** Rotating Snakes Illusion

The rotating snakes illusion is another optical illusion where circles with asymmetric luminance gradients, more simply striped concentric circles, appear to rotate [9]. The underlying mechanism that causes the illusion is unknown, but it is conjectured that the rotations may be caused by a

difference in the processing speeds of signals based on color contrast [2]. Further, it is known that the illusory strength is correlated to fixational drift [9]. To investigate the rotating snakes illusion, we will investigate how the model perceives rotating snakes when saccading in a small area, simulating fixation.



**Figure 3.** Troxler's fading Illusion

Lastly, Troxler's fading is an illusion in which fixating on the illusion causes it to fade [8]. This phenomena occurs due to a feature of neurons named neural adaptation, where a neuron constantly receiving the same stimulus will gradually decrease in responsiveness, therefore, fixating on the same location in the illusion causes it to slowly fade. To investigate Troxler's fading, we will investigate the perception of the illusion given the addition of a neural adaptation layer into the model.

## 2 Hypotheses and Approach

Since these illusions are visually "simple" stimuli, without high-level semantics, we are inspired to consider the early visual system. Specifically, we note that the eye saccades, and that visual adaptation begins in the retina [3]. Moreover, vision is recurrent [6]. Based on these straightforward biological facts, we hypothesize that they are sufficient mechanisms to reproduce the illusions in question. Specifically, we hypothesize a visual system with:

- Limited field-of-view, that is unable to view the entire image noiselessly.
- Recurrent perceptual state, that updates through time in a noisy fashion.
- Neural adaptation, such that it prefers to activate less frequently

will reproduce the Uniformity, Peripheral Drift, and Troxler's fading illusions.

To model this, we can either use a theoretically optimal (but still noisy) visual system which makes predictions about image content, or we can train a network to make

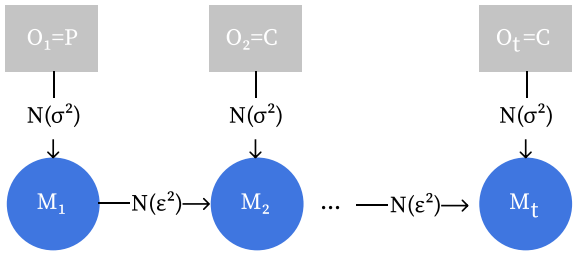


**Figure 4.** Graphical model of saccading and fixation.

predictions about image content. Since illusion stimuli are quite artificial (e.g. rotating snakes are very infrequent in nature), we also hypothesize that learning to predict illusion stimuli is sufficient *i.e.* generalization is not a necessary ingredient of the illusions. In either theoretical or simulated models, we can flexibly control model focus to mimic the fixation condition needed to induce the illusions.

## 3 Theoretical Results

As premises for our model we assume that we have noisy observations of limited fields of view of our model, and moreover that our memory state is updated noisily. While noisy observation is standard, we use a noisy memory update because maintaining perfect information through sustained activity (*i.e.* since perceptual information is not entirely learned/stored in synapses) would be inefficient. On top of this, we assume that image content at different locations are correlated (this is certainly true in illusions but also in the natural world [12]). As a final condition, we assume that our noise models are constant independent of the content being noised.

### 3.1 Optimal Estimation and Uniformity Decay

We assume we have an optimal estimator fixating on the center of an image $C$ after an initial glimpse of the periphery $P$, as shown in Fig. 4. Note that instead of a dense grid of pixels we have simplified "limited FOV" into an "image" where we can only observe one of two dependent random variables. In this model we simplify the relevant quantities to scalars, and assume we maintain a scalar memory $M_t$ representing our perceptual state.

To model a uniformity effect, we assume the memory is solely an optimal minimum mean squared error (MMSE) estimator of P, *i.e.*

$$M_t = \hat{P}_t = \mathbb{E}[P|M_{t-1} + N(\epsilon^2)]$$

. In practice, the periphery $P$ would be a random vector such as pixels in a certain part of an image, and the memory will also be a random vector which must encode other content (a percept of the whole image).

We are interesting in bounding the error of this estimator, $E(\hat{P}_t) = \mathbb{E}[(P - \hat{P}_t)^2]$. Throughout the proof, we will assume gaussian observation error, implying the MMSE es-

timator is also max-likelihood (ML). The uniformity illusion implies that our error will initially be near observation error, *i.e.* $\sigma^2$, but over time, the error will approach that of a contextual estimator $\dot{P} = \mathbb{E}[P|C]$.

