

Research Project in Mechatronics Engineering

**A CONVOLUTIONAL NEURAL NETWORK THAT
EXTRACTS DEPTH FROM IMAGES**

Samuel Reedy

Project Report ME7-2023

Co-worker: Nic Zwager

Supervisor: Dr Luke Hallum

13 October 2023



ENGINEERING
DEPARTMENT OF MECHANICAL
AND MECHATRONICS ENGINEERING

A CONVOLUTIONAL NEURAL NETWORK THAT EXTRACTS DEPTH FROM IMAGES

Samuel Reedy

ABSTRACT

Navigating through our three-dimensional world demands effective depth perception, a complex task seamlessly executed by the human visual system (HVS). The Binocular Energy Model (BEM), which simulates disparity-tuned neurons in the primary visual cortex (V1), has been a widely acknowledged model in this domain. However, its limitations, particularly in handling certain visual stimuli like anti-correlated random dot stereograms (RDS), and its somewhat antiquated structure spotlight a compelling opportunity for enhancement and re-exploration through the lens of modern computational methodologies.

In this investigation, a convolutional neural network (CNN), inspired by the foundational principles of the BEM, was developed, aiming to delve deeper into the intricacies of the HVS. The CNN was subjected to a rigorous training regimen, utilising a dataset of 50,000 RDSs, categorised into correlated and anti-correlated types, over four epochs. The training emphasised the development and subsequent analytical unpacking of convolution filters, with the overarching goal of unearthing new insights into the HVS's internal mechanisms.

The finalised model exhibited the capability to extract depth from both correlated and anti-correlated RDSs, even those absent during the training phase. It demonstrated an ability to attenuate its output to anti-correlated RDS and adeptly navigate the correspondence problem, which entails the accurate matching of features to each eye while disregarding incorrect matches-areas where the BEM falters. The analysis of the model illuminated several biological implications, suggesting the potential applicability of a hybrid model in the V1 and revealing that a mere two subunits, contrary to the BEM's suggested four, may be sufficient for depth extraction.

While the findings from this investigation pave a promising path forward in understanding depth perception and the role of V1 neurons, they also underscore the imperative for further research and validation. The necessity to solidify these implications and explore potential enhancements in both the architecture and functionality of the model is evident. This exploration lays the foundation for further research into the complex mechanisms underpinning visual perception and depth extraction in biological systems.


DECLARATION

Student

I Samuel Reedy..... hereby declare that:

1. This report is the result of the final year project work carried out by my project partner (see cover page) and I under the guidance of our supervisor (see cover page) in the 2023 academic year at the Department of Mechanical and Mechatronics Engineering, Faculty of Engineering, University of Auckland.
2. This report is not the outcome of work done previously.
3. This report is not the outcome of work done in collaboration, except that with a project sponsor as stated in the text.
4. This report is not the same as any report, thesis, conference article or journal paper, or any other publication or unpublished work in any format.

In the case of a continuing project: State clearly what has been developed during the project and what was available from previous year(s):

Signature: 
Date: 13/10/23

Supervisor

I confirm that the project work undertaken by this student in the 2023 academic year ~~is~~ / **is not** (*strikethrough as appropriate*) part of a continuing project, components of which have been completed previously.

Comments, if any:


Signature: 
Date: 13/10/23

Table of Contents

Acknowledgements	vii
Glossary of Terms	viii
Abbreviations	viii
1 Introduction	1
2 Litreature Review	1
2.1 Biological Mechanisms for Depth Perception	1
2.1.1 Visual Cortex	1
2.1.2 The Correspondence Problem	2
2.2 Random dot Stereograms	2
2.3 Existing Depth Perception models	4
2.3.1 Binocular Energy Model	4
2.3.2 BEM Extensions	5
2.4 Convolutional Neural Networks	5
2.5 Research Gaps	6
3 Methods	6
3.1 RDS Generation	6
3.1.1 Correlated RDS	6
3.1.2 Anti-correlated RDS	6
3.2 CNN Architecture	7
3.2.1 Overall Structure and Workflow	7
3.2.2 Subunit Structure and Functionality	8
3.3 Training	9
3.4 Model Analysis	9
3.4.1 Binocular Receptive Fields	9
3.4.2 Correspondence Problem	10
4 Results	11
4.1 Model Performance and Validation	11
4.2 Analysis of Filters and Subunits	12
4.2.1 Filter Characteristics and Deviations	12
4.2.2 Subunit Disparity Detection Mechanism	12
4.2.3 Response to Correlated and Anti-correlated Disparities	13
4.3 Addressing the Correspondence Problem	14
4.4 Effect Subunit Quantity Adjustment	15
4.4.1 Comparative Performance	16
4.4.2 Efficacy and Limitations	17
4.4.3 Characteristics and Functionality of Filters	17
5 Discussion	18
5.1 Model Behaviour and Disparity Encoding	18
5.2 Biological Implications and Filter Characteristics	19
6 Conclusions	19
7 Future Work and Further Exploration	20

References	20
Appendix A Supporting Figures	24

List of Figures

Figure 1	Illustration of the correspondence problem, showcasing the broad RFs of complex cells that encompass multiple targets within a visual scene. The visual system must discern the correct binocular matches from numerous possibilities (all intersections of rays), with only the matches within the horizontal ellipse deemed accurate. Image adapted from [1].	3
Figure 2	Example of a correlated random-dot stereogram (CRDS) and an anti-correlated random-dot stereogram (ACRDS). Image sourced from [2].	3
Figure 3	Schematic of the Binocular Energy Model, tuned for a disparity of 0, and composed of 4 simple subunits (S). Each simple cell is modelled using Gabor functions, with positive and negative parts representing the ON and OFF regions respectively [3], and convolves over the left and right input images. The convolution outputs are summed, half-wave rectified, and squared ($y = pos(x)^2$), generating the output for each subunit. The outputs from all four subunits are then aggregated to form the complex cell (Cx). Displayed subunits are in quadrature, meaning the Gabor functions' phases are shifted by $\frac{\pi}{2}$ between them. Image sourced from [4].	4
Figure 4	Two potential disparity encoding methods for BEM simple cells. (a) Positional encoding and (b) Phase encoding. Image sourced from [5].	4
Figure 5	Output of the BEM in response to correlated and anti-correlated RDS, represented as the column sum of the output. The graph demonstrates the model's inability to attenuate to anti-correlated RDSs, evidenced by a ratio of approximately 1 between the peak from the correlated RDS and the dip from the anti-correlated RDS.	5
Figure 6	Example of a 25x25 pixel correlated RDS alongside its corresponding disparity map. The left and right images act as CNN inputs, while the disparity map, emphasizing the figure with a value of 1 due to its 1-pixel disparity shift, is used to assess the model's performance.	7
Figure 7	Example of a 25x25 pixel anti-correlated RDS, featuring left and right images. Despite having a disparity of 1, it encodes values of 0.5 in the no disparity region and 0 in the disparity area, intending to prompt the model to reduce its output when faced with an anti-correlated RDS.	7
Figure 8	A schematic of the general CNN architecture, illustrating that the left and right input are passed to every subunit. The outputs of subunits are then added together to produce the predicted disparity map. The architecture allows modification of the number of subunits as needed.	8
Figure 9	A schematic of a subunits architecture. The subunit takes two inputs which undergo convolution, are then added together, and finally undergo activation. The dimension of the input images it retained between the input and the output.	8

