Modelling fixational eye movements to achieve superresolution in deep neural networks

A Thesis

submitted to

Indian Institute of Science Education and Research Pune in partial fulfilment of the requirements for the BS-MS Dual Degree Programme

by

Kirubeswaran O.R

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

April 2023

Supervisor: Dr. Tim Kietzmann

Certificate

This is to certify that this dissertation "**Modelling fixational eye movements to achieve superresolution in deep neural networks"** towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents the study/work carried out by Kirubeswaran O.R at University of Osnabrück under the supervision of Dr. Tim Kietzmann, Full Professor for neuro-inspired Machine Learning, University of Osnabrück, during the academic year 2022-2023.

Dr. Tim Kietzmann

Committee:

Dr. Tim Kietzmann

Dr. Collins Assisi

This thesis is dedicated to my past and older self.

Declaration

I hereby declare that the matter embodied in the report entitled **Modelling fixational eye movements to achieve superresolution in deep neural networks** are the results of the work carried out by me at the Department of neuro-inspired Machine Learning, University of Osnabrück, under the supervision of Dr. Tim Kietzmann and the same has not been submitted elsewhere for any other degree

Kirubes.

Kirubeswaran O.R

Date: 1 April 2023

Acknowledgments

I thank Dr. Tim Kietzmann for their supervision and for providing the opportunity to collaborate on this project. I thank Dr. Adrien Doerig for his constant mentorship, be it with writing code or providing useful suggestions on every roadblock during the course of the project. I would like to acknowledge Mathis Pink, who helped me get familiar with the technical resources necessary to get started with my project and provided support and guidance throughout the duration of my project.

I would like to thank my friends Arun, Harini, Pranav, Adish, Pavitra, and Madesh, who provided me with emotional support, listened to all my unhinged ramblings, and validated my thoughts. I would also like to thank all the people I met during the No Garland Neuroscience conference who became instant besties. Their passion and dedication to neuroscience inspired and provided me with the motivation to complete this project. Additionally, I'm grateful to all the interesting people I have come in brief but close contact with for the past year. Their diverse backgrounds, experiences, and perspectives have enriched my life and broadened my horizons. The brief conversations and interactions we had were invaluable in shaping my worldview and inspiring me to pursue new passions.

Finally, I'd like to thank my parents, Ramesh babu, and Vijayalakshmi, for their unwavering support and encouragement in my academic pursuits and my sister, Visali, for ordering me food every now and then.

Contributions

Contributor name	Contributor role
Tim Kietzmann, Adrien Doerig	Conceptualization Ideas
Kirubeswaran, Tim Kietzmann, Adrien Doerig	Methodology
Kirubeswaran, Tim Kietzmann, Adrien Doerig	Software
Kirubeswaran, Adrien Doerig	Validation
Kirubeswaran	Formal analysis
Kirubeswaran	Investigation
Kirubeswaran, Tim Kietzmann, Adrien Doerig	Resources
Kirubeswaran, Adrien Doerig	Data Curation
Kirubeswaran	Writing - original draft preparation
Kirubeswaran, Adrien Doerig	Writing - review and editing
Kirubeswaran, Tim Kietzmann, Adrien Doerig	Visualization
Tim Kietzmann, Adrien Doerig	Supervision
Tim Kietzmann, Adrien Doerig	Project administration
Tim Kietzmann, Adrien Doerig	Funding acquisition

This contributor syntax is based on the Journal of Cell Science CRediT Taxonomy[1].

[1] https://journals.biologists.com/jcs/pages/author-contributions

Abstract

Our eyes are constantly moving. Every second, we make three large saccadic movements, and in between these saccades, small fixational eye movements continuously occur. These eye movements are not random and serve crucial computational roles by focussing on relevant parts of the environment and allowing information to be integrated between eye movements. This active sampling of information is a hallmark of human visual processing but is currently difficult to model. Indeed, Deep Neural Networks (DNNs), the current state of the art for modelling the visual system, commonly lack eye movements and process static images in a single feedforward sweep. Understanding how to model eye movements and how to integrate this information over time is an important avenue of research.

The following thesis focuses on modelling the small fixational eye movements that continuously occur between saccades. It has been shown that these fixational eye movements allow the visual system to reach superresolution to detect features of higher spatial frequency than what would be possible under static fixation. To model this process, we used a recurrent DNN combining supervised learning and deep reinforcement learning that can learn where to look in images. Reproducing the experiments conducted on humans, we trained the network to classify down-sampled high spatial frequency psychophysical stimuli that cannot be discriminated from the static image. We show that the network is able to learn useful fixational eye movements to achieve human-like superresolution on these stimuli and test to what extent this model can explain experimental data about human fixational eye movements. Finally, we show that this method can be applied to reach superresolution on naturalistic images.

Table of contents

Page

Acknowledgements……..……………………………………………………………5

Abstract……………………………………………………………………………7

List of figures……………………………………………………………..……9

Introduction………………………………………………………………….**11**

Primer to fixational eye movements………………………………………12

Functions of Fixational Eye Movements…………………………..……14

Primer to Artificial Neural Networks………………………………………..18

Need for the study…………………………………………………………...19

Methods……………………………………………………………**...**………**.21**

Dataset preparation………………………………………………………….21

Network…………………………………………………………………………….22

Experiment……………………………………………………………………….25

Baseline model…………………………………………………………………….26

Consequences of selective retinal stabilization…………………………..27

Consequences of partial retinal stabilization……………………………...28

Applications of FEMs on naturalistic images……………………………..29

Results…………………………………………………………………………**.31**

Baseline model…………………………………………………………………….31

Consequences of selective retinal stabilization……………………….…32

Consequences of partial retinal stabilization……………………………...35

Applications of FEMs on naturalistic images……………………………..37

Discussion……………………………………………………………**...**…**..39**

References………………………………………………………………….…**42**

List of Figures

Page

Figure 1: Visual acuity as a function of retinal eccentricity……………………………………………………...……12

Figure 2: Consequences of selective retinal stabilization……………….15

Figure 3: Consequences of partial retinal stabilization…………………..16

Figure 4: Basic Architectures of Artificial Neural Networks…………..….18

Figure 5: Example images from training dataset…………………………22

Figure 6: Fixational eye movement network architecture……...……..23

Figure 7: Baseline model architecture……………………………………..25

Figure 8: Examples of average pooled images from the training dataset…………………………………………………………….26

Figure 9: Experimental setup for partial retinal stabilization Experiment……………………………………………………………….28

Figure 10: Examples of average pooled images from the GTSRB training dataset…………………………………………………..29

Figure 11: Baseline model performance on different versions of average pooled gratings……………………………………………….31

Figure 12: Consequences of selective retinal stabilization on FEMNet………………………………………………...…………32

Figure 13: Example of fixation pattern executed by a FEMNet instance……………..…………………………………………..…..34

Figure 14: Histogram of angles between successive Fixations made by a FEMNet instance………..………………………35

Figure 15: Consequences of partial retinal stabilization on FEMNet………………..…………………………………………………...36

Figure 16: Consequences of selective retinal stabilization
on FEMNet on the GTSRB dataset………..………………………………37

Introduction

One of the fundamental aspects of our everyday experience is our natural and the effortless ability to see the world around us. But the perceptual phenomena seem incredibly complex when demanded with a scientific explanation, considering the fact that the rich perceptual experience arises from a two-dimensional pattern of light stimulating the photoreceptors and other neurons at the back of our eyes. Kurt Koffka, a prominent German psychologist, succinctly captured the essence of this problem by asking the fundamental question, "Why do things look as they do?". Building upon Koffka's question, this thesis aims to provide a narrow inquiry into how the visual system processes and interprets visual information, with a particular focus on the role of specific types of eye movements during visual perception.

Where we choose to look is a selective process that is influenced by both bottom-up and top-down factors. The salience of visual stimuli, such as their contrast and motion, can capture our attention involuntarily through bottom-up processes, while top-down processes, such as our goals or the relevance of the task, can guide our attention to focus on relevant aspects of a scene (Katsuki and Constantinidis, 2014). These factors play an important role in perception by guiding our attention toward important or relevant parts of the scene because the visual system has anatomical constraints that limit our ability to see the entire scene at once. Instead, we selectively attend to parts of the scene where we fixate our gaze. In many primates, including humans, a small region of the retina called the fovea is specialized for high-acuity vision. This region provides the highest visual resolution and is critical for reading, recognizing faces, and other tasks that require detailed visual information.

The non-uniform acuity can be attributed to the unequal distribution of the photoreceptor cells in the human retina. The Foveola, a small region of about 0.3mm in diameter at the centre of the fovea has the highest acuity, consisting exclusively of colour-sensitive cone photoreceptors. The receptor density and hence the spatial as well as the chromatic resolution drops rapidly towards the visual periphery, away from the foveal region, where the image appears progressively blurry (Figure 1). Thus, the brain constantly sends commands to the eye muscles to make rapid eye movements called saccades that help us orient the gaze towards the areas of interest such that it falls sequentially on the fovea.

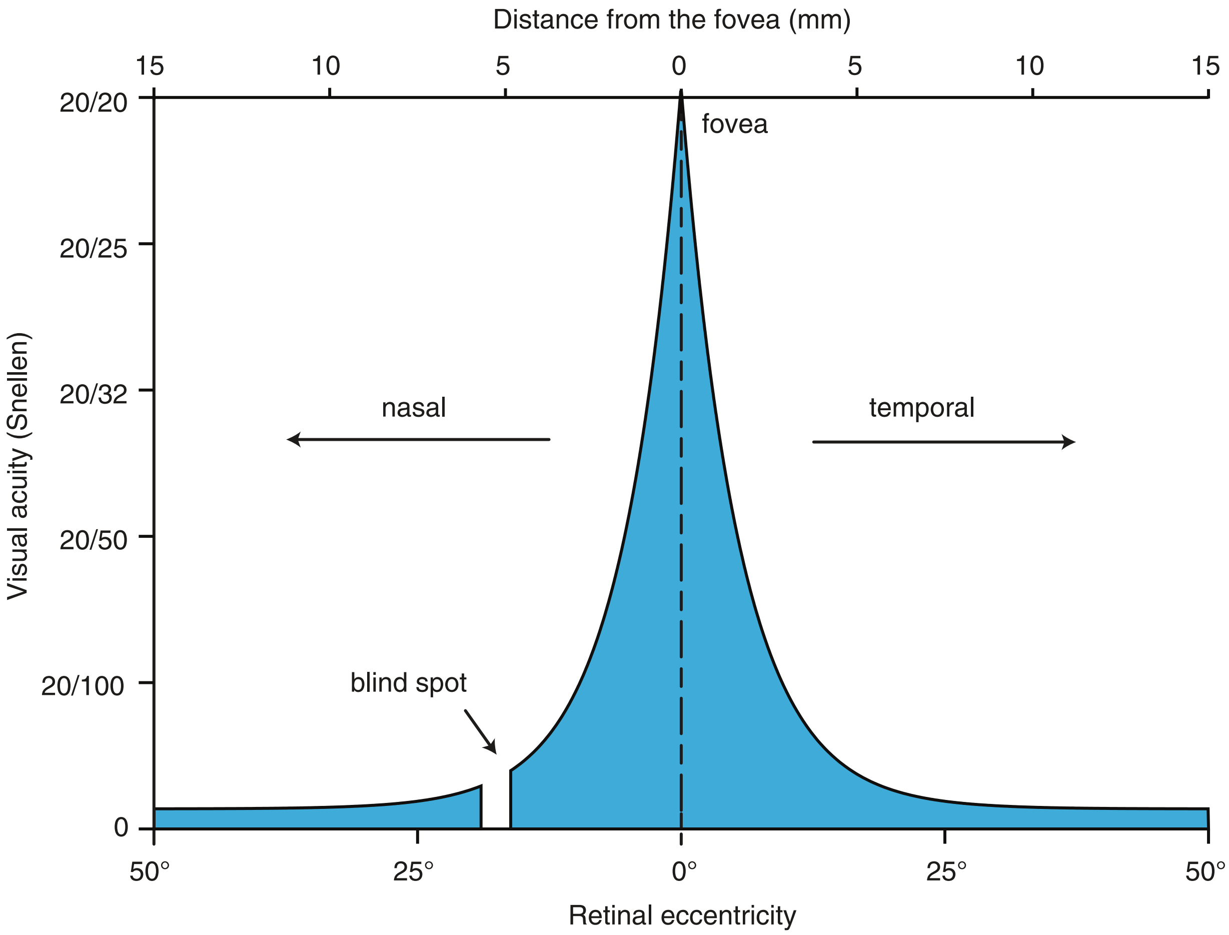


Figure 1. *Visual acuity as a function of degrees of retinal eccentricity, the degree to which an object is located away from the centre of the visual field/ fovea.* (Lambertus *et al.*, 2017)

Humans make several saccades per second. However, it is only during the periods of fixation in which we gain information about the scene in front of us. The main purpose of saccades is to bring the relevant visual information present in the periphery within the fovea. Despite what the name suggests, our eyes are never completely at rest during fixation; Tiny involuntary eye movements such as microsaccades (a much smaller saccade in terms of amplitude that keeps the attended scene within the foveola), drifts (slow continuous motion of the eye during the intersaccadic interval), and tremors (tiny high-frequency motion) still occur, together referred to as the Fixational Eye Movements (FEMs) (Ratliff and Riggs, 1950; Otero-Millan *et al.*, 2014; Rucci and Victor, 2015). Humans and other species constantly make these sequences of saccades and FEMs (Martinez-Conde and Macknik, 2008) in order to scan the visual scenes present in front of them. Out of the three, microsaccades have received more research attention, primarily because they are relatively larger and faster in magnitude, making them more detectable and distinguishable using non-invasive video trackers.