While we do not prove the limit, we show error monotonically increases and is bounded from above by the contextual estimator. We show this occurs in three steps:

1. $E(\hat{P}_1) = \sigma^2$
2. $E(\hat{P}_{t+1}) > E(\hat{P}_t)$
3. $E(\hat{P}_t) < E(\dot{P})$

.

Note that the fact that the MMSE estimator given condition $C$ is the conditional mean follows directly from $\min_g E[(Y-g)^2|C]$. Further note that the expected error is equivalent to the conditional variance, again from definitions and linearity of expectation. For example, since our model has $P_1$ conditioned on the observation $O_1 = N(P, \sigma^2)$, $E(\hat{P}_1) = V[P|O_1]$. Moreover, since we assume uniform distributions over image content $P, C$ absent other cues (which is generally true for artificial illusions), Bayes' rule allows us to directly flip the probability:

$$N(P, \sigma^2) = p(O_1|P) = p(P|O_1) = N(O_1, \sigma^2)$$

. From here we see not only that $\hat{P}_1 = \mathbb{E}[P|O_1] = O_1$ but also $E(\hat{P}_1) = V[P|O_1] = \sigma^2$.

To show that error monotonically increases, we can potentially view each new observation as a Bayes' update. (This proof is missing some serious steps) Specifically, if we assume that our previous $\hat{P}_{t-1}$ provides a normal estimate of $P$, i.e. $P \sim N(\hat{P}_{t-1}, E(\hat{P}_{t-1}))$, we can view it as a Gaussian prior in our update. If we sample noise $n \sim N(0, \epsilon^2)$:

$$p(P|O_t = C) = \frac{1}{\alpha} p(C|P)p(P) \quad (1)$$

$$P \sim N(\hat{P}_{t-1} + n, E(\hat{P}_{t-1}) + \epsilon^2) \quad (2)$$

Since the memory noise is independent of our memory, the noisy memory is simply a version of our original Gaussian prior with larger variance. To get the errors, we are interested in the variance of our new distribution relative to our old distribution. Unfortunately, the relationship between the $C|P$ and the prior is unclear, so it is also difficult to say how the error updates. However, the Bayes' framework provides some intuition that our estimate at each step will continue to accrue probability at $\text{argmax}_p(p(C|P))$.

Notably this proof doesn't really follow through since our updated distribution is immediately no longer normal, but we do not know how to fix it.

The estimators $\hat{P}_t$ and $\dot{P}$ are both MMSE-optimal, hence their expected errors are the conditional variances. According to the graphical model, we are comparing:

$$\mathbb{E}_{p\sim\hat{P}}[V[P|C, \hat{P} + N(0, \sigma^2) = p]] \overset{?}{<} V[P|C]$$

By the law of total variance, the additional observation conditioning can only reduce our expected error, so we have non-strict inequality. The law of total variance also provides strict inequality so long as $\hat{P} + N(0, \sigma^2), P$ are not conditionally independent given $C$; however it is unclear whether we can prove this without assuming a simple-enough error model wherein we can determine the optimal update, even though it is clear that $\hat{P}_0$ is independent of $C$.

Overall, the theoretical result is very constrained, but provides some justification for the notion of memory decay.

### 3.2 Sensory Adaptation

In Troxler's fading it has been shown that retinal sensory neurons desensitize to unchanging stimuli, demonstrating neural adaptation [8]. They explain that the receptive fields in peripheral vision are larger than those produced during fixational eye movements. When the eyes are fixated on a single point, the amplitude of eye movements (saccades) is reduced significantly, at which point the visual stimulation is not great enough to activate these peripheral receptive fields, resulting in fading. Visual sensory neurons have also been shown to normalize to background conditions including lighting [1]. Generally, we know that the brain is information efficient, and it has evolved to be information efficient. There are also information limits on how much data can be transmitted from the vision system at any given time.