Figure 10	Approach for deriving binocular RF. Utilising a bar stimulus with independent positional shifts for the left and right bars, spanning from -10 to 10 pixels, a 21 x 21 stimulus grid was constructed (grid sizes can differ based on selected disparities). The four separate grids symbolise combinations: dark-dark, bright-bright, dark-bright, and bright-dark (black = -1, white = 1, grey = 0). Displayed is the bright-dark grid. Disparity tuning curves were formulated by summing each pixel along a constant disparity line, mirroring the relative disparity compared to the left bar. This produced 41 data points, of which the central 21 were used as they contain the most relevant data. The method is adapted from [4].	10
Figure 11	Method for adjusting filters to calibrate the BEM (with a 1x33 filter length) to different disparities. Although diagrams are shown for subunit 1, equivalent shifts are executed across all subunits to guarantee thorough model calibration.	11
Figure 12	Predicted disparity maps corresponding to the provided 25x25 correlated and anti-correlated random dot stereograms, each featuring a 10-pixel-wide figure and a disparity of 1.	11
Figure 13	Visualisation of convolution filters for the 4-subunit model, each featuring a major and a minor extrema.	12
Figure 14	Plots of the eight filters within the model, showcasing the impulse response for each bar combination without disparity and the output from each subunit during the bar experiment, underscoring the receptive fields of each filter.	13
Figure 15	Illustration of the method to reduce the no disparity response while preserving the peak correlated disparity response and minimising the anti-correlated response, facilitating a mathematically attenuated response to anti-correlated RDS. The left graph represents the BEM's response, while the right graph depicts the trained model's response. Note that the response represents the column sum of an ideal disparity map.	14
Figure 16	Comparative exploration of the BEM variants and the trained model in navigating the correspondence problem. Each model was exposed to the same 100 correlated RDS at each disparity, with the plots illustrating the mean and standard deviation of the responses.	15
Figure 17	Depiction of RFs and the corresponding disparity tuning curves for both the BEM and the trained 4-subunit model, detailed for each bar permutation.	16
Figure 18	Response of models with a range of subunit counts (from 1 to 4) to identically correlated and anti-correlated RDS To enhance the visualisation of attenuation effects, the response is displayed as the sum of all pixels in each column.	17
Figure 19	Visualisation of convolution filters for the 1 and 2-subunit models, demonstrating the 1-subunit model trained to have identical filters to one of the subunits in the 2-subunit model.	17
Figure 20	Plots of the 4 filters within the 2-subunit model, showcasing the impulse response for each bar combination without disparity and the output from each subunit during the bar experiment, underscoring the receptive fields of each filter.	18
Figure A1	Visualisation of 25x25 correlated RDS with a disparity of 1 being passed through the trained model.	24
Figure A2	Full analysis of the BEM subunits.	25

Acknowledgements

I would like to thank the following people:

Nic Zwager, my project partner, for all the continuous work and time he put into this project.

Dr Luke Hallum, our supervisor, for his insightful guidance and consistent support throughout the project

The New Zealand eScience Infrastructure (NeSI), for enabling our research through access to their high-performance computing facilities.

Glossary of Terms

Anti-correlated RDS	A variant of RDS where pixel values within the figure are inverted, altering visual stimulus and depth perception.
Binocular Cue	A depth or distance visual cue that involves input from both eyes.
Binocular Disparity	The difference in image location of an object seen by the left and right eyes.
Binocular Energy Model	A model explaining aspects of disparity-tuned neurons in V1, involving rigid constraints and parallel, excitatory elements.
Complex Cell	A neuron in the visual system that exhibits spatial phase invariance and responds robustly to oriented gratings across various spatial phases.
Convolutional Neural Network	A class of deep neural networks, most commonly applied to analysing visual imagery.
Correlated RDS	A type of RDS where a selected disparity region is shifted in one image relative to the other, inducing a depth illusion.
Correspondence Problem	A challenge in stereoscopic vision related to determining corresponding parts of two images received by each eye.
Depth Perception	The visual ability to perceive the world in three dimensions and estimate the distance to an object.
Epoch	One complete forward and backward pass of all the training examples in machine learning.
Gabor Function	A Gaussian-windowed sinusoidal waveform.
Kernel	A matrix used to apply effects to an image via convolution.
Major Extremum	The point on the graph with the maximum magnitude, irrespective of the sign.
Minor Extremum	The point on the graph with the second largest magnitude, irrespective of the sign.
Monocular Cue	A visual cue to depth or distance that can be used by one eye alone.
Phase Shift	In the context of the BEM, altering the phase of one of the Receptive Fields within a subunit to encode disparity.
Positional Shift	In the context of the BEM, horizontally shifting one of the RFs within a single subunit to encode disparity.
Random-dot Stereogram	A pair of images consisting of random dots, which when viewed together, produce a perception of depth and 3D scene.
Receptive Field	A region in the visual field which, when stimulated, modulates the firing rate of a neuron.
Simple Cell	A neuron in the visual system primarily involved in processing visual information about oriented edges or gratings.
Stereopsis	The perception of depth produced by the reception in the brain of visual stimuli from both eyes.
Visual Cortex	The part of the cerebral cortex responsible for processing visual information.

Abbreviations

ACRDS	Anti-Correlated Random Dot Stereogram
BEM	Binocular Energy Model
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CRDS	Correlated Random Dot Stereogram
CUDA	Compute Unified Device Architecture
cuDNN	CUDA Deep Neural Network library
GPU	Graphics Processing Unit
HVS	Human Visual System
MSE	Mean Squared Error
RDS	Random Dot Stereogram
ReLU	Rectified Linear Unit
RF	Receptive Field
SGD	Stochastic Gradient Descent
V1	Primary Visual Cortex
V2	Secondary Visual Cortex
V3	Tertiary Visual Cortex

1. Introduction

Navigating our three-dimensional world is a complex task made easier by a process known as depth perception. Depth perception seamlessly integrates monocular and binocular cues to facilitate spatial awareness and object recognition. Monocular cues, such as size, shadow, and overlay, provide depth information from the perspective of a single eye, while binocular cues leverage the slightly different images perceived by each eye to enhance depth perception through a phenomenon known as stereopsis. Despite the apparent ease with which we perceive depth, the underlying mechanisms, especially those involving binocular disparity and stereopsis, remain complex and not fully understood, despite extensive research. This perceptual complexity has become a significant exploration point in computational modelling and robotics, where replicating human-like depth perception can enhance machine interaction and navigation within physical spaces.

This project is based on the Binocular Energy Model (BEM), a foundational computational model that provides insights into depth perception mechanisms within the primary visual cortex (V1). Recognising the BEM's roots in a computationally limited era, this research employs modern computational technologies by developing a convolutional neural network (CNN) aimed at mimicking the human visual system. Utilising Random-dot Stereograms (RDS), known for inducing depth perception through stereopsis, the project navigates through the intricacies of exploring and emulating visual depth perception mechanisms.

The ultimate goal is to analyse the trained CNN, exploring its mechanisms and learning pathways, in pursuit of uncovering new insights into the human visual system's complex mechanisms.

2. Litreature Review

This literature review explores the multifaceted domain of depth perception, focusing on the biological mechanisms and computational models, notably the Binocular Energy Model (BEM), that seek to understand and replicate it. Special attention is given to the utilisation of Random Dot Stereograms (RDSs) and the application of Convolutional Neural Networks (CNNs) in modelling and investigating visual perception and depth analysis in both biological and computational realms.

2.1 Biological Mechanisms for Depth Perception

Depth perception is integral to human vision, enabling the recognition of a three-dimensional environment and significantly influencing object recognition and spatial awareness [6]. Depth perception is informed by both monocular and binocular cues, emphasising binocular disparity for discerning fine-depth differences [7]. Originating from the horizontal separation of the eyes, binocular disparity not only provides slightly varied visual inputs to each eye due to the differential positioning on the left and right retina but also encompasses disparities within the image structures themselves [8]. This facilitates a robust perception of depth, even in the absence of other cues [9]. The interpretation of binocular disparity to perceive depth is known as stereopsis.

2.1.1 Visual Cortex

The primary visual cortex (V1) is pivotal in the initial processing of visual data for stereopsis, serving as the first stage in the visual pathway where neurons can be activated

by stimulation from either eye and exhibit substantial binocular interactions when both eyes receive concurrent stimuli [10]. Particularly, neurons in V1, which can be categorised into two main types: simple and complex cells, have localised receptive fields and typically demonstrate tuning to orientation and spatial frequency [11,12]. These neurons, especially the binocular ones, compare the relative position of visual stimuli in the left and right eyes, establishing a foundational framework for the brain to compute essential depth information for stereopsis. This depth information, once processed in V1, is conveyed to higher-level neurons in the secondary (V2) and tertiary (V3) visual cortex, where more complex aspects of depth perception are computed, facilitating a refined interpretation of three-dimensional space and bolstering visual accuracy [13].