After the proper characterization and establishment of FEMs in the 1950s, a number of researchers utilized optical techniques, such as retinal stabilization, which counteracts any motion in the retinal image caused by eye movements, to effectively eliminate FEMs (especially microsaccades) and gain insight into their functional role in visual perception. One of the important results from these studies was that lack of FEMs causes the vision to fade away within a matter of seconds in a laboratory setting (Ratliff and Riggs, 1950; Ditchburn and Ginsborg, 1952) and when the subjects were allowed to move their eyes without the retinal stabilization condition, vision returned back to normal. These studies showed that FEMs counteract neural adaptation, a phenomenon where the neuronal activity decays in response to repeated to prolonged stimulation, by constantly shifting the luminance information of the visual input on the retina, thus preventing visual fading.

After a brief period of quiescence during the 1950-80s (Kowler and Steinman, 1979), the investigation of FEMs has recently gained renewed momentum and widespread recognition in the field of vision research. This resurgence is partly due to advancements in highly precise non-invasive, and user-friendly eye-tracking technology, combined with the use of mathematical and computational modelling techniques has provided researchers with a powerful tool for generating new hypotheses, as well as re-evaluating older ones. Recent work on microsaccades and drifts has also shown its impact on several perceptual and cognitive functions, including its role in visual acuity (Rucci *et al.*, 2007), exploration of small spatial areas (Rolfs, 2009), engagement in attentional and cognitive processes (Martinez-Conde and Macknik, 2007; Otero-Millan *et al.*, 2008; Martinez-Conde *et al.*, 2013) and resolving perceptually ambiguous stimuli (Troncoso *et al.*, 2008; Rolfs, 2009; Otero-Millan *et al.*, 2012). Thus, further in-depth investigation into microsaccades (and other FEMs) could lead us to a better understanding of the fundamental mechanisms that govern visual perception, both in normal and pathological vision

(Martinez-Conde, 2006). In addition to these functions, FEMs also play a vital role in reducing redundancy and extracting features (Kuang *et al.*, 2012), which altogether establishes the fact that FEMs are not merely refreshing the input to retinal receptors but instead are an essential stage of information processing.

In this thesis, we focus on modelling two proposed functions of FEMs, namely, their ability to enhance the perception of fine spatial details (Rucci *et al.*, 2007) and encode information spatiotemporally (Rucci and Victor, 2015) to improve spatial resolution. Humans have the remarkable ability to distinguish between objects that are separated by just a few seconds of an arc, even though the resolution of our retinal neurons is typically limited to about 1' (one arc minute) due to the receptor density and shape of our retina (Carney and Klein, 1997). This ability to perceive details beyond the limit of the resolution set by the retina is known as hyperacuity (Geisler, 1984), which allows us to achieve superresolution and perceive visual stimuli at a higher resolution than what would be possible under static fixation in the fovea.

Various studies have proposed a potential role for FEMs in explaining this phenomenon of superresolution (Hennig and Wörgötter, 2003; Rucci *et al.*, 2007). These studies suggest that the visual system may leverage the continuously changing temporal input generated by FEMs, even when observing a static image, to encode spatial information in a spatiotemporal manner (Kuang *et al.*, 2012; Rucci and Victor, 2015). This is a departure from conventional theories of neuroscience, which propose that information is encoded only in a spatial format.

(Rucci *et al.*, 2007)'s research was instrumental in revealing how FEMs contribute to enhancing spatial detail. Classical experiments designed to investigate the role of fixational eye movements in enhancing spatial detail using stabilization techniques faced technological constraints. The inability to selectively stabilize FEMs in their natural context necessitated experiments conducted in suboptimal conditions that resulted in the subjects being required to maintain fixation for extended periods during the stabilized condition, leading to induced visual fading.

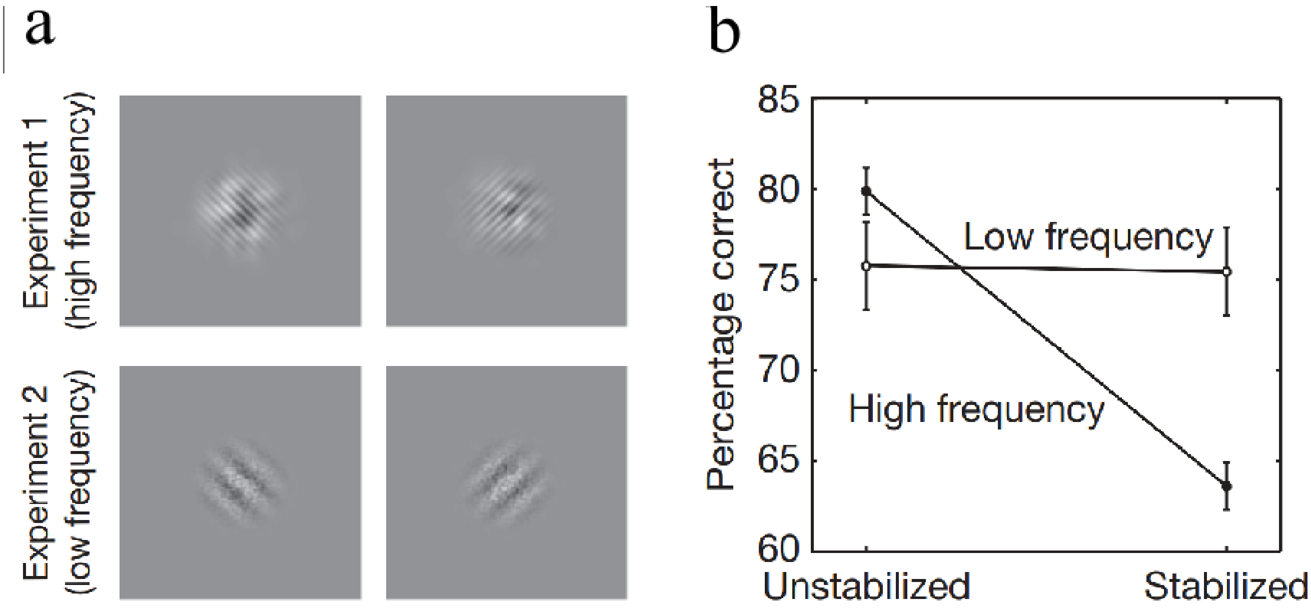


Figure 2. *Consequences of selective retinal stabilization* (Rucci *et al.*, 2007). *(a) shows the examples of stimuli used in the experiments, gratings with either high (experiment 1) or low (experiment 2) spatial frequency masked with noise fields with frequency opposite to the spatial frequency of the gratings. (b) Plot shows the performance of subjects in a forced choice task where they have to report the orientation of the gratings in a free viewing condition vs. retinal stabilization condition where the FEMs were selectively impaired using a technique that processes eye-movement signals in real-time. Orientation prediction performance on high-frequency stimuli is highly impaired in the case of retinal stabilization, while it stays the same for low-frequency stimuli.*

In a seminal study showing the role of FEMs for superresolution, Rucci et al. investigated the impact of FEMs on a classification task where the participants had to report the orientation of the visual stimuli that consisted of tilted gratings in ±45 degrees from the vertical axis and their orientation prediction performance was measured in two conditions: with and without retinal stabilization. They were able to employ a modern retinal stabilization method which effectively selectively eliminates the fixational eye movements performed by the subject by moving the stimulus real-time, thus allowing them to study the influence of FEMs in a free viewing condition (Figure 2). Two kinds of stimuli were presented, one with high and the other with low spatial frequency gratings. Both of them were masked with noises inversely proportional to the grating frequencies in order to simulate the power spectrum of natural images. Unlike the inconclusive results of the classical experiments, this study shows that the presence of FEMs helps participants perform significantly better