Let's say each neuron n at time t sees light value $k \in K$, where K is finite. The probability that a neuron stays in the same state $P(X_{n,t} = k|X_{n,t-1} = k)$ is the complement of the probability that the neuron changes state $P(X_{n,t} \neq k|X_{n,t-1} = k)$. If we have a model that assumes that non-transmitted states have not changed, it is obvious that transmitting only the information from states that have changed is more efficient than transmitting the information from all of the states. Pixels in natural images are known to be highly spatially correlated. In this case, we expect $P(X_{n,t} = k|X_{n,t-1} = k)$ would be higher than if the pixels were spatially independent. The higher $P(X_{n,t} = k|X_{n,t-1} = k)$ is, the greater we would expect the efficiency gain to be. Normalizing to the most recent stimuli decreases the probability that unchanging input will be transmitted in visual sensory neurons, and this increases information efficiency relative to transmitting all information.Sensory adaptation is information efficient and part of sensory adaptation is normalizing to recent stimuli.The sensory adaptation architectures explained in the Troxler's fading section are designed based on these principles.

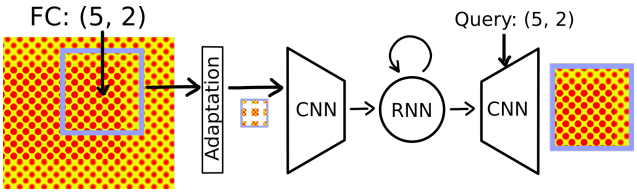


**Figure 5.** Deep network model of a saccading observer.

## 4 Methods and Simulation Results

While an optimal observer model can help explain information decay in simple settings, how can we study our illusions with more generality? In this section we explore whether a deep network model that applies the hypothetically sufficient principles can recreate the illusion. In comparison with the theoretical model, we now must fully rely on self-supervision to learn how to make conditional estimates. For simplicity we do not inject additional noise into the RNN update, *i.e*. memory update noise is only from floating point arithmetic error.

The model is illustrated in Fig. 5. While the brain implements vision differently than a deep network model and even though better deep network approximations can be made, we believe the model captures critical components for our temporal illusions. The misalignment may act as additional compelling evidence that the components of the model are sufficient ingredients rather than other details of biological vision.

Specifically, our model receives a limited and noisy field of view of the image. Since human field of view encompasses every standard illusion, one approach to modeling could be to provide the whole image with a noise profile that shifts depending on the observed focus. However, in order to make learning less noisy, we instead use an even sharper peripheral vision falloff and omit information from outside the smaller field of view. This input is optionally passed through an adaptation layer in which each pixel is "sensed" by a size 1 recurrent cell. This layer can, for example, gate unchanging, redundant visual stimuli to conserve energy. After adaptation, the view is downsampled through a convolutional network (CNN) and fed to a recurrent network (RNN) along with the coordinates of the focus (*i.e*. a form of proprioception). The RNN's state represents the model's percept of the image, and a subsequent upsampling CNN can query it for partial views of the image.

The model is trained by self-supervised predictive coding, *i.e*. the model predicts what it will see during the next timestep and learns from the prediction error with respect to the input to the adaptation layer. In most experiments we use pixel-wise mean-squared error as our objective. During training, the model saccade locations are determined either randomly or according to a random walk. During evaluation, we can simply specify saccade coordinates, e.g. allowing it to first randomly saccade an image and then fixate by sampling points near the image center.

**Additional helpful tricks.** We found it helpful to mix different saccades over a single image in a batch, instead of batching over different images. In general, a small batch size is insufficient for learning. Further, we found that expressing model proprioception with fourier features [13] to be helpful in sharpening image output, even though our illusions are relatively low-frequency signals. Negative results are reported Sec. 4.4.

Our model was prepared using the `pytorch-lightning` framework, using a gated recurrent unit (GRU) with hidden layer size of 512 for the memory module and a 4-layer CNN for both downsampling and upsampling the image before and after memory encoding respectively. All images are converted to grayscale, zero-centered, and resized to 64x64 before being input to the model, and from there a 32x32 window is provided at each timestep.

### 4.1 Uniformity Illusion

**Goal** In emulating the visual response to this illusion, it was important for two key phenomena to emerge. First, the uniformity illusion occurred only when fixating. When we saccade we get information from the periphery that breaks our perception of uniformity, and our model should do the same. Second, that upon fixation, the image appears to become more uniform over time. As we stare at the center pattern, it takes time for the image to appear uniform, so our model should do the same. This would align not just empirically but also with the optimal estimator model introduced earlier, which proves uniformity is a temporal process.