Simple cells in the V1 play a pivotal role in initial visual information processing, utilising their unique receptive fields (RFs) that contain distinct light-sensitive (ON) and dark-sensitive (OFF) regions. These cells are adept at identifying oriented edges or bars of light in visual stimuli due to their ability to sum inputs and suppress interactions between the ON and OFF regions, thereby exhibiting a preference for specific orientations [14].

Additionally, the response of simple cells is influenced by the phase of visual stimuli, such as sinusoidal gratings, adjusting their activity based on the alignment of wave-like patterns with their RFs. This nuanced, phase-specific interaction is vital for the spatial processing of visual stimuli, contributing to the complex visual information processing network within the V1 [15].

Conversely, complex cells in V1 navigate through significant nonlinear spatial integration and are recognized for their spatial phase invariance, enabling them to generate robust responses to oriented gratings across various spatial phases, provided they align with the cell's preferred orientation [12]. The diversity of complex cells is pronounced, with their receptive fields forming a heterogeneous population. They are defined by exclusion, being identified as any cortical neuron that does not possess a simple receptive field [16]. Compared to simple cells, complex cells generally present well-defined disparity tuning curves, indicating their nuanced role in visual perception [17].

2.1.2 The Correspondence Problem

In the domain of stereopsis, the correspondence problem emerges as a pivotal challenge, intricately tied to the phenomenon of binocular disparity, which results in slightly varied images being perceived by each eye. The complexity of the problem is rooted in the necessity to accurately correlate features between the two eyes, a task that is notably complicated by instances where responses to inaccurate matches (false matches) can be as pronounced as those to true matches [18].

Complex cells, distinguished by their notably large RFs, are especially vulnerable to the correspondence problem. This vulnerability stems from the likelihood of several image features in a visual scene, which are optimally excitatory to the cell, being situated within each complex cell's RF. This situation amplifies the challenge of accurately identifying matches and navigating through the myriad of possible binocular correspondences [1,4].

2.2 Random dot Stereograms

RDSs have been pivotal in exploring depth perception and stereopsis, comprising two nearly identical random dot images with a specific disparity region, or figure, shifted in one image relative to the other, thereby generating a disparity [9]. This method uniquely isolates binocular depth perception by crafting depth through the binocular combination

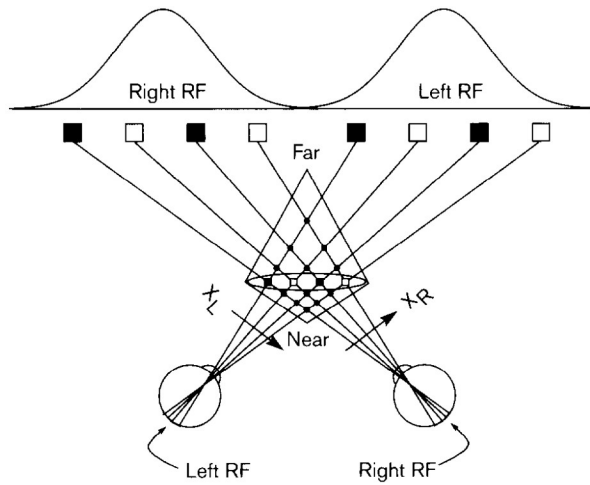


Figure 1 Illustration of the correspondence problem, showcasing the broad RFs of complex cells that encompass multiple targets within a visual scene. The visual system must discern the correct binocular matches from numerous possibilities (all intersections of rays), with only the matches within the horizontal ellipse deemed accurate. Image adapted from [1].

of the monocular half views, providing a simplified yet effective context for studying the intricate HVS [9, 19, 20]. Ensuring the efficacy of an RDS involves adhering to several guidelines, such as maintaining dot consistency, implementing a subtle horizontal shift, and avoiding pattern repetition, all crucial for inducing a depth illusion while minimising visual noise [21]. This project zeroes in on correlated and anti-correlated RDSs, with the correlated RDS adhering to the format previously described and the anti-correlated variant inverting the pixel values within the figure, as illustrated in Figure 2.

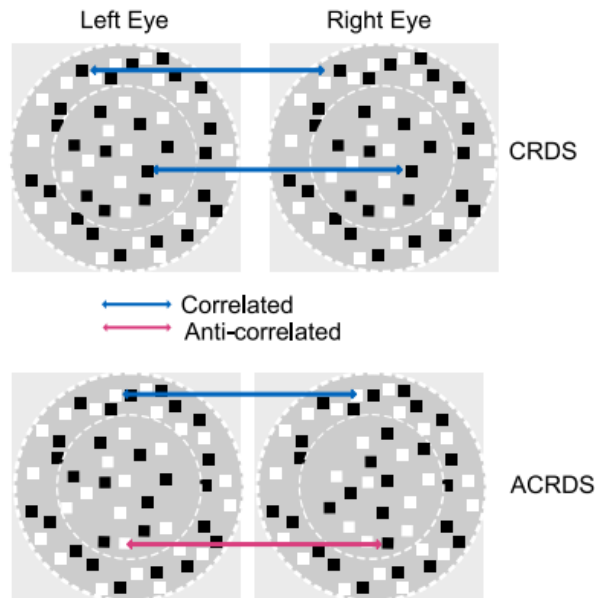


Figure 2 Example of a correlated random-dot stereogram (CRDS) and an anti-correlated random-dot stereogram (ACRDS). Image sourced from [2].

While some psychophysical studies have observed a reversed perceived depth direction in anti-correlated RDSs compared to correlated ones, attributing this to opponent processing where responses to neurons tuned equally but oppositely are computed [22–25], others have found no evidence of depth perception in anti-correlated RDSs [26–28]. Despite the ambiguity in depth perception, anti-correlated RDSs activate disparity-tuned neurons in the

visual cortex, with the expected high positive correlations at the accurate disparity being inverted, resulting in an inverted disparity tuning function [2,6]. Notably, neurons in V1 exhibit this inversion effect while also demonstrating an attenuated response magnitude [6,18].

2.3 Existing Depth Perception models

2.3.1 Binocular Energy Model

The Binocular Energy Model (BEM) has garnered widespread recognition for its capability to illuminate various aspects of disparity-tuned neurons within the V1. It mandates the existence of two parallel, excitatory elements that form a quadrature pair [3,29]. Figure 3 provides an outline of the model, which can encode disparity in binocular simple cells through either phase or positional shifts. Both methods necessitate that each subunit be tuned to the same disparity to accurately tune the model to a singular disparity [3,5].

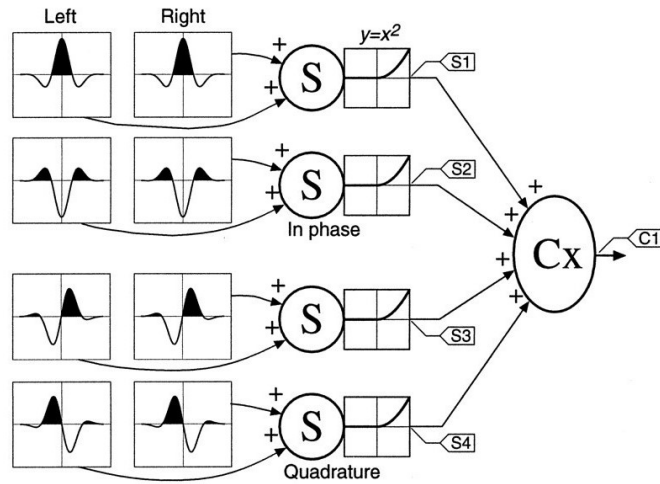


Figure 3 Schematic of the Binocular Energy Model, tuned for a disparity of 0, and composed of 4 simple subunits (S). Each simple cell is modelled using Gabor functions, with positive and negative parts representing the ON and OFF regions respectively [3], and convolves over the left and right input images. The convolution outputs are summed, half-wave rectified, and squared ($y = pos(x)^2$), generating the output for each subunit. The outputs from all four subunits are then aggregated to form the complex cell (Cx). Displayed subunits are in quadrature, meaning the Gabor functions' phases are shifted by $\frac{\pi}{2}$ between them. Image sourced from [4].