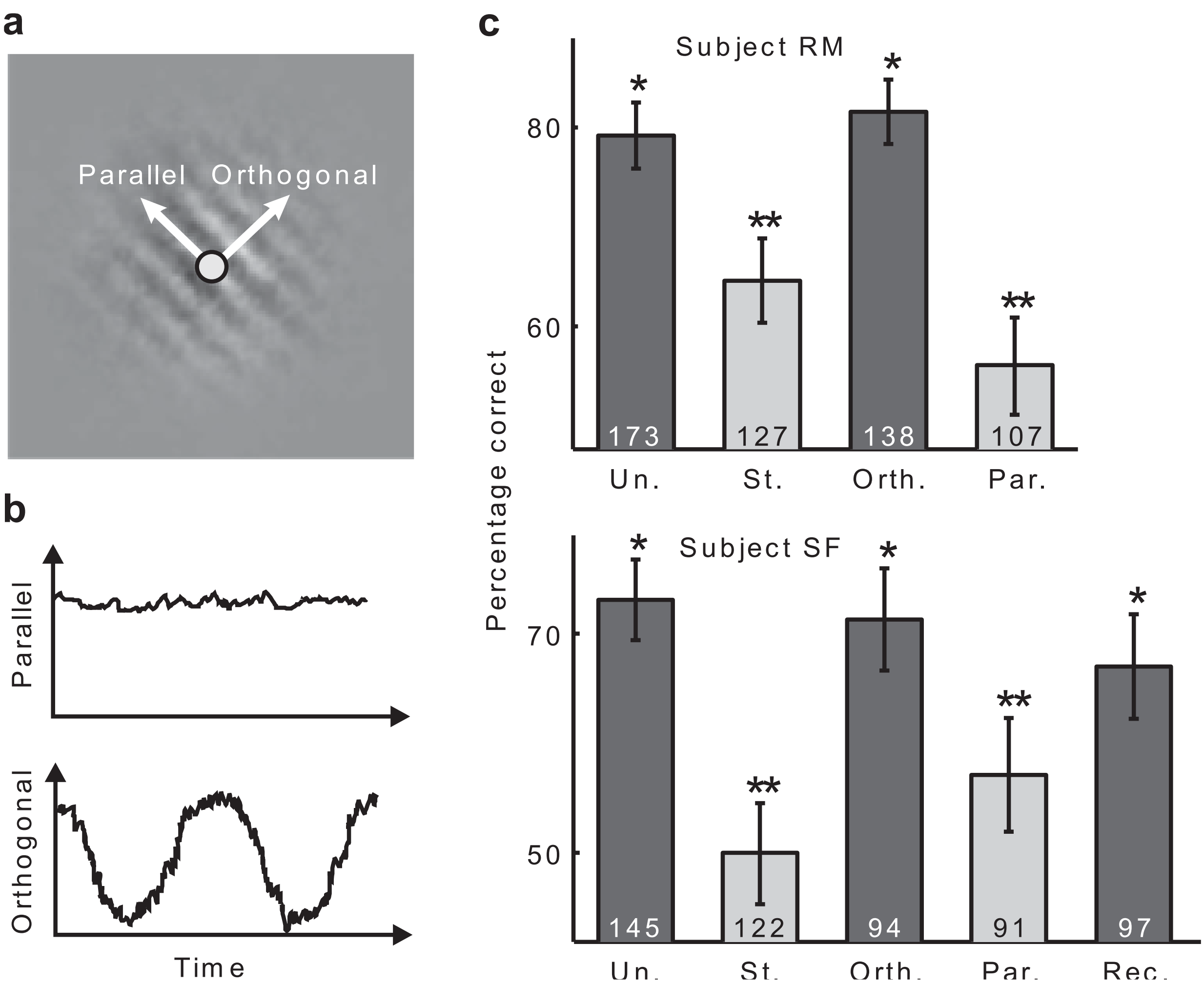


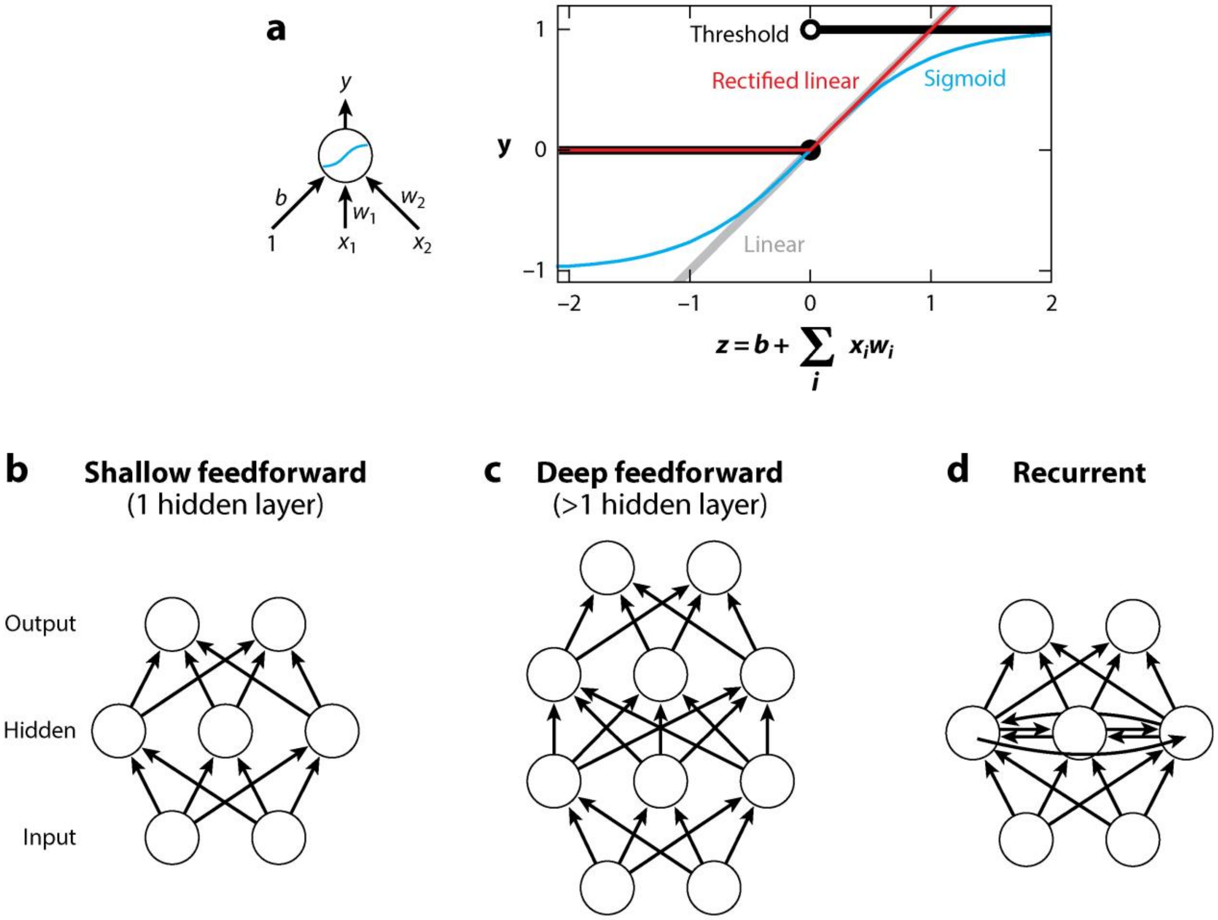
Figure 3. *Consequences of partial retinal stabilization* (Rucci *et al.*, 2007)*. (a) shows the retinal stabilization confined to a single axis that is either orthogonal or parallel to the stimulus (b) shows the changes in luminance information experienced by a retinal receptor (depicted by a circle) along the stimuli when it's confined to an axis parallel to the orientation of the grating (top panel), revealing little information regarding the stimuli, and to an axis orthogonal to the orientation of the grating (bottom panel), which in contrast, has a comparatively high signal to noise ratio. (c) The mean accuracies of correct orientation classification for the high spatial frequency stimuli are impaired significantly when FEMs are restricted to the parallel axis and remain unaffected when they are restricted to the orthogonal axis.*

in the orientation prediction task for the high spatial frequency gratings in the normal fixation condition than when the retinal image motion is selectively eliminated through retinal stabilization since the prediction accuracy almost drops to chance level in the latter condition. In contrast, the performance accuracies in both the unstabilized and stabilized conditions remained similar in the case of low spatial frequency gratings. The finding of this study implies that FEMs are integral to our visual perception of high spatial frequency stimuli. It also indicates that the lack of FEMs, such as when the retinal image motion is selectively eliminated through retinal stabilization, can significantly impair our ability to perceive these stimuli. On the other hand, the presence or absence of FEMs has minimal effect on our perception of low spatial frequency stimuli.

In another experiment, the FEMs were selectively eliminated via retinal stabilization along a particular direction, either orthogonal or parallel, relative to the orientation of the stimuli in order to establish more concrete evidence for the effect of FEMs on classification performance. Eliminating fixational modulations of luminance in a direction orthogonal to the gratings significantly reduced the mean classification performance for high-frequency gratings, while the orientation prediction accuracy of the gratings remained almost the same when the motion of the FEMs was restricted in a direction orthogonal to the gratings, which reveals the most information about the stimuli.

Based on the results of these two experiments, it can be clearly inferred that the absence of fixational modulation facilitated by FEMs impairs the ability to accurately classify high spatial frequency stimuli. Additionally, it also revealed that the temporal modulations induced by FEMs improved the signal-to-noise ratio for high-frequency gratings in comparison to low-frequency gratings. Overall, this study illustrates an additional function enabled by the tiny oculomotor movements apart from preventing visual fading, which is that we are able to take advantage of the luminance modulations produced by the FEMs to enhance fine spatial details and how FEMs act as a mechanism for achieving superresolution by using temporal modulations to encode spatial resolution, otherwise known as spatiotemporal encoding (Rucci *et al.*, 2007).

In recent years, Artificial Neural Networks (ANN) (Figure 4) have been influential as a framework for developing neuroscientific models of the biological vision and the brain in general since they capture aspects of how it processes information, its neural activity, and its behavioural outputs (Khaligh-Razavi and Kriegeskorte, 2014; Yamins *et al.*, 2014). One of the biggest advantages of using ANNs as computational models is to formulate novel hypotheses about information processing in the brain and then testing competing hypotheses under rigorous conditions using different types of learning rules (both supervised and unsupervised learning regimes), architectures, input data, and objective functions that define the ANN (Doerig *et al.*, 2022).



Kriegeskorte N. 2015.
Annu. Rev. Vis. Sci. 1:417–46

Figure 4. *Basic architectures of Artificial Neural Networks (ANN)* (Kriegeskorte, 2015)*. (a) A single unit takes a linear combination of its inputs x_i and bias b and applies a set of weights w_i to calculate an output value z. This is passed through an activation function to generate an output y (image on the right) which adds non-linearity to z. Feedforward neural networks process information in a single direction, from input to output (b,c), while recurrent neural networks have loops in their connections, allowing information to flow in a cyclical manner and enabling the network to maintain a type of memory (d). Feedforward networks that have at most one hidden layer are shallow (b), and more than one hidden layers are deep (c). Non-linearity functions in between hidden layers of ANNs enable the network to capture complex relationships between the inputs and the outputs, allowing it to learn and model highly nonlinear and continuous functions.*

However, despite significant progress in the field, there are still many aspects of biology that remain largely unexplored. Developing biologically plausible models of the perceptual system not only facilitates our understanding of the underlying mechanisms of vision but can also help improve the robustness, efficiency, and accuracy of existing computer vision models. For example, current state-of-the-art models rely heavily on large datasets and extensive training to perform basic object recognition tasks, whereas humans can effortlessly perform these tasks with minimal world knowledge.

One key aspect that distinguishes biological vision from current deep neural network (DNN) models is our ability to actively explore our surroundings using eye movements, integrating visual information across fixations to form a comprehensive representation of the world. This characteristic is not present in most DNN models, which process the entire image input at once, making them incapable of handling dynamic and continuous visual inputs. Incorporating this biological feature into DNNs has the potential to significantly enhance their efficiency by reducing the required pixel count and task complexity by disregarding irrelevant visual features in the input image. By incorporating models of fixational eye movements in particular, computer vision models can gain additionally proposed computational benefits, such as the ability to reveal information at superresolution that would not be discernible with static inputs (Rucci *et al.*, 2007). Such models can not only bridge the gap between the static image-based deep learning models and the continuous and dynamic visual processing in the human brain, but this biological detail allows us to test and further investigate the crucial computational role of FEMs in the early stages of neural information processing.

Recurrent Neural Networks (RNN) are a class of neural networks that can effectively integrate information from temporal sequences of data. Unlike feedforward networks, RNNs can handle sequential data by passing information from one time step to the next through hidden states. This makes RNNs essential for modelling the spatiotemporal encoding aspect of the human perceptual system, which involves high recurrent information flow in which lateral and recurrent connections are widely present (Kriegeskorte, 2015; Schrimpf *et al.*, 2018; Kietzmann *et al.*, 2019; Spoerer *et al.*, 2020). The task of learning where to look effectively in the input images using the neural network model given the portion of the image that the network is focused on (since the FEM model can only look at parts of the image similar to human percept) is framed as the 'control problem.' The control problem in the context of our model refers to the challenge of directing attention to relevant parts of the image. This can be framed as a decision-making process where the agent (in this case, the RNN model of FEM) must select actions (i.e., eye movements/ fixations) to maximize some objective (i.e., information gain or task performance). Reinforcement learning (RL) can be used to solve this control problem by training an agent to learn the optimal action policy through trial-and-error interactions with the environment. Specifically, RL algorithms aim to maximize a cumulative reward signal obtained from the environment in order to learn an optimal policy that maps states to actions. In the context of our model, RL can be used to learn an optimal eye movement policy that maximizes information gain and improves task performance. Therefore, we model FEMs using a biologically plausible deep RNN combining supervised learning and deep Reinforcement Learning (RL) that can learn where to look in images. Our work shares similarities with other attempts that use deep learning to incorporate attentional processing and saccadic eye movements (Mnih *et al.*, 2014; Choi *et al.*, 2022). However, very few models with FEMs exist.

The remainder of this thesis is organized in the following sections. In the methods section, we first describe the datasets used to train the model and then outline different in-silico psychophysics experiment setups. These experiments test whether the neural network model of FEM can learn human-like FEM and reproduce psychophysical data. The results section provides an analysis of the performance of the model on these experiments and interprets the results. Finally, the discussion section summarizes the main findings of the thesis and provides an outlook on the results. It discusses the similarities between the model and human vision and also highlights the future directions of this project, including potential areas for improvement and further research.