**Modeling** We emulate this phenomena, specifically the blur illusion as described in the introduction, with our saccading model by attempting to reconstruct different instances of the illusion. We initially perform self-supervised training with a generated dataset of uniform patterns, consisting of 50000 64x64 grayscale images with different tilings of unique shapes and lines. This encodes a prior towards uniformity, as the model learns to reconstruct uniform patterns as it saccades over parts of the image. Below is an example from the dataset of the illusion, and its uniform and blurred counterparts.

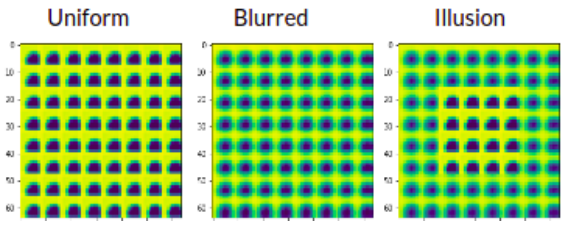


**Figure 6.** Uniform image is for self-supervised training and the illusion image is for inference,

The model is not specifically designed to reconstruct the

whole image. It is fed only the field of view from the saccade location, and from that input produces a predictive patch of what the image will look like at the next saccade. So to do a full reconstruction, we perform an inference for a patch at 9 different locations in the image. We then take these predicted patches and stitch these together at their proper locations to construct the full predicted image.

**Metrics** To quantify the effectiveness of our reconstruction, we use the structural similarity index measure (SSIM) which is commonly used to measure how similar two images are to each other. It has been applied to the domains of lossy image compression and image restoration and found to be a more effective measure for image comparison than mean squared error [15]. Given images $x$ and $y$,

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where $\mu_x$ is the mean of the image, $\sigma_x$ is the variance, $\sigma_{xy}$ is the covariance between the two images and $c_1$ and $c_2$ are small constants to stabilize the denominator. These constants are $(0.01\dot{L})^2$ and $(0.03\dot{L})^2$ respectively, where $L$ is the dynamic range of the pixel values as expressed by $2^n - 1$, $n$ being the number of bits per pixel. This formula is derived from the product of three comparison functions that measure luminance, contrast, and structure between images, and the value can be from -1 to 1.

**Findings** We found that predicted images when the model fixated were more similar to the uniform image than they were to the original illusion. This was expected behavior, as it is consistently fed precepts from the center pattern. In saccading mode, on the other hand, it was expected that the model reproduced the image as faithfully as possible, as it is getting information from the periphery and the center pattern. This was also the case, as the output of the model was more similar to original input than either the uniform pattern or its blurred counterpart. An example of these phenomena are shown below.

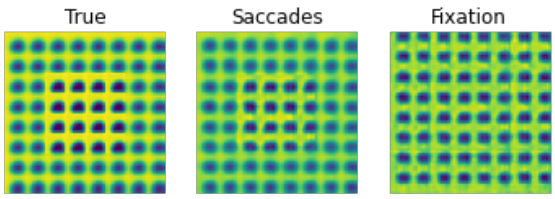


**Figure 7.** Illusion input (left), Reconstruction from sacadding model inference (center), and Reconstruction from fixating model inference (right)

From these final reconstructions we can see that the uniformity pattern emerges under fixation. While it does appear to be somewhat blurry compared to the sharper center image in the original, these artifacts can be attributed to a lack of specific examples to more robustly learn this pattern. The model had to learn to recurrently autoencode from a variety of different patterns, and with more computational resources to train on a larger dataset for more epochs, it is feasible to get a higher resolution reconstruction. Some other example illusions and their fixated reconstructions are shown below.