Within the BEM, disparity encoding of the binocular simple cells can be done in one of two ways: phase or positional shifts. Figure 4 illustrates the two methods of encoding disparity for the BEM simple cells, showcasing both positional and phase encoding. Each subunit must be tuned to the same disparity for the model to be accurately tuned to a single disparity [1].

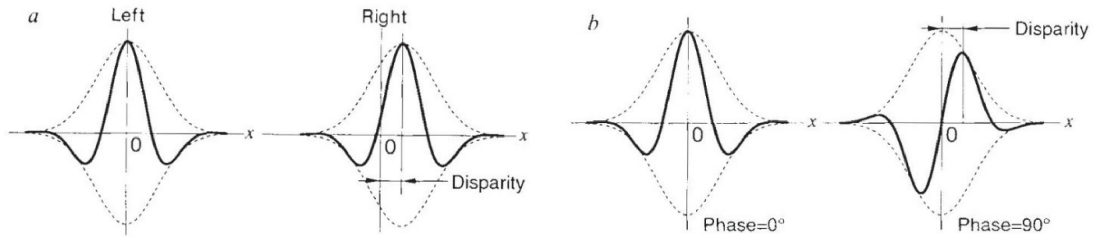


Figure 4 Two potential disparity encoding methods for BEM simple cells. (a) Positional encoding and (b) Phase encoding. Image sourced from [5].

However, despite its broad acceptance, the BEM is not without limitations. It struggles

with the correspondence problem and does not attenuate its response to anti-correlated RDSs [6, 18, 30]. Figure 5 illustrates the BEM’s inability to attenuate to anti-correlated RDSs, revealing a ratio of approximately 1 between the peak from the correlated RDS and the dip from the anti-correlated RDS.

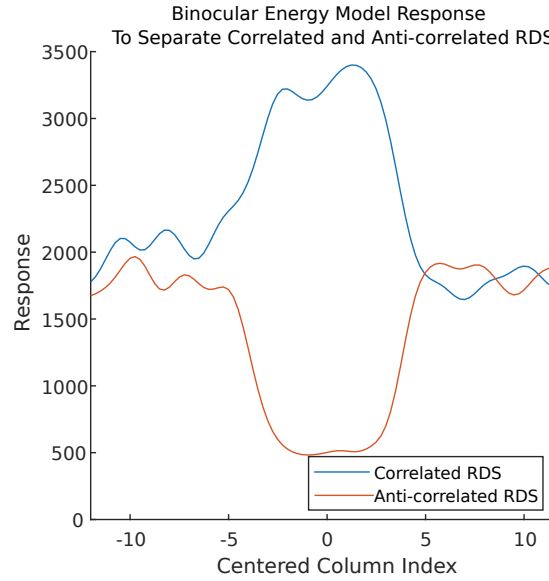


Figure 5 Output of the BEM in response to correlated and anti-correlated RDS, represented as the column sum of the output. The graph demonstrates the model’s inability to attenuate to anti-correlated RDSs, evidenced by a ratio of approximately 1 between the peak from the correlated RDS and the dip from the anti-correlated RDS.

2.3.2 BEM Extensions

Numerous extensions and modifications have been explored to enhance the BEM and address its inherent limitations. To navigate through the correspondence problem, various strategies have been employed. These include assigning a probability to each potential match between left and right retinal images [30], introducing suppressive subunits arranged in a push-pull organisation [31], and developing a hybrid model that encompasses both position and phase disparity [32,33]. Efforts to modify the BEM to attenuate responses to anti-correlated stimuli have included the incorporation of a monocular simple cell, applying non-linearities at both the monocular and binocular levels [34], and introducing second-order filters that utilise monocular energy responses as input [2]. However, achieving a biologically plausible model that replicates the attenuation effects observed in odd-symmetric neurons has proven to be a complex challenge [6].

2.4 Convolutional Neural Networks

CNNs are specialised neural networks designed for processing grid-like data, notably images, and have been instrumental in image analysis and feature extraction. Essentially, CNNs consist of convolutional layers that apply filters to input images and may contain pooling layers that reduce spatial dimensions, and fully connected layers that typically perform classification. Activation functions, such as the Rectified Linear Unit (ReLU), introduce non-linearity, enabling the network to learn complex patterns efficiently [35]. In depth perception research, CNNs have been utilised for various applications, including extracting depth from RDSs and simulating cyclopean perception, showcasing potential uses in computer vision and robotics [36]. Furthermore, data augmentation techniques like Generic Data Augmentation (GDA) can be employed to enhance model robustness and mitigate overfitting during training [37]. With their adept image processing and feature extraction

capabilities, CNNs offer a promising path for ongoing research in visual perception and related applications.

2.5 Research Gaps

The HVS has undeniably been at the forefront of numerous studies and computational model developments, as evidenced throughout this literature review. Although RDSs have been utilised in various visual studies, particularly for their ability to singularly isolate binocular depth perception, a discernible gap is observed in the exploration of training a biologically plausible CNN using RDSs to unearth novel insights into the HVS. Thus, the ensuing research, which ventures into this unexplored intersection, establishes its novelty in the field.

3. Methods

3.1 RDS Generation

RDSs were generated to serve as the primary data for training and evaluating the CNN. Two types of RDSs were utilised: correlated and anti-correlated.

3.1.1 Correlated RDS

Correlated RDSs were generated by manipulating several parameters, each playing a crucial role in defining the characteristics of the stereograms. Table 1 outlines the parameters considered for the generation of a correlated RDS.

Table 1 Outline of parameters considered for RDS generation, along with a description and their unit

Parameter	Definition	Unit
Image Size	Specifies the width and height of the RDS in pixels. It determines the overall spatial resolution and size of the generated stereogram.	pixels
Figure Width	Defines the width of the disparity region (the figure) within the RDS. It affects the perceptual prominence and visibility of the figure against the background.	pixels
Disparity Shift	Determines the amount by which the figure is shifted to create binocular disparity. It directly influences the perceived depth of the figure in the stereogram.	pixels

During the generation of correlated RDSs, the corresponding disparity map was also produced, calculating disparity as the disparity shift relative to the original random dot pattern. The RDSs served as inputs for the CNN, while the disparity map provided a reference to evaluate the model’s output using Mean Squared Error (MSE). An example of a correlated RDS and its disparity map is depicted in Figure 6.

3.1.2 Anti-correlated RDS

In contrast to merely shifting the figure relative to the other image as with a correlated RDS, an anti-correlated RDS also inverts the pixels within the figure, thereby creating an anti-correlation. It is generated with the same parameters as the correlated RDS seen in

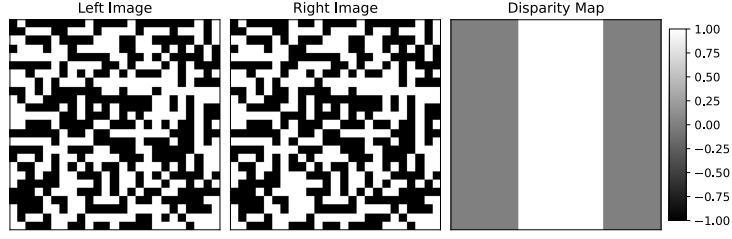


Figure 6 Example of a 25x25 pixel correlated RDS alongside its corresponding disparity map. The left and right images act as CNN inputs, while the disparity map, emphasizing the figure with a value of 1 due to its 1-pixel disparity shift, is used to assess the model’s performance.

Table 1. Figure 7 presents an example of an anti-correlated RDS alongside its disparity map.