Methods

Dataset Preparation:

We create a dataset to reproduce Rucci et al. (2007) 's experiments with Gabor stimuli. However, it is necessary to deteriorate the spatial information on these images before they get processed by the deep neural network model of fixational eye movements in order to accurately represent the lower resolution of the image on the fovea before any fixations compared to the actual perceived resolution of the image by the brain. Therefore, to assess if the model can achieve superresolution, we apply average pooling as a technique to down-sample an image by dividing it into small sub-regions and computing the average value of each sub-region. This technique serves as a front end to the network, allowing us to test if the model can learn the appropriate sequence of fixational eye movements and spatiotemporally integrate information across these eye movements similar to humans to recover the necessary information from the pooled image in order to classify them correctly.

The model was trained and tested on images of Gabor patches similar to the visual input designed and used by Rucci et al. (2007). The training set consists of 7200 images of both high and low spatial frequency gratings, respectively. The high spatial frequency gratings and the low spatial frequency gratings had a frequency of 6.7 cycles and 2.5 cycles, respectively. The gratings placed at the centre of the image were tilted ±45° from the vertical axis and occupied precisely half the size of the 64x64 pixels image with a grey background. A Gaussian filter and a random high-frequency noise were added to the stimuli with low spatial frequency and vice versa to simulate the power spectrum of natural images whose power of the noise is inversely proportional to the square of the spatial frequency. The 7200 images in the training set covered all possible phase values between 0 and 180, with ten images for each phase value split equally between the two possible orientations (-45 and +45) and spatial frequencies (high and low). The test set contained 2000 images of only high or only low-spatial frequency stimuli with random phases and orientation (Figure 5). All of these images are grayscale, and they were generated using Python.

All the images along with the prediction labels and the respective image sizes were stored in a HDF5 file.

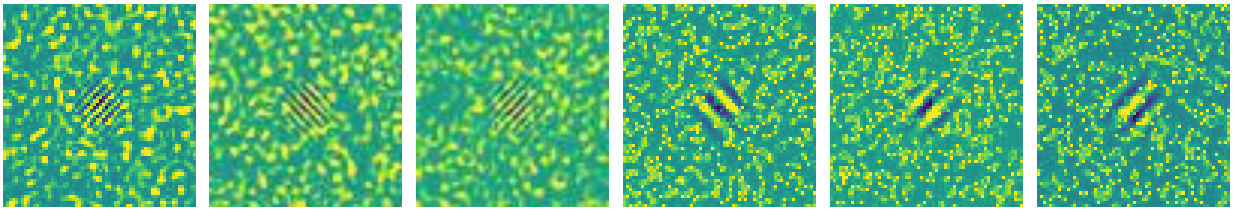


Figure 5. *Examples of high spatial frequency gratings (first three images) and low spatial frequency grating data (last three images) with added noise and Gaussian filter included in the training set.*

Network:

The network that we used to model fixational eye movements is adapted from the Recurrent Attention Model (RAM) (Mnih *et al.*, 2014), written in RLlib (a software library to manage reinforcement learning algorithms), TensorFlow, and Python, was provided by Dr. Adrien Doerig. RAM leverages deep reinforcement learning and recurrent neural network to create a hybrid network model that learns to control eye movements while aiming to accurately classify the input images (Figure 6). Thus, instead of attending all the pixels of the input image at once, the RAM sequentially attends to different regions within the image at each time step then dynamically learns and updates its appropriate representation of the environment using the information gathered from the previously attended locations and uses this history along with the task demands to choose where to attend in the future. We decide to call it FEMNet (FEM-Network)

At each timestep, the agent receives an input image from the training set which is pre-processed by the RLlib framework. This input is a cropped version of the image with dimensions 51x51 pixels, whereas the original size of the image is 64x64 pixels. Therefore, the agent does not get to observe the complete environment but only a portion of them. However, the image label, in addition to the image, is input to the model, so it can use the label information to do the category supervised learning. This label is passed as input concatenated to the image information. Before any further computations, the cropped image is down-sampled via average pooling, as mentioned in the previous section. The amount of down-sampling (output size of the

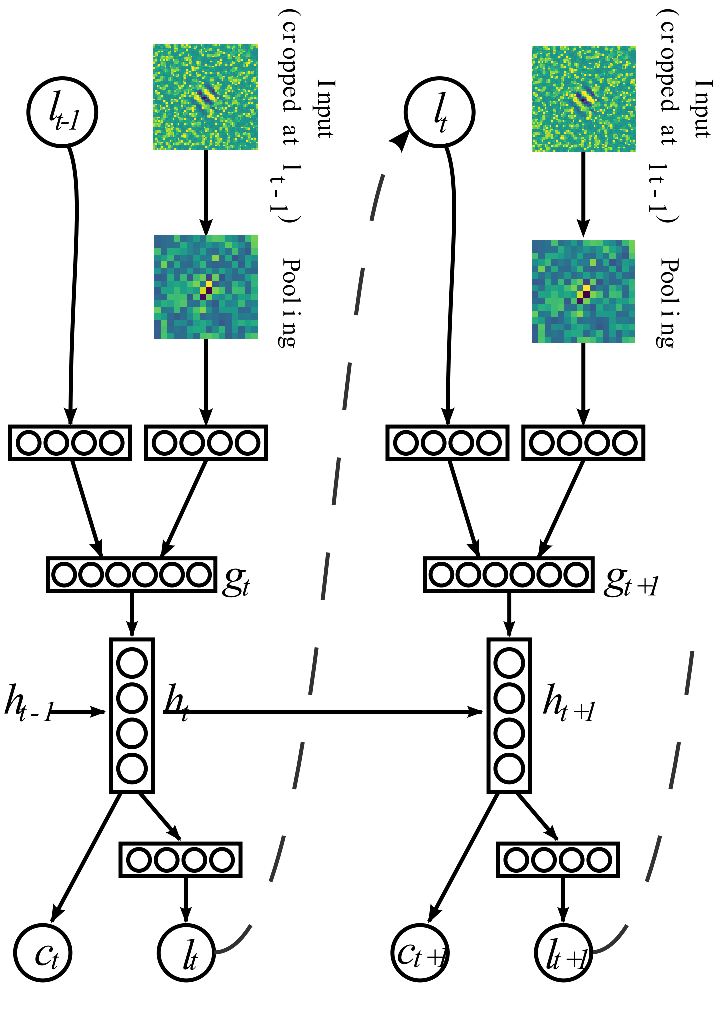


Figure 6. *Figure and description adapted from* (Mnih *et al.*, 2014)*. The environment class in the RLlib provides a flattened cropped image (However, an actual image is shown here for illustration purposes) and the location information to the network. Additionally, it also provides the classification label for the image since the network is trained in a supervised setting. The cropped image and its location are then mapped into a hidden space using independent linear layers and ReLU activation functions, producing a single vector containing information from both components. The neural network model for fixational eye movements consists of three main components: a core network, an attention head (l_t), and a category head (c_t). The core network takes the current gaze position g_t as input, along with the internal state of the model at the previous time step ((h_{t-1}), to produce the new internal state of the model (h_t). The location network and action network use the internal state (h_t) to produce the next location to attend to ((l_t) and the action or classification (c_t), respectively. This iterative process of the basic RNN model is repeated for a variable number of steps.*

image after passing it through the pooling filter) is determined by the size and the stride value of the filter. The size denotes the pixel dimensions of the patch whose average value represents the output for that patch, whereas the stride represents the filter displacement. The pooled input, along with the current fixation location, is then passed through a series of dense layers whose final output is a single vector g_t that contains information from both the image as well as the fixation location, where $g_t = Rect\left(Linear\left(h_g\right) + Linear(h_l)\right)$, where $Rect(x) = \max(x, 0)$ and $Linear(x) = Wx + b\ for\ weights\ W\ and\ bias\ b. h(l) = Rect(Linear(l)$ and $h(g) = Rect(Linear(average\_pooled\_input))$ respectively. h_g and h_l contain 128 (weights denoted by θ_g^0 and θ_g^1) neurons, while g contains 256 neurons (weights denoted by θ_g^2).

g_t is then fed into the core network, which consists of a recurrent neural network that has an internal state which builds a representation of the environment based on the history of fixations. h_t denotes the internal state of the RNN represented by its hidden layers, which are recurrently updated over time through $h_t = f_h(h_{t-1}) = Rect(Linear(h_t - 1) + Linear(g_t)$. The hidden layer acts as an input to two other networks (category head and attention head) whose actions affect the state of the environment. At each timestep, the action network outputs the classification prediction using a dense layer followed by a softmax activation function (a function that converts the scores determined by the action network for each class into probability distributions). The next fixation location is determined by the location network, which also contains a single dense layer similar to the action network. The final location is stochastically drawn from a normal distribution with location output as the mean and a small fixed variance. This is done in order to tame the exploration of the policy. This discrete fixation jumps from one fixation to the other, making the model non-differentiable; however, we use a reinforcement learning algorithm to address this problem. The action network is trained using supervised cross-entropy loss, while the location net is trained using APPO (Petrenko *et al.*, 2020) (Asynchronous Proximal Policy Optimization) with a binary reward (0 if the model predicts the wrong label from the input image after six timesteps and one otherwise) so that the agent learns to maximize the total reward it can attain when interacting with the environment. Both losses are backpropagated through the whole network so that the early shared layers learn features useful for both classification and fixation location.

Overall, the current model differs from the original RAM in the following ways: (1) the cropped images in the network were average pooled before any linear transformations were applied to them, (2) the location network is trained with Asynchronous Proximal Policy Optimization (APPO) algorithm instead of the traditionally used REINFORCE algorithm (Williams, 1992) since the former algorithm is faster in terms of convergence and better in terms of sample efficiency and exploration purposes.

Experiments:

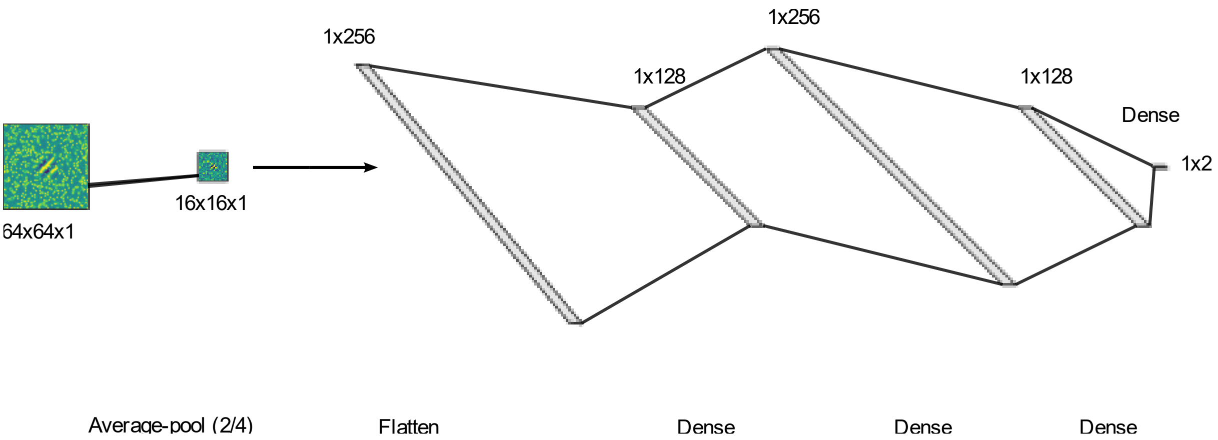


Figure 7. *A feedforward neural network is designed to measure the appropriate amount of average pooling (filter size/ stride value) necessary to eliminate the orientation information from high spatial frequency gratings. Similar to FEMNet, an average pooling layer is present in the front end of the network. The pooled image is flattened and passed through a series of dense layers (4 layers with 128, 256, 128, and 2 units, respectively) that incorporate non-linearities using rectified linear units (ReLU), and the final classification output is generated using a softmax classifier. The model is trained for 10 epochs with the aim of accurately classifying images by using a cross-entropy loss function with the Adam optimizer and a learning rate of 0.001.*

The following experiments primarily test the functional validity of the fixational eye movements model, that is, if they are able to reproduce the functions of FEMs established in humans, specifically based on the ones proposed by Rucci et al.