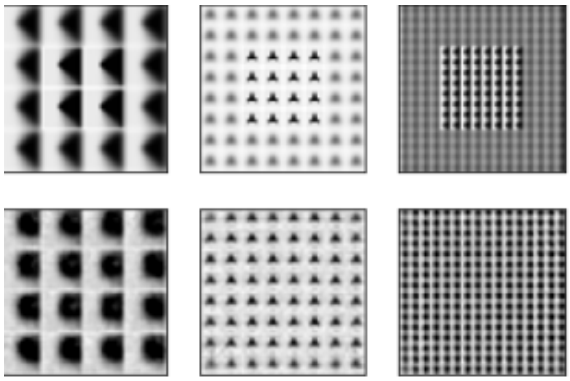

**Figure 8.** Illusion inputs (top) and Reconstruction from fixating model inference (bottom)

To show that the uniformity result from Figure 7. arose over time, we plotted the SSIM of reconstructed patches to their respective patches in the uniform image over different time steps. The plot below demonstrates that while fixating, the model on average gets better at reconstructing the uniform pattern over time. This is analogous to the filling in of the uniform pattern to the periphery when we fixate on the center of a uniform illusion.

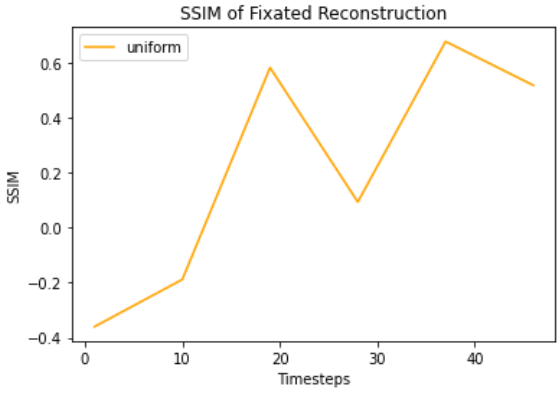


**Figure 9.** SSIM of fixated reconstructions to uniform pattern

Global SSIMs were also calculated comparing saccading and fixation modes across the dataset to test the reconstruction's similarity to the illusion and its blurred and uniform pattern components. This was done by calculating the SSIM for each reconstructed image to the aforementioned patterns. The mean taken across the images for these different comparisons, and are shown in the table below.

| | Original Illusion | Central Uniform Pattern | Peripheral Blurred Pattern |
|---|---|---|---|
| Saccading Mode | **0.457** | 0.412 | 0.448 |
| Fixation Mode | 0.397 | **0.442** | 0.421 |

**Table 1. Global SSIM of images to model output** Relative values support hypotheses and qualitative findings. Blurred similarities are noticeably higher due to artifacts from reconstructed patches.

The results demonstrate that there is a marked relative difference between fixation and saccading for model inference. Saccading models on average more effectively reconstructed the original illusion, while fixation resulted in better similarity to the uniform pattern as described previously. The SSIM values are rather low, however, suggesting that the reconstructions are not similar to the input images at all. This can be attributed to an inability to properly reconstruct some of the images in the dataset. We noticed that on images with more complex shapes and intricate details, the model would perform very poorly on reconstruction. An example is shown below of a spiral pattern not even closely being reconstructed. These sets of images cause the mean SSIM to be weighted down accordingly.

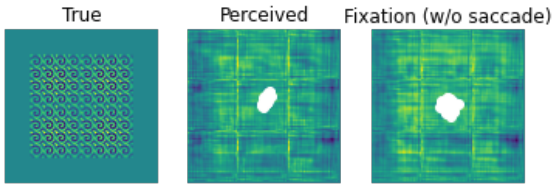


**Figure 10.** Failed reconstruction from a more complex example.

This poor performance can be due to many factors. There may not have been enough examples in the data to properly learn a generalizable function that can handle all the different shapes. More training may also have been required to capture intricate details at a higher resolution. It is also plausible that neither the saccading mechanism nor the grid-based reconstruction can adequately capture these details as readily as some simpler patterns. These failings are elaborated on in later sections.

### 4.2 Peripheral Drift/Rotating Snakes

The peripheral drift illusion, specifically the rotating snakes illusion, is a class of illusions constructed out of one or more circles each consisting of concentric striped circles. Upon looking at the illusion, the circles will appear to rotate by themselves. We start with a simplified version of the rotating snakes illusion, consisting of a variably shaded circle, as seen below.

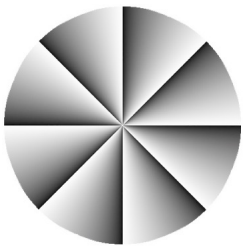

**Figure 11.** Simplified Rotating Snakes Illusion

The first step towards recreating the illusion through the aforementioned saccading model would be to verify that the model can, to a reasonable extent, reconstruct the whole input image given the entirety of the image, essentially acting as an auto encoder. To verify this, we first trained the model on the above simple rotating snakes illusion, and were able to replicate the illusion reasonably well even with saccading and a restricted field of view as seen below.