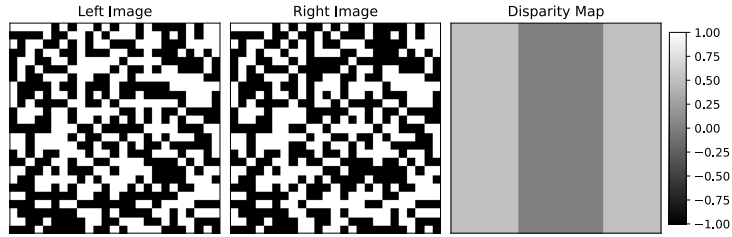


Figure 7 Example of a 25x25 pixel anti-correlated RDS, featuring left and right images. Despite having a disparity of 1, it encodes values of 0.5 in the no disparity region and 0 in the disparity area, intending to prompt the model to reduce its output when faced with an anti-correlated RDS.

Although the RDS itself possesses a disparity of 1, the disparity map encodes 0 in the disparity region. While anti-correlated RDSs do not drive a clear perception of depth, they do activate disparity-tuned neurons in the visual cortex [6]. The choice for a value of 0.5 in the no disparity region was in order to mimic the attenuation output of response noted from cortical neurons [6]. Section 2.2 explores the literature surrounding the inversion of the response to anti-correlated RDSs, which was the driving choice for the inverted disparity map for anti-correlated RDSs.

3.2 CNN Architecture

Drawing substantial influence from the BEM, detailed in Section 2.3.1, the CNN architecture was devised to harness modern computing techniques.

3.2.1 Overall Structure and Workflow

The CNN, depicted in Figure 8 and 9, processes left and right images of an RDS through respective input channels, utilising subunits to perform mathematical operations while preserving the input dimensions in the output. The final model output, a predicted disparity map, is formulated by summing the outputs from all subunits. The model was set up such that the number of subunits can be modified. The model’s performance is quantitatively measured and optimised by computing the MSE between the model’s output and the ideal disparity map. While accuracy is a consideration, the primary emphasis is placed on model analysis over exclusively high-accuracy attainment. The Stochastic Gradient Descent (SGD) optimiser, chosen for its ability to produce the lowest final MSE compared to alternatives like the ‘Adam’ algorithm, ensures efficient model convergence during training and the achievement of a stable, minimal loss.

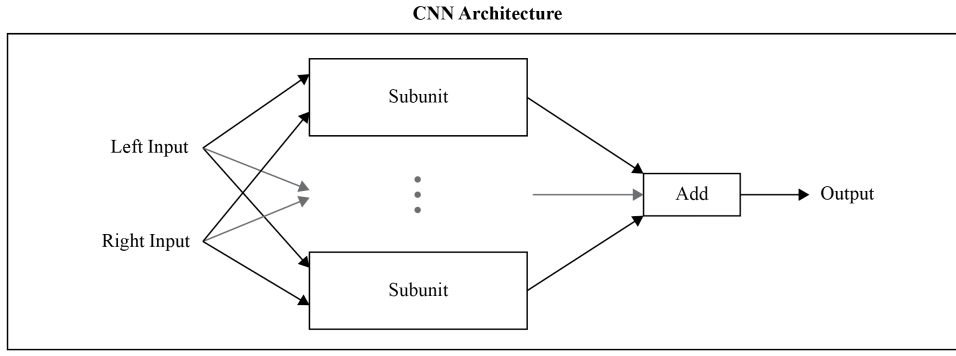


Figure 8 A schematic of the general CNN architecture, illustrating that the left and right input are passed to every subunit. The outputs of subunits are then added together to produce the predicted disparity map. The architecture allows modification of the number of subunits as needed.

3.2.2 Subunit Structure and Functionality

Within the CNN architecture, each subunit encompasses two convolution layers (one for each input), an add layer, and an activation layer. Inputs are convolved within the subunit, and the resulting left and right convolved outputs are subsequently summed and activated. Figure 9 and Table 2 outline the architecture as well as the layers used in the subunit.

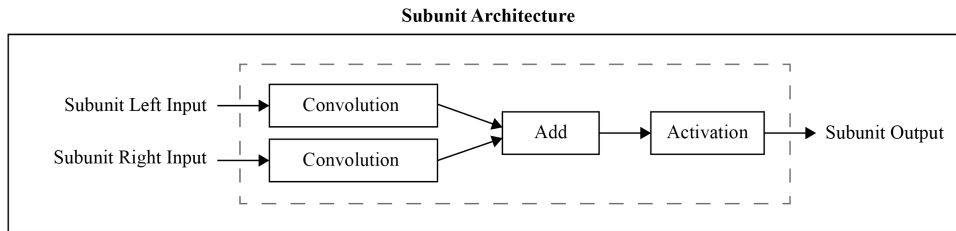


Figure 9 A schematic of a subunit's architecture. The subunit takes two inputs which undergo convolution, are then added together, and finally undergo activation. The dimension of the input images is retained between the input and the output.

Table 2 Outline of layers in the subunit, along with their description.

Layer	Description
Convolution	Utilises a 1x11 kernel to ensure full encapsulation of the figure during convolution and employs zero-padding to preserve spatial dimensions. The random uniform initialiser was chosen for its minimal final MSE in preliminary tests. Each convolution layer has 11 learnable parameters (the weights of the filter). Bias was not used in this layer, to ensure alignment with the BEM and to ensure biologically plausible filters. Overall, each subunit has 22 trainable parameters.
Add	Combines the outputs from the left and right convolution layers, summing them element-wise to produce a single output that retains the spatial dimensions. This step is crucial for merging information from both input channels.
Activation	Involves half-wave rectification followed by squaring, aligning with the BEM and ensuring the model can emulate the non-linear response of a simple cell effectively.

The CNN model was constructed using the Keras API, integrated within TensorFlow. The selection of TensorFlow over alternative prominent machine learning libraries was dictated

by its inherent flexibility and adaptability, essential due to the non-standard design of the CNN architecture [38].

3.3 Training

Utilising the NeSI high-performance computing cluster, all models underwent training, leveraging the computational capabilities of an NVIDIA A100 GPU. The incorporation of NVIDIA's CUDA and cuDNN libraries significantly expedited the training process, capitalising on the GPU's abundant cores and proficient multi-processing capabilities, thereby providing a pronounced advantage over CPU training in terms of speed and processing efficiency.

A substantial dataset, proportionally large relative to the count of learnable parameters (22 per subunit), was employed to mitigate the risk of overfitting [39]. Specifically, a total of 50,000 RDSs were designated for training purposes, complemented by an additional 10,000 reserved for validation throughout the training phase. The RDSs were characterised by a 25x25 pixel height and width and a 25x10 figure size. The dataset encompassed 25,000 correlated RDSs, each with a disparity of 1 (refer to Figure 6), and 25,000 anti-correlated RDSs, each with a disparity of 1 and an attenuated disparity map (refer to Figure 7). Notably, the disparity was exclusively applied to the right image.

The training was executed over 4 epochs, utilising a batch size of 8. Given that the model was trained singularly on a single stimulus type, which exclusively contained black or white pixels, and only had 22 trainable parameters per subunit, the classification task was relatively straightforward. Aiming to prioritise model analysis over mere accuracy attainment, a span of 4 epochs was deemed adequate to stabilise the model's loss and ensure an efficient training trajectory.

3.4 Model Analysis

3.4.1 Binocular Receptive Fields

To form a link with extant biological models and facilitate a comprehensive understanding of the model's filters and complex cell output, an analysis of the RFs was employed. This was accomplished by modifying the bar experiment, as delineated in [4], and is schematically represented in Figure 10. The experiment utilised a bar stimulus, defined as a 1 x 33 pixel image with a bar width of 1 pixel and populated with permutations of black (1) and white (-1) while maintaining all other pixels at a value of 0 to inhibit model activation in those regions.

A model modification was implemented, replacing convolution operations with the sum of the dot product to generate a single scalar output. The scalar outputs from the left and right were amalgamated, subjected to half-wave rectification, and squared, adhering to the CNN architecture. This produced a singular scalar output value, which was subsequently plotted on the RF graph, contingent on the shifts of the left and right bars. The modifications were implemented to accentuate the specific image regions that elicited responses from the filters. The methodology generated a 21 x 21 stimulus grid, from which the disparity tuning curve of the neuron was extracted by summing along lines of constant disparity (the positive diagonals). It is of note that this process can be executed at both the simple cell and complex cell output levels, and both are utilised in this report.