(Rucci and Victor, 2015). In addition, we also test if the fixational patterns learned by the model can help it attain superresolution on naturalistic.

Experiment 1: Baseline model

In order to make sure that the FEMNet effectively employs spatiotemporal encoding to classify the Gabor dataset, it is crucial to determine the appropriate level of downscaling required for the images, especially for the high spatial frequency gratings. This is essential to prevent the FEMNet from relying on classifying the images just because the network is able to see the orientation information in the input stimuli. Therefore, a feedforward model with the same number of units as the FEMNet network is constructed without the recurrent component (Figure 7).

This feedforward network was trained on the Gabor dataset, which contains both high as well as low-frequency gratings with different filter sizes as well as stride lengths (in the pooling layer – 2/2, 2/3, 3/1, 3/2, 3/3), and we tested the performance of the network on the Gabor test data which either contains only high frequency or only low-frequency stimuli. The underlying idea of the experiment is that the feedforward model would be unable to categorize the high-frequency test data because it is incapable of integrating information from these images temporally, as hypothesized.

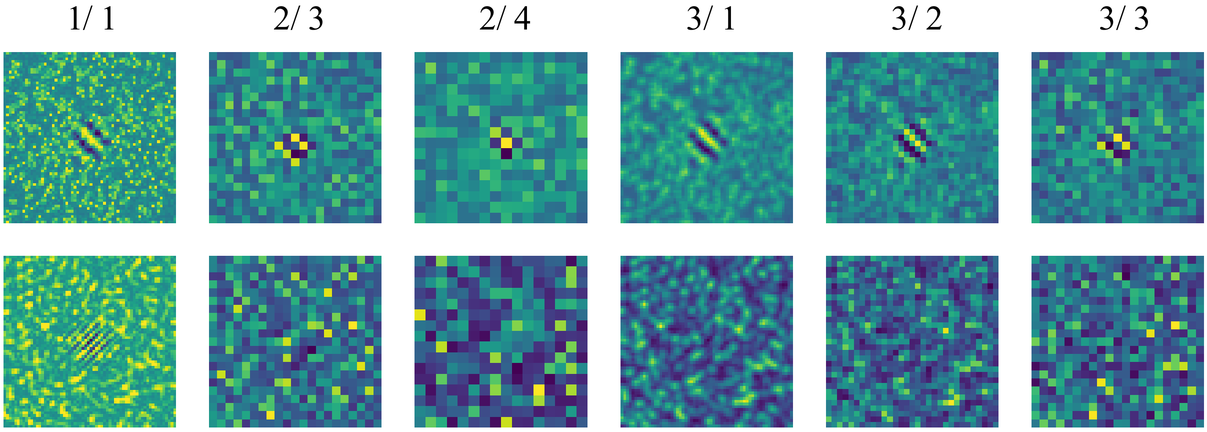


Figure 8. *Examples of average pooled images of both low (top panel) and high (bottom panel) spatial frequency gratings for different kernel sizes and stride lengths. 1/1 (1st column) shows the original image (64x64 pixels), while the subsequent columns display the images with different versions of pooling. All the images were resized to the same dimensions for illustration purposes*

Consequently, the baseline model would fail to learn the relevant features necessary for the classification task on the high-frequency gratings, unlike the low-frequency gratings, where the orientation information would still be visible even after pooling.

Experiment 2: Consequences of selective retinal stabilization

Once the appropriate set of pooling parameters are obtained from the baseline model, we then sought to reproduce Rucci et al. (2007) 's first experiment, where the FEMNet's performance is measured in two cases, with and without stabilization of fixations to see (1) If the network can learn superresolution and (2) If yes, are the eye movements crucial for this feat. In order to implement this, we trained the FEMNet on the dataset consisting of both high and low-frequency gratings. Subsequently, the orientation classification performance of this trained FEMNet is them measured in two kinds of test data, consisting only of high or low-frequency gratings under two conditions, one where the FEMNet is free to make fixations (unstabilized condition) and the other where all the fixations are fixed at the centre of the gratings (which is also the centre of the image), thus replicating the retinal stabilization condition. Testing the model with only low spatial frequency stimuli checks if it is capable of classification when the information on the stimulus is not entirely destroyed. On the other hand, testing it on only high spatial frequency stimuli checks whether the network can make use of the recurrence as well as the fixational modulations to achieve superresolution and recover the orientation information even though it is removed from the static image by pooling (Figure 8).

Next, we test if the FEMs learned by the network are crucial for classification performance by testing the accuracy of FEMNet when the model is forced to fixate at the center of the gratings throughout all time steps instead of being allowed to make eye movements (stabilized condition).

All the experiments involving FEMNet also contain two control models. The first control is a FEMNet that is untrained. The second control is a trained FEMNet, but when testing, the model is forced to make totally random eye movements within the image. This control model helps to determine if the fixation pattern learned by the trained FEMNet is significantly more effective than spatiotemporally integrating information using any random sequence of eye movements.

When retraining deep neural network architectures with different random initial states, significant variations in performance and learning internal representations by the network can occur (Mehrer *et al.*, 2020). Thus, we train six instances of FEMNet (in comparison to six human subjects in Rucci et al. (2007) 's study) and repeat the experiments on every instance of the model. Training multiple instances of the same network and comparing their performances on the classification task as well as their fixation patterns, helps to test the robustness of the FEMs displayed by the FEMNet.

Experiment 3: Consequences of partial retinal stabilization

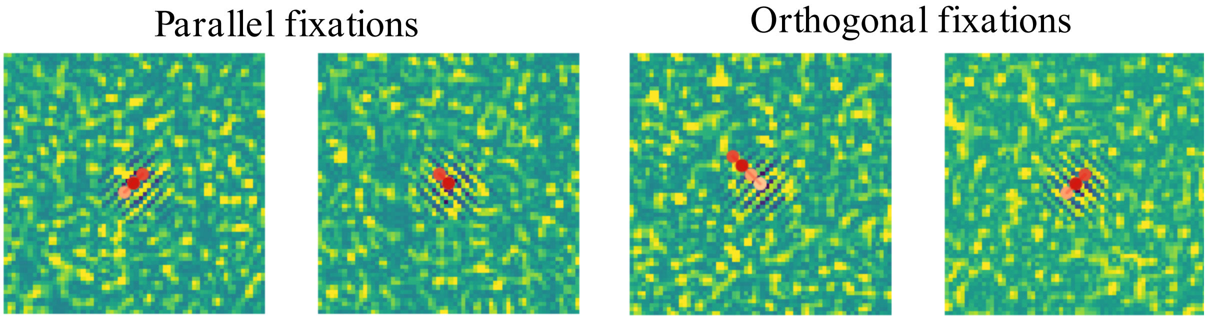


Figure 9. *Experimental setup for partial retinal stabilization experiment. First fixations are always at the centre of the Gabors (yellow dot). Successive fixations made by the FEMNet are restricted to fixate either along the axis parallel or orthogonal to the high spatial frequency test images.*

In the next experiment, we investigate how the FEMNet makes use of FEM orientation by constraining the FEMs to a single axis that is either orthogonal or parallel to the gratings. This experiment uses the same model that was used in the previous experiment, except we now test this FEMNet on the dataset containing only high-frequency images under two conditions. At test time, the model is forced to make fixations along an axis parallel/ orthogonal to the orientation of the grating. We hypothesize that the model that is constrained to make orthogonal fixations will perform better in the classification task since the orthogonal motion along the grating provides maximum information about the stimulus orientation. On the other hand, any sequence of fixational eye movements that are constrained to be parallel to the stimulus only reveal changes in noise patterns rather than providing information about luminance. This experiment provides additional support for the cause-and-effect relationship between fixational eye movements and the performance of the FEMNet on the orientation classification task.

Experiment 4: Applications of FEMs on naturalistic images

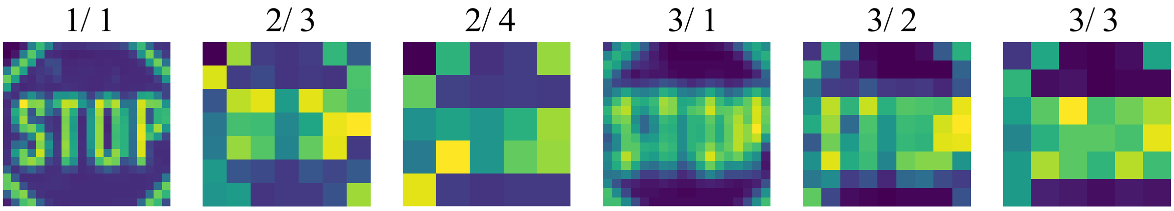


Figure 10. *Example of average pooled images of one of the 43 classes in the GTSRB training dataset for different kernel sizes and stride lengths. 1/1 (1st column) shows the original image (20x20 pixels). All the images were resized to the same dimensions for illustration purposes.*

Finally, we move on from the simplistic Gabor stimuli and test the effectiveness of the FEMNet model in a naturalistic setting. Therefore, we use a new dataset consisting of street signpost images called the GTSRB dataset (German Traffic Sign Recognition Benchmark)(Stallkamp *et al.*, 2011). This dataset consists of 39,209 images in the training set and 12,630 images in the test set with up to 43 classes of signpost images with dimensions 20x20 pixels. A baseline model similar to the one before is made to test if a feedforward model can classify the images properly after sufficiently pooling the dataset. Additionally, we created a simple Recurrent Neural Network Baseline model by attaching a recurrent unit with six timesteps in the penultimate layer of the feedforward model. This recurrent baseline model tests the importance of FEMs for the FEMNet, independent of the recurrent integration of information. However, a selective retinal stabilization experiment similar to experiment 1b is also performed with the FEMNet trained on the GTSRB dataset with stabilized and unstabilised fixations, and similar control models are used. As in the previous experiments, six instances of the FEMNet are trained on the street sign dataset and the robustness of their responses is compared similarly to the previous experiment.