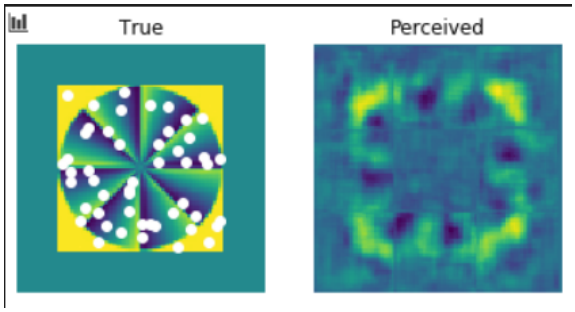


**Figure 12.** Auto-encoding with the rotating snakes illusion

The next step would be to attempt to generalise the auto-encoder to produce correct results for small perturbations of the image such as shifts or rotations. However, after several iterations and tweaks, we were ultimately unable to create a model that could auto-encode rotations of the rotating snakes illusion.

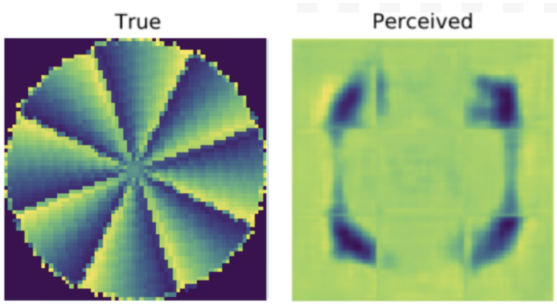


**Figure 13.** Failure to Generalize Onto Rotations

We therefore found that due to the complex nature of the rotating snakes illusion, we would need much larger training data sets, as well as much longer compute time and resources to reproduce the rotating snakes.

### 4.3 Troxler's fading

It has been established that the Troxler's fading illusion is due to neural adaptation [8]. During fixation, the amplitude of eye movements (saccades) is reduced, at which point the stimulation is not great enough to activate the peripheral receptive fields, resulting in fading. We planned to start by evaluating the model on the most basic Troxler illusion shown in Figure 3. We planned to evaluate the illusion quantitatively with L2 distance. For success, during saccades, the model output would be closer to the input than the input background. Upon fixation, the model would be closer to the input background than the input itself. The input refers to the pixel values of the illusion and the input background refers to an image consisting of solely the illusion background without the ring.

Unfortunately, for most trials the model was not able to reproduce basic stimuli. The model especially struggled with ring shapes and the output was not qualitatively good enough to measure the image. The failure to reproduce input stimuli was not a product of the adaptation layer. The model did not learn the input stimuli without the sensory adaptation layer either. At the last minute, our team developed modifications to the architecture and training procedure that did actually reproduce output sufficient to test, as shown below. A brief qualitative evaluation was completed, indicating that the illusion was not reproduced. However, there was no time to tune the sparsity loss parameters so this is not a fair assessment of the sensory adaptation architectures. The model was primarily trained on the third sensory adaptation architecture: RNN with Sparsity loss.

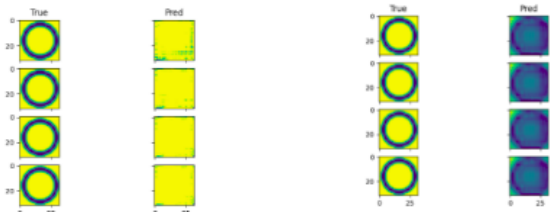


**Figure 14.** Preliminary Reconstruction
**Left**: most models did not yield any results for the ring.
**Right**: after making some architectural modifications, the model reproduced a plausible output for the ring.

Each sensory adaptation architecture is a function of each individual pixel and respective previous timesteps. We want to compute a simple function of each input individually so that it is not computing a complex function over the entire image. Sensory adaptation architecture 3, had 151 learnable weights, the most out of all the adaptation layers. The entire model has approximately 1.1 million learnable weights. In short, each adaptation layer contains a very small fraction of the parameters in the entire model. They are small and lightweight.