The derived binocular RF and disparity tuning curves provide insights into the model's response to various stimuli and its disparity tuning, respectively. This analysis not only sub-

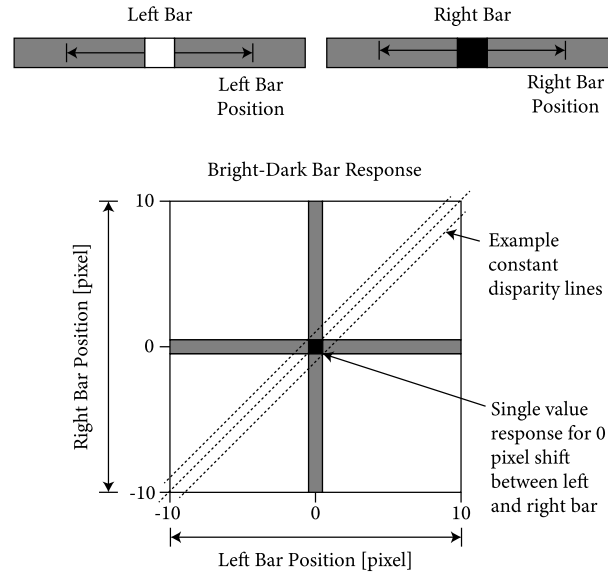


Figure 10 Approach for deriving binocular RF. Utilising a bar stimulus with independent positional shifts for the left and right bars, spanning from -10 to 10 pixels, a 21 x 21 stimulus grid was constructed (grid sizes can differ based on selected disparities). The four separate grids symbolise combinations: dark-dark, bright-bright, dark-bright, and bright-dark (black = -1, white = 1, grey = 0). Displayed is the bright-dark grid. Disparity tuning curves were formulated by summing each pixel along a constant disparity line, mirroring the relative disparity compared to the left bar. This produced 41 data points, of which the central 21 were used as they contain the most relevant data. The method is adapted from [4].

stantiates the model’s capability to extract depth information from RDSs but also provides a platform for comparing its performance and characteristics with biological systems, thereby bridging the computational model with biological plausibility.

3.4.2 Correspondence Problem

Navigating the intricacies of the correspondence problem is paramount in crafting models that aptly mirror the human visual system, and the model’s proficiency in discerning accurate disparities while negating incorrect matches within a visual context becomes a critical evaluation metric. The methodology employed to address this involves tuning the model to specific disparities by implementing a position shift between the left and right filters, with the magnitude of the shift determining the disparity to which the model is attuned. This approach facilitates a thorough evaluation of the model’s response to the correspondence problem by generating a series of models, each tuned to disparities ranging from -3 to 3, inclusive of 0, through shifting the right filter by a specified pixel amount while maintaining the filter length by appending zeros opposite to the shift direction. Figure 11 provides a visual representation of the filter-shifting methodology, illustrated on a filter from the original BEM with a 1x33 filter length.

Subjecting each disparity-tuned model to a consistent set of 100 RDSs, all possessing no disparity, a multi-step analysis process is employed. Initially, the output is normalised, and the model’s response is calculated as the mean pixel value. This is followed by a subsequent normalisation of all responses to standardise the scale across models, thereby mitigating convolution effects, and culminating in the computation of the mean and standard deviation of the 100 responses for each disparity-tuned model. A model that adeptly navigates the correspondence problem should exhibit a singular peak at the no disparity index, signalling optimal firing in response to the no disparity RDS and effective filtering of incorrect disparities. Conversely, a model that falters in addressing the correspondence problem

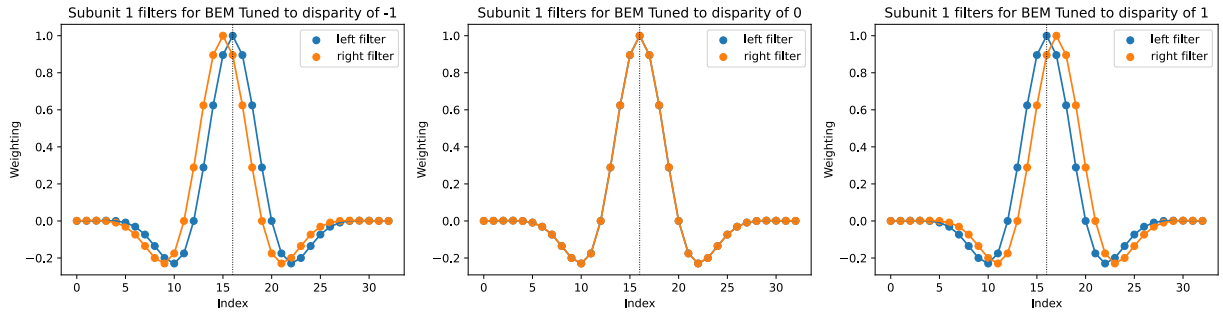


Figure 11 Method for adjusting filters to calibrate the BEM (with a 1x33 filter length) to different disparities. Although diagrams are shown for subunit 1, equivalent shifts are executed across all subunits to guarantee thorough model calibration.

might display a local maximum in the no disparity region, potentially alongside additional local maxima equal to or surpassing the no disparity response, indicative of challenges in filtering out erroneous disparities.

4. Results

4.1 Model Performance and Validation

a model with 4 subunits was successfully trained, demonstrating a capability to extract depth from RDSs by generating a predicted disparity map. Figure 12 illustrates the model’s adeptness in extracting correlated depth and attenuating anti-correlated RDS, as evidenced by the diminished magnitude response when compared to the correlated output. A detailed visualisation of a correlated RDS navigating through the trained model is provided in Figure A1.

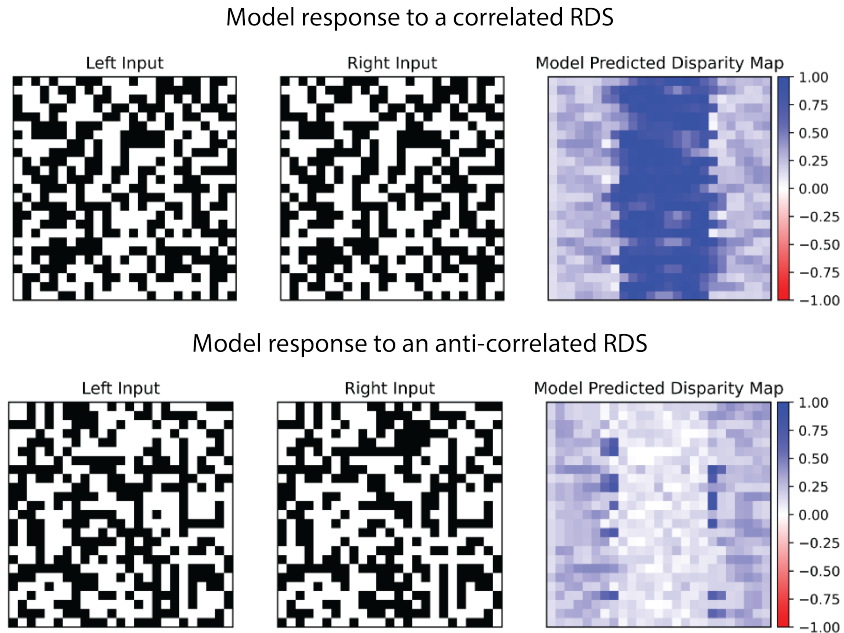


Figure 12 Predicted disparity maps corresponding to the provided 25x25 correlated and anti-correlated random dot stereograms, each featuring a 10-pixel-wide figure and a disparity of 1.