Results

Experiment 1: Baseline model

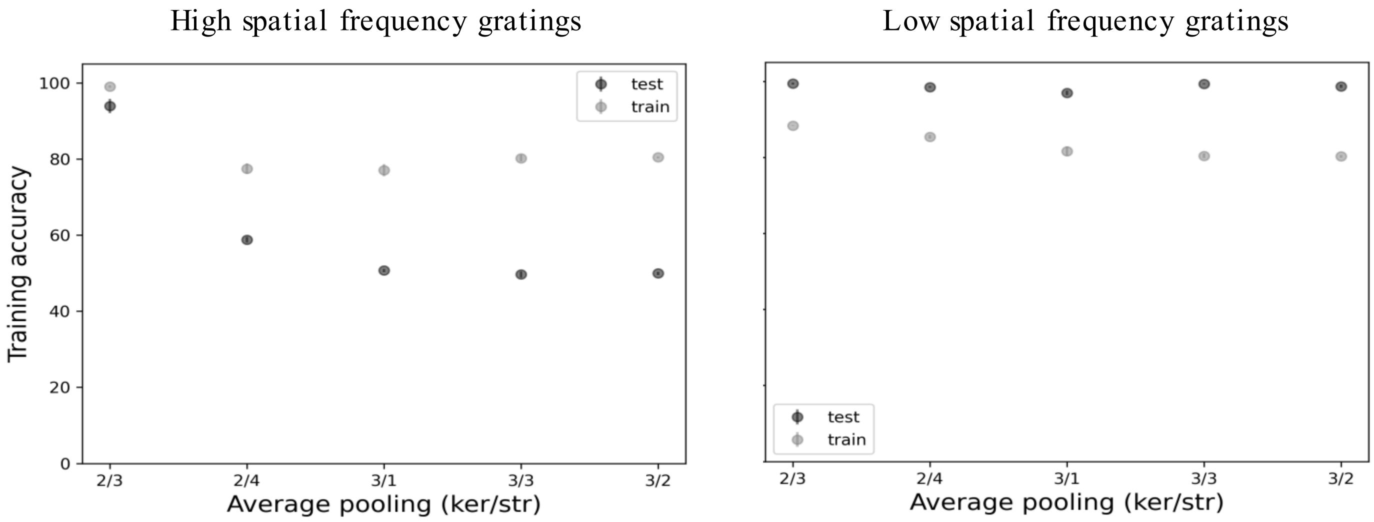


Figure 11. *Plot on the left shows the performance of the baseline model (trained on both high and low-frequency gratings) on a test set that only consists of high spatial frequency gratings for different parameters of average pooling (kernel size and stride length denoted by ker/str) whereas the plot on the right shows the performance of the baseline model on a test set that only consists of low spatial frequency gratings.*

The baseline model acts as a sanity check. It can be observed that the training, as well as the test accuracy, declines and remains relatively similar for the feedforward model after the average pooling parameters (kernel size and stride value, denoted as ker/str in Figure 11) 2/4, for test data consisting of only high spatial frequency gratings. This suggests that pooling destroys information in the high freq gratings, preventing the ff control net from categorizing them. Hence, the FEM-net would need to learn useful eye movements to classify high-frequency pooled gratings. In contrast, the baseline model is able to achieve high training as well as test accuracies on the low-frequency gratings (>80% test and training accuracy on all the versions of pooling), which shows that the when orientation information in the gratings is not completely destroyed by after pooling, the baseline model is able to perform the classification task. Therefore, for the upcoming experiments, it is justified to consider any pooling size in the set of [2/4, 3/1, 3/2, 3/3] in the front end of the FEMNet to accurately represent the images perceived by the lower resolution of the fovea. In the following, a 2/4 pooling was used.

Experiment 2: Consequences of selective retinal stabilization

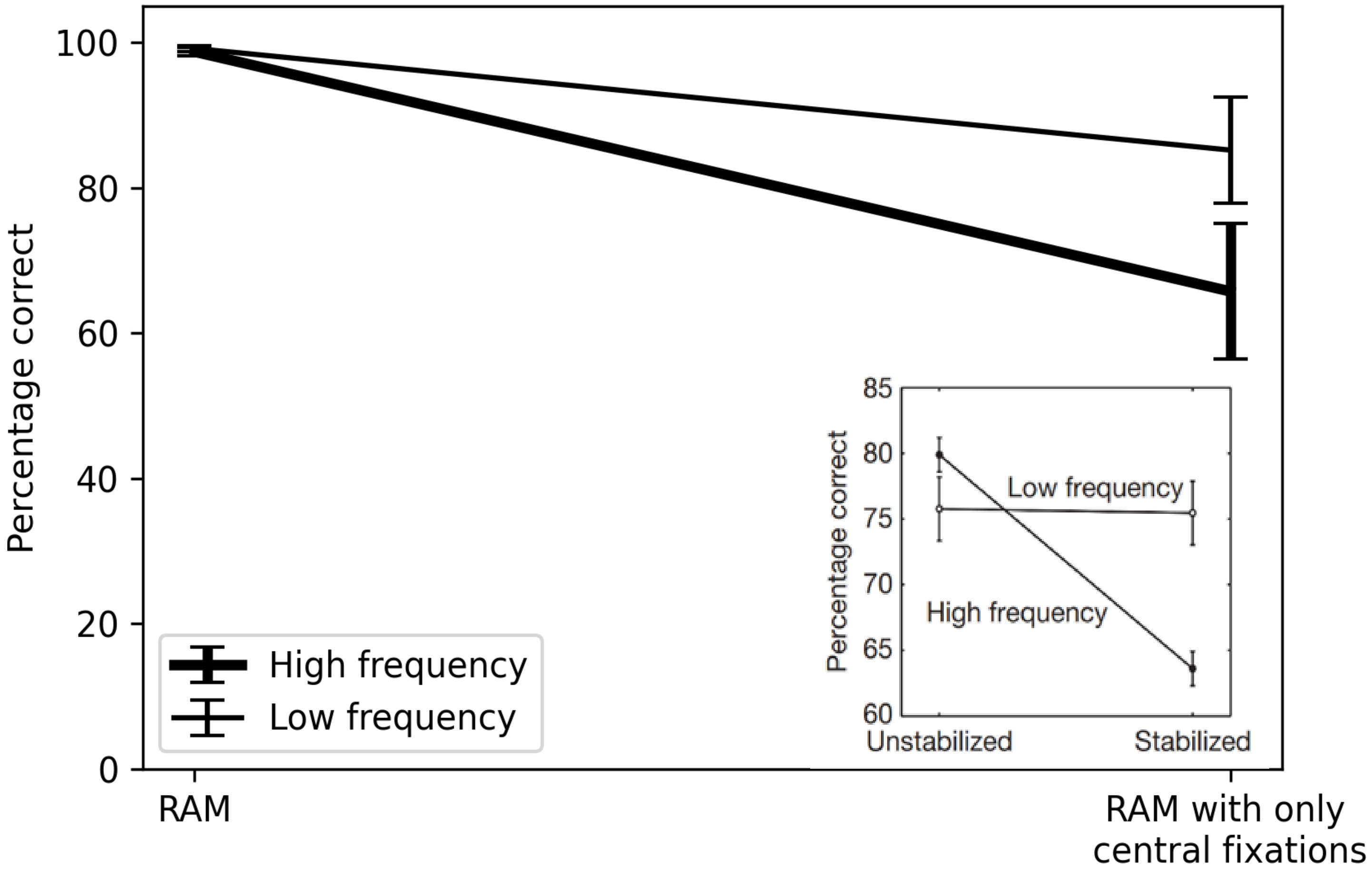


Figure 12. *Plot illustrates the performance of FEMNets under two different viewing conditions. On the left-hand side, the FEMNet model's accuracy is measured under free-viewing conditions, whereas on the right-hand side, its accuracy is measured under the retinal stabilization conditions. In this latter condition, the model is forced to make only central fixations in all six timesteps. High spatial frequency vision is impaired in the stabilization condition, whereas the performance of the FEMNet remains relatively unchanged for low spatial frequency stimulus. The inset shows the performance of humans in a similar psychophysical study that investigates the function of enhancement of fine spatial detail by FEMs (Figure 2b), conducted by* (Rucci *et al.*, 2007)*. Error bars denote the standard deviation of the six instances of FEMNets.*

The FEMNet is capable of performing the task of classifying high spatial frequency gratings when fixational eye movements (FEMs) are allowed. This is demonstrated by the high accuracy achieved by the model in the unstabilised condition, as shown in Figure 12. However, when the retinal stabilization condition is replicated in the FEMNet by forcing the model to fixate at the centre of the image for all timesteps, the model's performance drops significantly in the central fixation case for high spatial frequency stimuli. This result supports the hypothesis that FEMs are crucial for enhancing high spatial frequency vision.

To further reinforce this hypothesis, a random fixation control experiment is performed where the FEMNet is allowed to make random fixations during testing without any spatial constraints. This experiment shows that the model's performance is significantly worse than the unstabilised condition (this control model achieves 58% on the high-frequency test set and 76% on the low-frequency test set), indicating that simply having any form of fixational eye movements is not sufficient for performing the task.

In contrast, for low spatial frequency stimuli, the FEMNet achieves almost 100% accuracy in both unstabilised and stabilised conditions, suggesting that FEMs may not be necessary for this task. These findings are consistent with Rucci et al.'s study and provide initial evidence that the FEMNet behaves similarly to human FEMs by enhancing high spatial frequency vision.

Additionally, the effectiveness of the learned sequence of fixations was tested with a control for this experiment that included an untrained network (that performs an arbitrary sequence of fixations) whose test accuracy was ~50% for both high and low spatial frequency test stimuli, which is close to chance level accuracy.

Similar to human FEMs, the nature of fixations displayed by FEMNet also changes with respect to the type of stimulus. This can be observed in Figures 12 and 13, where the fixation patterns executed by the FEMNet are visibly different depending on the orientations of the gratings. In the case of low spatial frequency stimulus, the angles between fixations are not very different. This might be explained by the fact that FEMNet is not necessarily required to execute fixations in a strategized manner, as shown in experiment 2. The model performs fixations that are relatively orthogonal to the gratings for high spatial frequency stimuli oriented at 45°, which

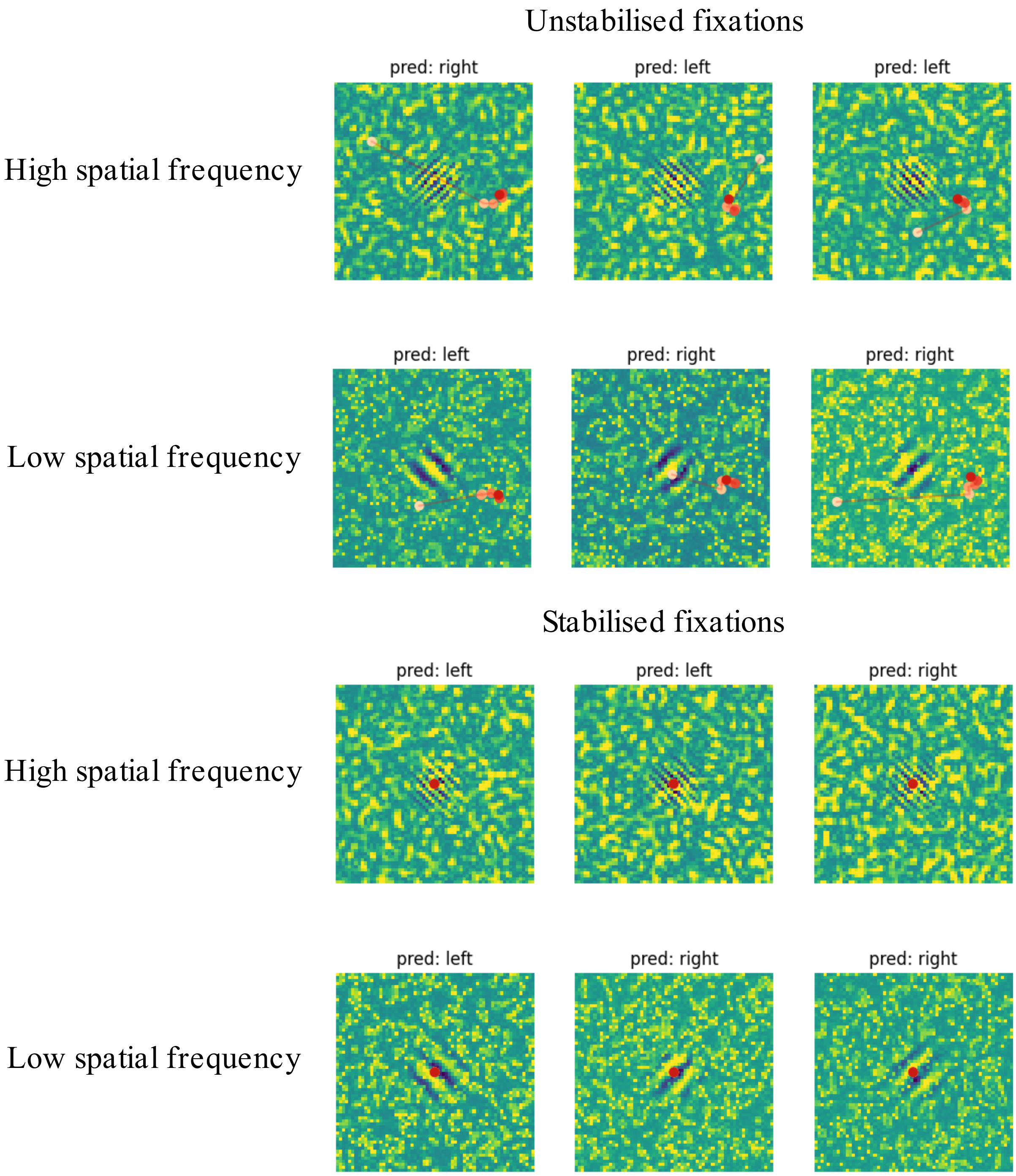


Figure 13. *The top half panel of the figure displays the fixation sequence of one of the FEMNet instances in unstabilized condition for both high and low spatial frequency stimuli. The yellow dot indicates the initial fixation, while the red dot shows the final (sixth) fixation. In the bottom half of the panel, the FEMNet's fixations are shown under stabilized conditions for high and low frequency gratings, where it is constrained to fixate at the centre of the stimulus at all time steps. The labels at the top of each stimulus are FEMNet's prediction for that image after it makes the sequence of fixations shown in the figure.*

provides the maximum information about the orientation of the stimuli. However, for the stimuli oriented at -45°, the FEMNet stops moving towards 45° (or) in a direction parallel to the gratings and shifts to other angles, as shown in Figure 14. The peak of peak frequency of the angles between successive fixations shifts significantly for different orientations of the gratings. This shows that the network changes its behaviour depending on the stimulus, adapting its visual sampling strategy to the current input.