Based on the concepts established in the sensory adaptation section: part of sensory adaptation is normalization to recent input, and that sensory adaptation is information efficient, we have designed three architectures for the sensory adaptation layer.

The first architecture, hardcoded, computed a feature of the absolute difference of the input with the mean of the previous j timesteps. $X'_{n,t} = |X_{n,t} - \frac{1}{j}\sum_{i=t-j}^{t-1} X_{n,i}|$ That feature was input into a threshold function, where if greater than a threshold, a one was output, otherwise a 0. This binary output was multiplied by the original image to form an out gate (the hadamard product of the binary thresholded output and the original image data). If the threshold was exceeded, the input would be transmitted to the model. We considered this to be the simplest form of normalization, and we hoped that this model would serve as a sanity check for our intuitions, that this kind of input normalization was in fact responsible for the sensory adaptation resulting in Troxler fading. This architecture was briefly tested.

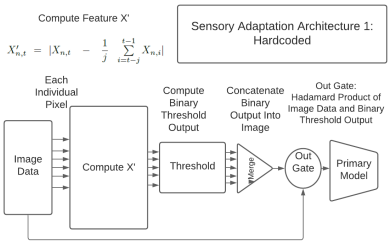


**Figure 15.** Hardcoded Layer

The second architecture has the same structure as the first model, except that the out gate was learned with a linear function of each neuron fed into a sigmoid activation. A sparsity loss was taken over the outputs of the linear layer or the activation depending on which loss function was used. This architecture was also briefly tested. Upon training on the images, very little signal reached the linear layer in the out gate, leading to concerns that the normalization input feature was insufficient. Perhaps this kind of input normalization is too simple to reproduce sensory adaptation.

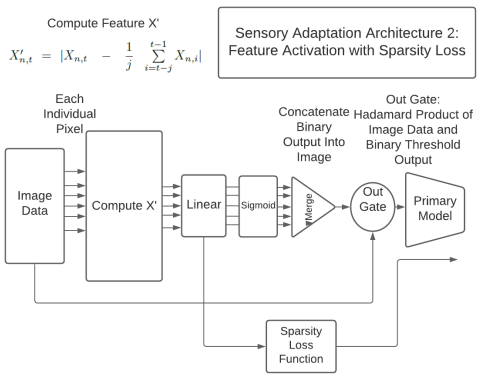


**Figure 16.** Feature Activation with Sparsity Loss

The third architecture feeds each individual pixel into it's own RNN, which also takes as input the location of the pixel in the image. The RNN outputs are fed into a linear layer followed by a ReLU activation. Again, a sparsity loss was taken over the outputs of the linear layer or the activation depending on which loss function was used. The model was primarily trained using this adaptation architecture, and was used for the best reconstruction of the input stimuli.

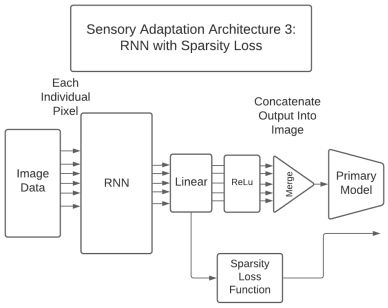


**Figure 17.** RNN with Sparsity Loss

The sparsity loss measures information efficiency. By restricting the functions to a simple input of the timeseries of each individual pixel, and by encouraging information efficiency with the sparsity loss, we hope the architecture will learn something similar to sensory adaptation which is also information efficient.

Two sparsity loss functions were used, L1, and we developed $-e^{-(x+6)^2} + .01x$, a modified version of $-e^{-(x)^2}$ from Olshausen et Al. 1996.[10] L1 encourages activations to be near zero, but not actually zero, and did not actually lead to ReLu activations turning off in tests of the adaptation architectures. The concern with low valued activations is that the model could simply learn high weights to compensate and these inputs would not actually be turned off. $-e^{-(x+6)^2} + .01x$ is used on the linear output instead of the activation and encourages outputs to be negative resulting in 0 outputs of the ReLU activation. The sparsity loss was weighted as a function of distance from the center of the image to be consistent with prior findings [8].

The model was initially trained on 40,000 randomly generated images with various random shapes including circles, rings, rectangles and lines added to images with a white background. The percentages of various shapes were modified in an attempt to improve the performance without success.