After 4 epochs and the stabilisation of the loss MSE, the model achieved a final MSE of 0.067, signifying a satisfactory alignment of its outputs with the desired depth maps. To further substantiate this performance, the standard deviation of MSE across varying figure

widths was tested to ensure that the model was not dependent on a specific figure width of 10. The model was tested across figure widths from 2 to 20, utilising 1000 correlated and anti-correlated RDSs each. The MSE remained consistent across all figure widths for both anti-correlated and correlated RDSs, with standard deviations of 0.0055 and 0.0130 respectively. This indicates the model is width-invariant and suitable for predicting disparity in various image and figure sizes.

4.2 Analysis of Filters and Subunits

4.2.1 Filter Characteristics and Deviations

The convolution filters, visualised in Figure 13, exhibited notable characteristics that differentiated them from the conventional BEM Gabor filters. When analysing the convolution filters, the major and minor extremums are used to determine the position shift between filters (the index offset between subsequent extremas) as well as if there is a phase shift by their orientation.

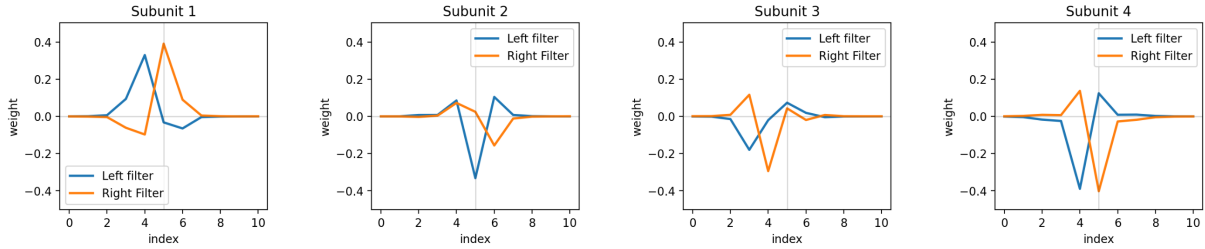


Figure 13 Visualisation of convolution filters for the 4-subunit model, each featuring a major and a minor extrema.

Analysing a single subunit's (such as subunit 4) filters in Figure 13 reveals that the major and minor extrema are offset by one index (representing a position shift) as well as flipped so that the minor extremum of the right filter has the same index as the major extremum of the left filter. This property is seen distinctly in subunits 1 and 4, and only partially in subunits 2 and 3.

4.2.2 Subunit Disparity Detection Mechanism

A comprehensive analysis of the subunits was conducted, utilising all possible bar permutations of black (B) and white (W): BB, BW, WB, and WW, with the bars being 1 pixel in width. Figure 14 provides an in-depth analysis, illustrating the impulse response of the filters within each subunit to different bar permutations without disparity, alongside the outputs from the bar experiment for each subunit.

The data revealed that the model trained certain subunits to be more responsive to specific bar permutations, as evidenced by some outputs being zero for certain permutations and non-zero for others.

Specifically, the following properties occur for each of the permutations:

- **Black, Black (BB):** Subunit 4 responds most strongly to the BB permutation. Due to both the left and right filter major extrema of subunit 4 being negative, the response to black (-1) produces a strong positive response. Subunits 2 and 3 also have a small response due to their negative major extrema. Subunit 1 has no response as both its major extrema are positive, resulting in a negative response, which is zeroed by the activation function.

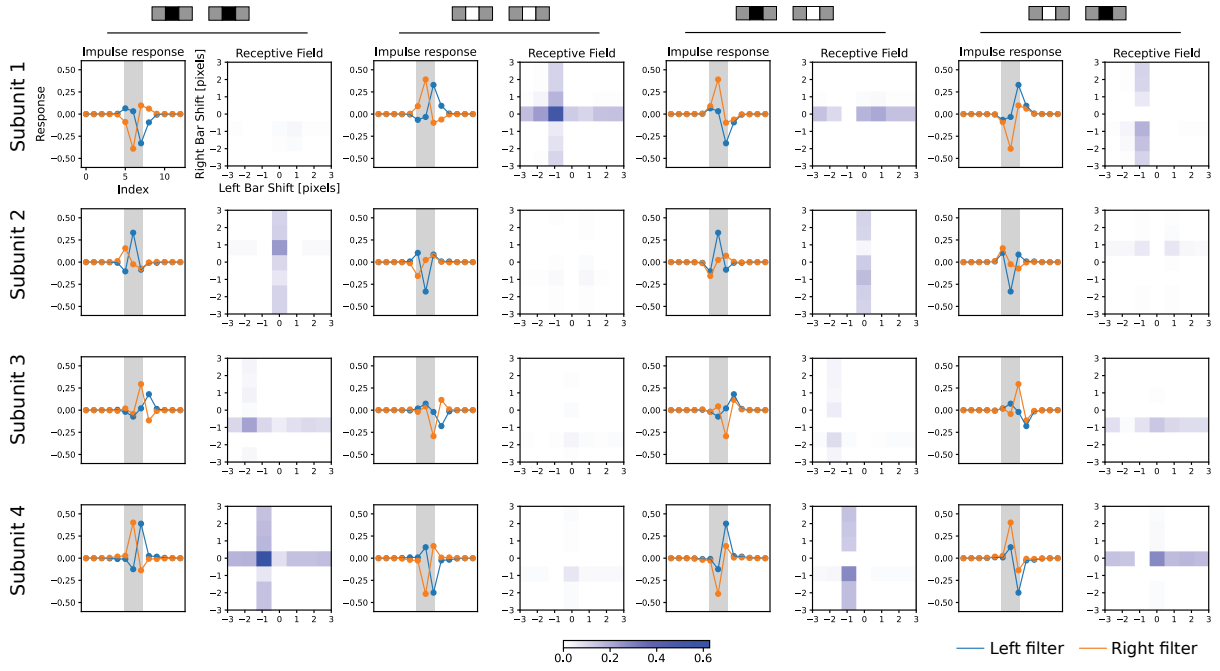


Figure 14 Plots of the eight filters within the model, showcasing the impulse response for each bar combination without disparity and the output from each subunit during the bar experiment, underscoring the receptive fields of each filter.

- **White, White (WW):** SSubunit 1 responds most strongly to the WW permutation. Due to both the left and right filter major extrema of subunit 1 being positive, the response to white (1) produces a strong positive response. Only subunit 1 responds due to being the only subunit with a positive major extrema. The rest are zeroed by the activation function.
- **Black, White (BW):** Subunits 1 and 4 respond most strongly to the BW permutation. This is due to the magnitude of their major extrema being significantly larger than their minor extrema, resulting in a stronger final response.
- **White, Black (WB):** Subunits 1 and 4 respond most strongly to the BW permutation. This is due to the magnitude of their major extrema being significantly larger than their minor extrema, resulting in a stronger final response.

The property of the minor extremum in both the left and right filters allows for attenuation. In all RF responses of the subunits that produce outputs for specific permutations, reduced response regions can be noted. The alignment of the major extrema produced the high-magnitude response for the specific disparity. The minor extrema reduces the response of the major extrema, allowing for attenuation.

4.2.3 Response to Correlated and Anti-correlated Disparities

The model exhibited a null response to anti-correlated disparities such as BW or WB, aligning with expectations. In a correlated RDS, the depth map within the disparity region presented non-zero values, whereas in an anti-correlated RDS, it was zero. This specific response pattern was achieved by having the magnitudes of the peaks for the left and right filters within a subunit be roughly equivalent. Two scenarios could arise that produced a minimal or zero output: when the major extremum was negative and the minor extremum was positive, which after half-wave rectification yielded an output of zero; or when the

minor extremum was negative and the major extremum was positive, their summation resulted in a small positive number, which when squared produced a near-zero output.