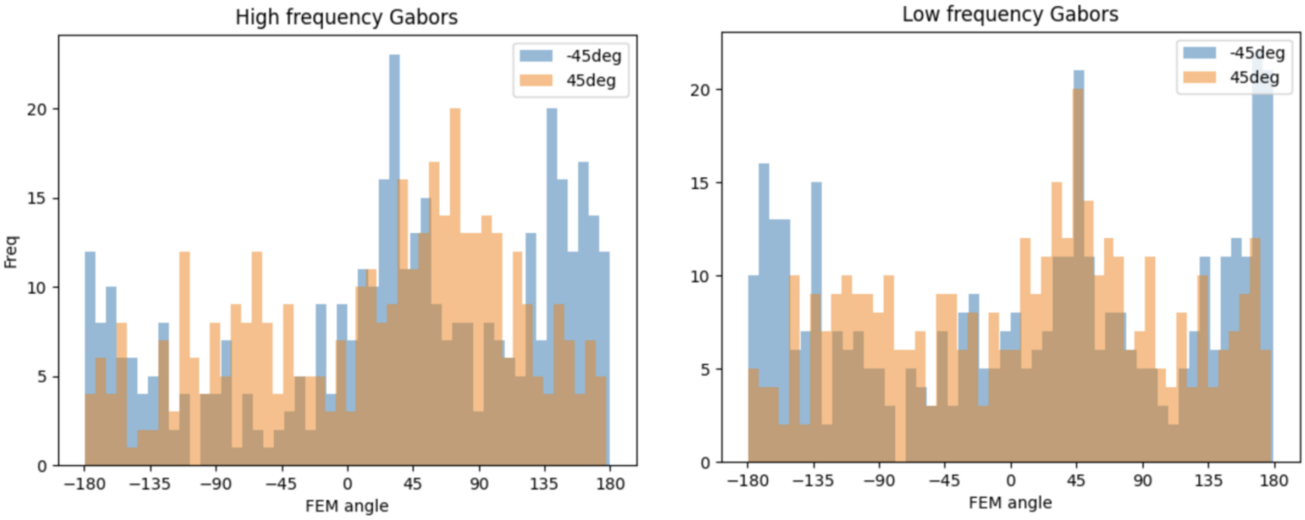


Figure 14. *Histogram of angles between successive fixations made by a FEMNet instance collected for 100 images in each class (left/ right oriented images) on both high and low frequency test datasets.*

FEMNets with different pooling versions were not able to reproduce the same effect observed in humans. FEMNets could not learn to distinguish the orientations of the high spatial frequency gratings better than chance level using FEMs for average pooling versions 3/1, 3/2, and 3/3. In these cases, the pooling was too high to be overcome with FEMs.

Experiment 3: Consequences of partial retinal stabilization

The difference in the performance between the FEMNet constrained to execute FEMs along a certain axis that is parallel or perpendicular to the direction of the grating orientation is not very high, as hypothesized (Figure 15). Although, the

FEMNet, which performed orthogonal fixations, achieved a higher orientation classification accuracy as expected (roughly 7%). Both of these models were able to outperform the FEMNet forced to fixate randomly across the image. This suggests that the learned fixation patterns of the FEMNet are crucial for achieving a high classification accuracy, and the direction of the fixation movement relative to the stimulus plays an essential role.

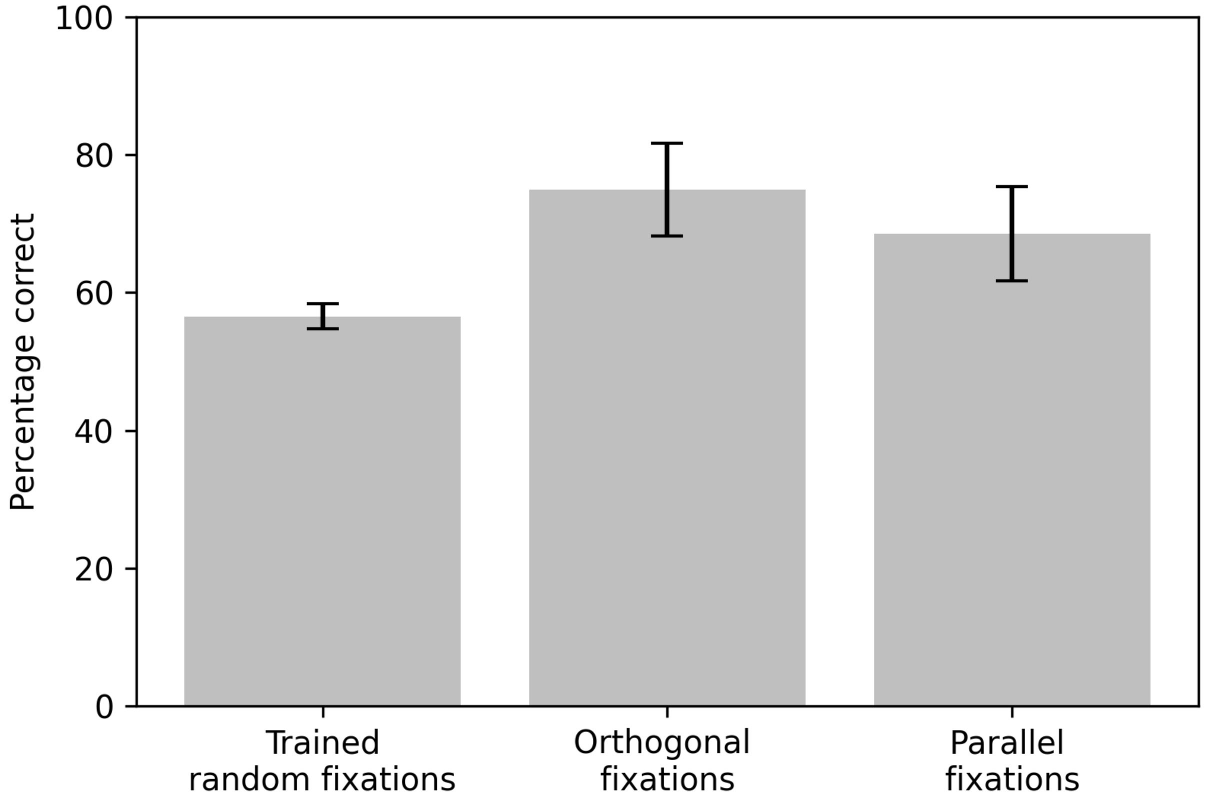


Figure 15. *Plot illustrates the performance of the FEMNets whose movement on the stimulus is restricted to an axis orthogonal or parallel to the grating. The performance of a trained FEMNet forced to perform random fixation is included as a control. Error bars denote the standard deviation of the six instances of FEMNets.*

While it is true that fixational movements that are orthogonal to the stimuli provide the most information about the orientation of the grating, the performance of the FEMNets forced to carry out such movements does not always reflect this idea. One of the factors that might contribute to this discrepancy is the FEMNet's ability to learn the most effective fixations during training. This is clear from the fixation sequences made by one of the FEMNet instances (Figure 14). Ideally, the model should learn to make orthogonal fixations to optimize information gain from the high spatial frequency gratings. However, it can be observed that the model has learned to make fixations that are not the most optimal for the classification task. This could be due to the fact that the training set/ the Gabor classification task is not as complex. Thus, the models that were trained on this task also had no obligation to learn the most optimal way to fixate so as to achieve a high classification accuracy during training. This is backed up by the fact that the mean reward received by the FEMNet during training quickly converges to 1.

Experiment 4: Applications of FEMs on naturalistic images

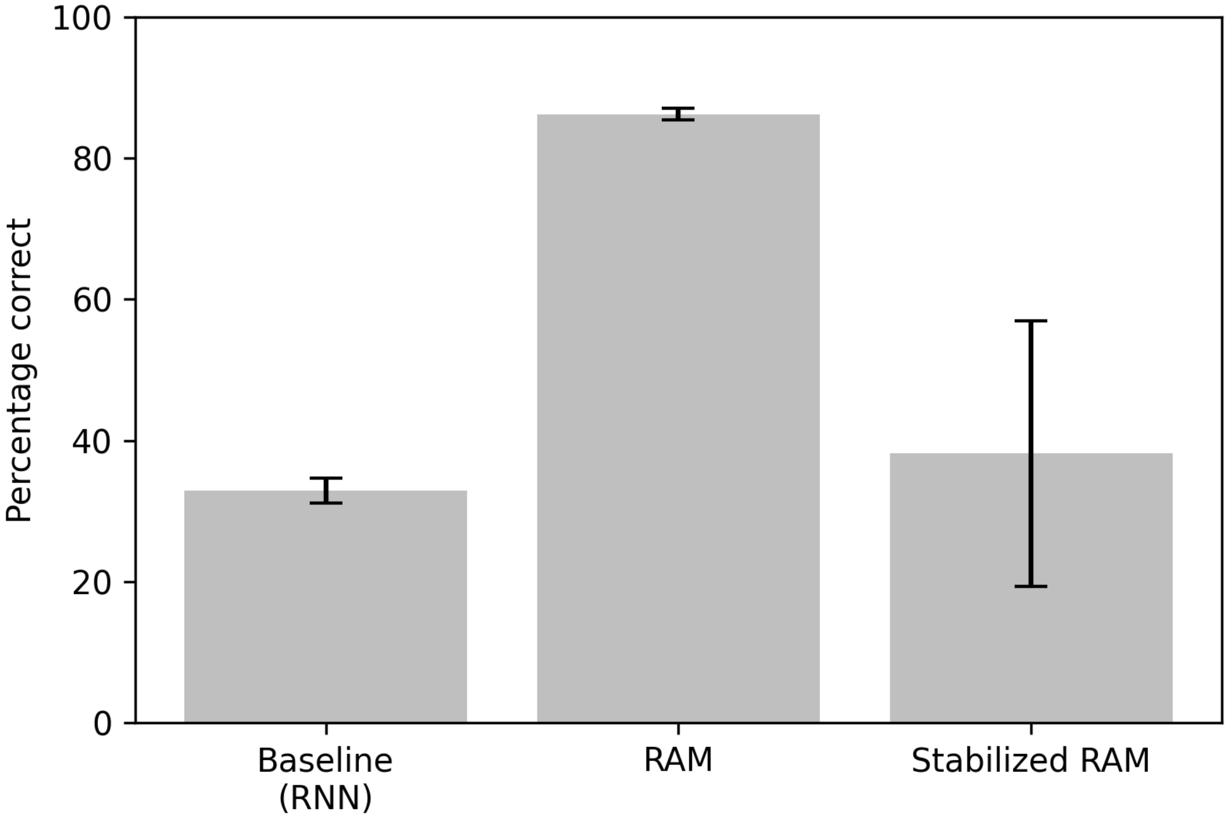


Figure 16. *Classification accuracy of the baseline model with Recurrence, FEMNets, and FEMNets with stabilized FEMs, respectively (from left to right) on the GTSRB test dataset. Error bars denote the standard deviation of the six instances of FEMNets.*

Since the stimuli in this experiment are different from the Gabors used in the previous experiments, the pooling parameters were chosen again to ensure that a control network without eye movement cannot classify the pooled images. The pooling parameter for this dataset was chosen to be 3/3 (kernel size/ stride value) after making sure that the baseline model, as well as the baseline model with recurrence, could not classify the images in the GTSRB test dataset, the recurrent baseline model, which lacked FEMs, was not able to perform the classification task effectively, achieving only 33% mean accuracy. This result highlights the importance of FEMs in visual processing. On the other hand, the FEMNet model, which incorporated FEMs, achieved significantly higher accuracy in the unstabilized condition with a mean test accuracy of 86.24%. Furthermore, the stabilized FEMNet was not able to perform the classification task effectively, in line with the hypothesis that FEMs are necessary for effective visual processing.