### 4.4 Negative Results

Training our networks turned out to be surprisingly difficult. The following approaches did not qualitatively improve network performance:

- SSIM/VGG/Adversarial perceptual loss: since pixel-wise mean squared error tends to smooth heavily penalizes high-frequency output, it may be preferable to use a higher level perceptual loss, as has been shown to help super-resolution [4, 7]. SSIM/CIFAR-pretrained VGG based losses failed to create generalized models, while adversarially training was unstable overall, as the discriminator quickly dominated (we did not have the expertise to tune).

- Expressing proprioception with polar coordinates or as deltas: modifying proprioception to polar coordinates did worse than fourier features, even on circular stimuli. Providing propioception deltas instead of absolute coordinates made no difference.

- More localized predictive coding as in Rao *et al.* [12], though more plausible, did worse an end-to-end back-propagation of prediction error.

## 5 Discussion

We set out to study how to understand why we can perceive static images as a changing percept. It quickly became evident that the root cause of this is because our vision is dynamic even if the observed target is dynamic: our percepts are maintained by recurrent and noisy activity and updated with adaptive sensory neurons. Moreover, our eyes constantly saccade and will make micro-oscillations on the target even when we are consciously fixating. We spend the work showing that several temporal illusions that occur on ostensibly static images are a result of these system properties.

To this end, we show an optimal estimator model must perceive a changing percept of a location it is not viewing over time, and empirically validate that a deep network trained in an entirely self-supervised manner does recreates the uniformity illusion. However, we struggle to reproduce the peripheral drift illusion due to a lack of data that allows us to generalize to the stimuli; it is yet unclear whether our mechanisms are sufficient for it but scaled-up training

should enable such a test. We fail to effectively recreate Troxler's fading due to a lack of time, and so similarly we cannot confirm whether sensory adaptation emerges out of efficient sparsity objectives, but such a result should fall quickly out of further experimentation. In general, it seems that visual illusions are difficult to study with deep neural networks trained solely on the illusion stimuli; *i.e*. to learn robust and good estimation parameters, the model appears to need more diverse stimuli (*e.g*. requiring pretraining on ImageNet) even though the illusion stimuli are themselves artificial.

It is unclear why this kind of dynamically updating visual system is preferred e.g. over a powerful, feedforward visual architecture that dominates computer vision today. Thus a pressing question would be to see what benefits arise from these design choices.

## References


[1] M. Carandini and D. J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1): 51–62, 2012.

[2] B. R. Conway, A. Kitaoka, A. Yazdanbakhsh, C. C. Pack, and M. S. Livingstone. Neural basis for a powerful static motion illusion. *Journal of Neuroscience*, 25(23):5651–5656, 2005.

[3] J. B. Demb. Functional circuitry of visual adaptation in the retina. *The Journal of physiology*, 586(18):4377–4384, 2008.

[4] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[5] K. Kar, J. Kubilius, K. Schmidt, E. B. Issa, and J. J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, 22(6):974–983, 2019.

[6] J. Kubilius, M. Schrimpf, A. Nayebi, D. Bear, D. L. Yamins, and J. J. DiCarlo. Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, page 408385, 2018.

[7] A. Lucas, S. Lopez-Tapia, R. Molina, and A. K. Katsaggelos. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Transactions on Image Processing*, 28(7):3312–3327, Jul 2019. ISSN 1941-0042. doi: 10.1109/tip.2019.2895768. URL http://dx.doi.org/10.1109/TIP.2019.2895768.

[8] S. Martinez-Conde, S. L. Macknik, and D. H. Hubel. The role of fixational eye movements in visual perception. *Nature reviews neuroscience*, 5(3):229–240, 2004.

[9] I. Murakami, A. Kitaoka, and H. Ashida. A positive correlation between fixation instability and the strength of illusory motion in a static display. *Vision research*, 46(15): 2421–2431, 2006.

[10] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[11] M. Otten, Y. Pinto, C. L. Paffen, A. K. Seth, and R. Kanai. The uniformity illusion: Central stimuli can determine peripheral perception. *Psychological Science*, 28(1):56–68, 2017.

[12] R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[13] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains, 2020.

[14] R. S. van Bergen and N. Kriegeskorte. Going in circles is the way forward: the role of recurrence in visual inference. *Current Opinion in Neurobiology*, 65:176–193, 2020.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.