Discussion

The Fixational Eye Movement Network (FEMNet) proposed in this thesis demonstrates human-like behaviour through its ability to selectively perform fixational eye movements and achieve superresolution. We replicated and extended the finding that FEMs help enable the enhancement of high spatial frequency vision using psychophysical stimuli (Gabor patches) that have helped shed light on the functions and mechanisms regarding the the role of FEMs in humans. One of the important aspects of computational modelling of biological phenomena is to understand the relevant features or components of the model which are necessary to reproduce the human like-behavior. In FEMNet, this could be attributed to its recurrent connectivity and its ability to perform attention-guided eye movements. The inability of the feedforward models with no adaptive eye movements to reproduce the experimental findings in humans further supports this claim.

The FEMNet learned to make tiny movements akin to FEMs without any explicit constraint. Not only do they enhance high spatial frequency vision, but they also adapt their fixation modulations according to the type of stimuli (in the case of Gabor stimuli, this was orientation), similar to the effect observed in humans(Intoy and Rucci, 2020). It is worth noting that the ability of FEMNet to mimic human-like visual processing mechanisms is dependent on certain pooling parameters. In fact, the FEMNet was unable to achieve superresolution in high spatial frequency images that were over-pooled. This suggests that the information contained in these images was simply insufficient to learn any useful representations relevant to the classification task, even though the model attempted to encode information in a spatiotemporal manner.

Moreover, the third experiment conducted on FEMNet highlights the importance of the direction of the FEM sequence. Interestingly, the difference in performance for FEMNet fixating only orthogonally versus only parallelly to the gratings is not as pronounced as in humans. It is also surprising that the FEMNet constrained to make fixations along the direction of the grating could learn to perform better than chance level test accuracy. One potential explanation given for this is the complexity of the training set used. This could be improved in many ways. Currently, the noise field, spatial frequency, and the phase of the Gabor patches are the only variable parameters in the training data. To improve the training data, the center of the Gabor patches could be randomly placed anywhere within the radius of an arbitrary pixel from the centre of the whole image, in addition to variations in the noise field, spatial frequency, and phase of the patches. This would add an extra parameter to the training set and increase the complexity of the input data during training, since the FEMNet fixates in and around the Gabor patch. This additional complexity could help the model avoid converging too quickly, as was observed in some FEMNet instances, and learn a fixation sequence that optimizes the amount of information gained from the stimuli. There were other minor details in the training data that could be improved. For example, the relative power of the noise to the gratings was not equal, and the spatial frequency of the stimuli was not exactly the same as the one used in Rucci et al. (2007) 's experiment (the differences between the frequencies were ±1). By addressing these issues and incorporating more varied and complex training data, the model's performance in the parallel versus orthogonal learning experiment could be improved.

Finally, FEMNets have shown that they have potential real-world applications (Experiment 4). Although the percentage accuracy of the FEMNet on the GTSRB dataset isn't flattering, it is important to note that the training procedures were not optimized for maximum accuracy in image classification tasks. Instead, the main focus of the study was to model FEMs in a biologically plausible manner. Future work could involve scaling up the model to make it more biologically plausible by including two separate streams (such as dorsal and ventral) to carry out the where vs what pathways. Additionally, a more accurate version of foveal sampling and peripheral vision could be incorporated into the model. Furthermore, the study could be extended to explore the importance of other features, such as efference copies and other parameters that affect the model's performance. Previous work in computer vision has demonstrated the significance of deep neural network models that incorporate ideas from biological vision. For instance, models that simulate retinal foveation and sample parts of the image similar to eye movements have been shown to outperform traditional models against adversarial examples (Gant *et al.*, 2021; Choi *et al.*, 2022). Additionally, these models have been found to act as biological proxies for data augmentation, leading to improved performance in self-supervised learning (Wang *et al.*). These findings suggest that incorporating biological principles into computer vision models can yield valuable insights and improvements in performance. Such applications could also be tested with the current or a scaled-up version of the FEMNet.

In conclusion, our study sheds light on the remarkable computational advantages that Foveated Eye Movements (FEMs) offer to the human visual system. By using fewer neurons, FEMs enable efficient processing of fine spatial details, which is especially crucial given the limited resolution of the retina. Our findings have shown that recurrent systems that learn targeted eye movements can achieve better classification performance via superresolution, replicating essential experimental findings on human FEMs, but also demonstrated substantial improvements in computer vision performance using naturalistic stimuli.

Importantly, our modelling work underscores the view that FEMs are not haphazard but rather a well-orchestrated strategy of the brain to exploit its recurrent connectivity for computational advantages as it naturally emerges in our simple network. Hence, our results suggest an explanation for the emergence of FEMs in human vision, which is to facilitate the efficient processing of detailed visual input despite the limited retinal resolution.

Therefore, our work adds to the growing body of evidence that integrating insights from biology can yield valuable advancements in machine learning and computer vision research and vice versa.

References

1. Carney, T, and Klein, SA (1997). Resolution Acuity is better than Vernier Acuity. Vision Research 37, 525–539.

2. Choi, M, Zhang, Y, Han, K, Wang, X, and Liu, Z (2022). Human Eyes Inspired Recurrent Neural Networks are More Robust Against Adversarial Noises.

3. Ditchburn, RW, and Ginsborg, BL (1952). Vision with a stabilized retinal image. Nature 170, 36–37.

4. Doerig, A et al. (2022). The neuroconnectionist research programme.

5. Gant, J, Banburski, A, and Deza, A (2021). Evaluating the Adversarial Robustness of a Foveated Texture Transform Module in a CNN.

6. Geisler, WS (1984). Physical limits of acuity and hyperacuity. J Opt Soc Am A 1, 775–782.

7. Hennig, MH, and Wörgötter, F (2003). Eye micro-movements improve stimulus detection beyond the Nyquist limit in the peripheral retina. In: Proceedings of the 16th International Conference on Neural Information Processing Systems, Cambridge, MA, USA: MIT Press, 1475–1482.

8. Intoy, J, and Rucci, M (2020). Finely tuned eye movements enhance visual acuity. Nat Commun 11, 795.

9. Katsuki, F, and Constantinidis, C (2014). Bottom-up and top-down attention: different processes and overlapping neural systems. Neuroscientist 20, 509–521.

10. Khaligh-Razavi, S-M, and Kriegeskorte, N (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLOS Computational Biology 10, e1003915.

11. Kietzmann, TC, Spoerer, CJ, Sörensen, LKA, Cichy, RM, Hauk, O, and Kriegeskorte, N (2019). Recurrence is required to capture the representational dynamics of the human visual system. Proceedings of the National Academy of Sciences 116, 21854–21863.

12. Kowler, E, and Steinman, RM (1979). Miniature saccades: eye movements that do not count. Vision Res 19, 105–108.

13. Kriegeskorte, N (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. Annual Review of Vision Science 1, 417–446.

14. Kuang, X, Poletti, M, Victor, JD, and Rucci, M (2012). Temporal Encoding of Spatial Information during Active Visual Fixation. Current Biology 22, 510–514.

15. Lambertus, S, Bax, NM, Fakin, A, Groenewoud, JMM, Klevering, BJ, Moore, AT, Michaelides, M, Webster, AR, Wilt, GJ van der, and Hoyng, CB (2017). Highly sensitive measurements of disease progression in rare disorders: Developing and validating a multimodal model of retinal degeneration in Stargardt disease. PLOS ONE 12, e0174020.

16. Martinez-Conde, S (2006). Fixational eye movements in normal and pathological vision. Prog Brain Res 154, 151–176.

17. Martinez-Conde, S, and Macknik, SL (2007). Windows on the mind. Sci Am 297, 56–63.

18. Martinez-Conde, S, and Macknik, SL (2008). Fixational eye movements across vertebrates: Comparative dynamics, physiology, and perception. Journal of Vision 8, 28.

19. Martinez-Conde, S, Otero-Millan, J, and Macknik, SL (2013). The impact of microsaccades on vision: towards a unified theory of saccadic function. Nat Rev Neurosci 14, 83–96.

20. Mehrer, J, Spoerer, CJ, Kriegeskorte, N, and Kietzmann, TC (2020). Individual differences among deep neural network models. Nat Commun 11, 5725.

21. Mnih, V, Heess, N, Graves, A, and kavukcuoglu, koray (2014). Recurrent Models of Visual Attention. In: Advances in Neural Information Processing Systems, Curran Associates, Inc.

22. Otero-Millan, J, Macknik, SL, and Martinez-Conde, S (2012). Microsaccades and blinks trigger illusory rotation in the “rotating snakes” illusion. J Neurosci 32, 6043–6051.

23. Otero-Millan, J, Macknik, SL, and Martinez-Conde, S (2014). Fixational eye movements and binocular vision. Front Integr Neurosci 8, 52.

24. Otero-Millan, J, Troncoso, XG, Macknik, SL, Serrano-Pedraza, I, and Martinez-Conde, S (2008). Saccades and microsaccades during visual fixation, exploration, and search: Foundations for a common saccadic generator. Journal of Vision 8, 21.

25. Petrenko, A, Huang, Z, Kumar, T, Sukhatme, G, and Koltun, V (2020). Sample Factory: Egocentric 3D Control from Pixels at 100000 FPS with Asynchronous Reinforcement Learning.

26. Ratliff, F, and Riggs, LA (1950). Involuntary motions of the eye during monocular fixation. Journal of Experimental Psychology 40, 687–701.

27. Rolfs, M (2009). Microsaccades: small steps on a long way. Vision Res 49, 2415–2441.

28. Rucci, M, Iovin, R, Poletti, M, and Santini, F (2007). Miniature eye movements enhance fine spatial detail. Nature 447, 852–855.

29. Rucci, M, and Victor, JD (2015). The unsteady eye: an information-processing stage, not a bug. Trends Neurosci 38, 195–206.

30. Schrimpf, M et al. (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? 407007.

31. Spoerer, CJ, Kietzmann, TC, Mehrer, J, Charest, I, and Kriegeskorte, N (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. PLOS Computational Biology 16, e1008215.

32. Stallkamp, J, Schlipsing, M, Salmen, J, and Igel, C (2011). The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In: The 2011 International Joint Conference on Neural Networks, San Jose, CA, USA: IEEE, 1453–1460.

33. Troncoso, XG, Macknik, SL, Otero-Millan, J, and Martinez-Conde, S (2008). Microsaccades drive illusory motion in the Enigma illusion. Proc Natl Acad Sci U S A 105, 16033–16038.

34. Wang, B, Mayo, D, Deza, A, Barbu, A, and Conwell, C On the use of Cortical Magnification and Saccades as Biological Proxies for Data Augmentation.

35. Williams, RJ (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 8, 229–256.

36. Yamins, DLK, Hong, H, Cadieu, CF, Solomon, EA, Seibert, D, and DiCarlo, JJ (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences 111, 8619–8624.