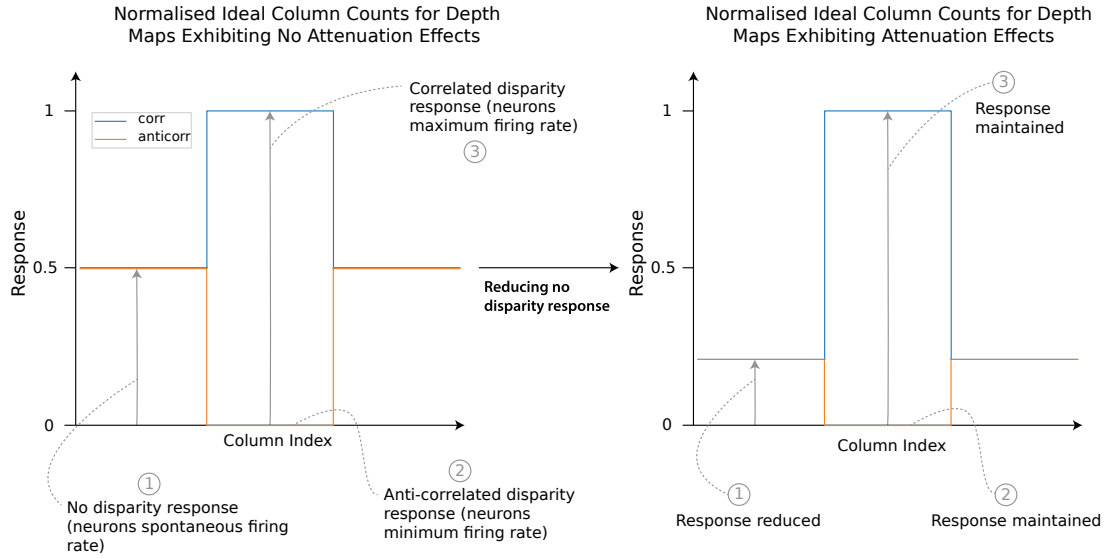


Figure 15 Illustration of the method to reduce the no disparity response while preserving the peak correlated disparity response and minimising the anti-correlated response, facilitating a mathematically attenuated response to anti-correlated RDS. The left graph represents the BEM’s response, while the right graph depicts the trained model’s response. Note that the response represents the column sum of an ideal disparity map.

The offset of 1 index between filter extrema indicates the model is tuned to a disparity of 1 pixel. This is reflected in the strong response RFs for the BB and WW permutations, where the peak response occurred at a left-right bar shift of $(-1, 0)$. Therefore, it indicates a tuning to a 1-pixel-correlated disparity shift. The same is seen for the anti-correlated region, whereby the 0 response regions at $(-1, 0)$ show the model is able to reduce the output of the anti-correlated disparity region when there is a 1-pixel shift between pixels.

4.3 Addressing the Correspondence Problem

The model’s proficiency in resolving the correspondence problem was evaluated, given the pivotal nature of this capability. The assessment involved juxtaposing the trained model with the binocular energy model, generating several models each tuned to disparities ranging from -3 to 3, inclusive of 0.

Figure 16 illustrates a stark contrast in the performance of the binocular energy model variants and the trained model in addressing the correspondence problem. Notably, the binocular energy model variants occasionally exhibit a larger output response from models not tuned to the true disparity, thereby failing to accurately resolve the correspondence problem.

A detailed exploration of the RFs and disparity tuning curves of both the BEM and the trained model is presented in Figure A2 and Figure 17. The RFs of the trained model, across all bar permutations, are markedly condensed compared to those of the BEM. This is manifested in the trained model’s pronounced response at specific coordinates for each permutation, whereas the BEM demonstrates a more dispersed response across various left and right bar coordinates.

Moreover, the disparity tuning curves of the trained model exhibit enhanced selectivity, showcasing a prominent peak at a single index with a sharp decline in response at neighbouring indices, thereby accentuating the model’s selectivity. Conversely, the BEM’s dis-

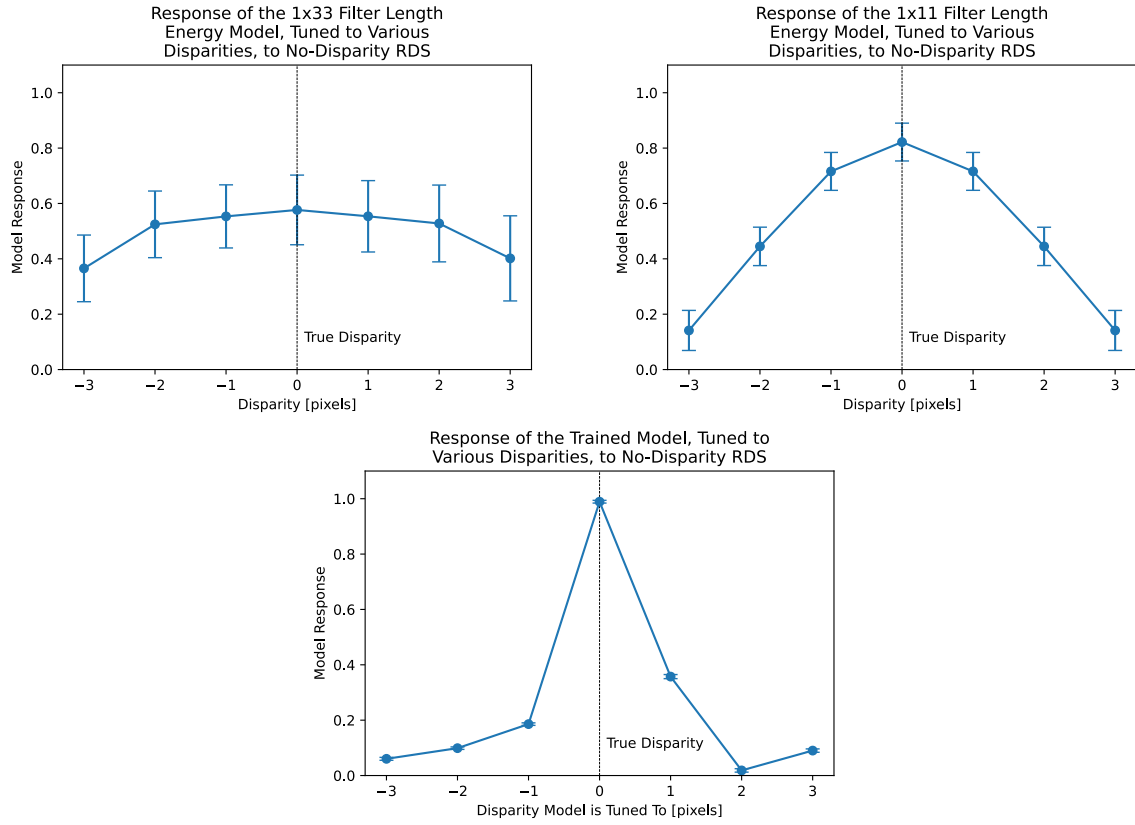


Figure 16 Comparative exploration of the BEM variants and the trained model in navigating the correspondence problem. Each model was exposed to the same 100 correlated RDS at each disparity, with the plots illustrating the mean and standard deviation of the responses.

parity tuning curves display a more gradual ascent to their peak, suggesting that substantial responses are retained at adjacent indices.

These distinct patterns underscore the trained model’s adeptness in addressing the correspondence problem, a challenge that proves to be a stumbling block for the BEM. The trained model, with its selective RFs and disparity tuning curves, demonstrates a heightened sensitivity to a singular disparity, offering diminished responses to alternative disparities. Consequently, it renders the strongest response to an RDS that aligns with its tuned disparity, while other disparities elicit weaker responses. In contrast, the BEM, with its gradual peak approach, indicates its potential to render notable responses even to RDSs with disparities that are off-tuned. This selectivity in the trained model underscores its efficacy in adeptly navigating the challenges posed by the correspondence problem, thereby enhancing its reliability and applicability in practical scenarios.

4.4 Effect Subunit Quantity Adjustment

Due to the performance characteristics of the 4-subunit model with regard to the potential underutilisation of subunits (seen in Section 4.2.2), three alternative models, each encompassing 1-3 subunits, were developed and subjected to the same training dataset as the original 4-subunit model. The goal of these models was to analyse the effect of reducing the number of subunits and their ability to continue to effectively extract depth from RDSs while maintaining the ability to attenuate anti-correlated disparities and address the correspondence problem.

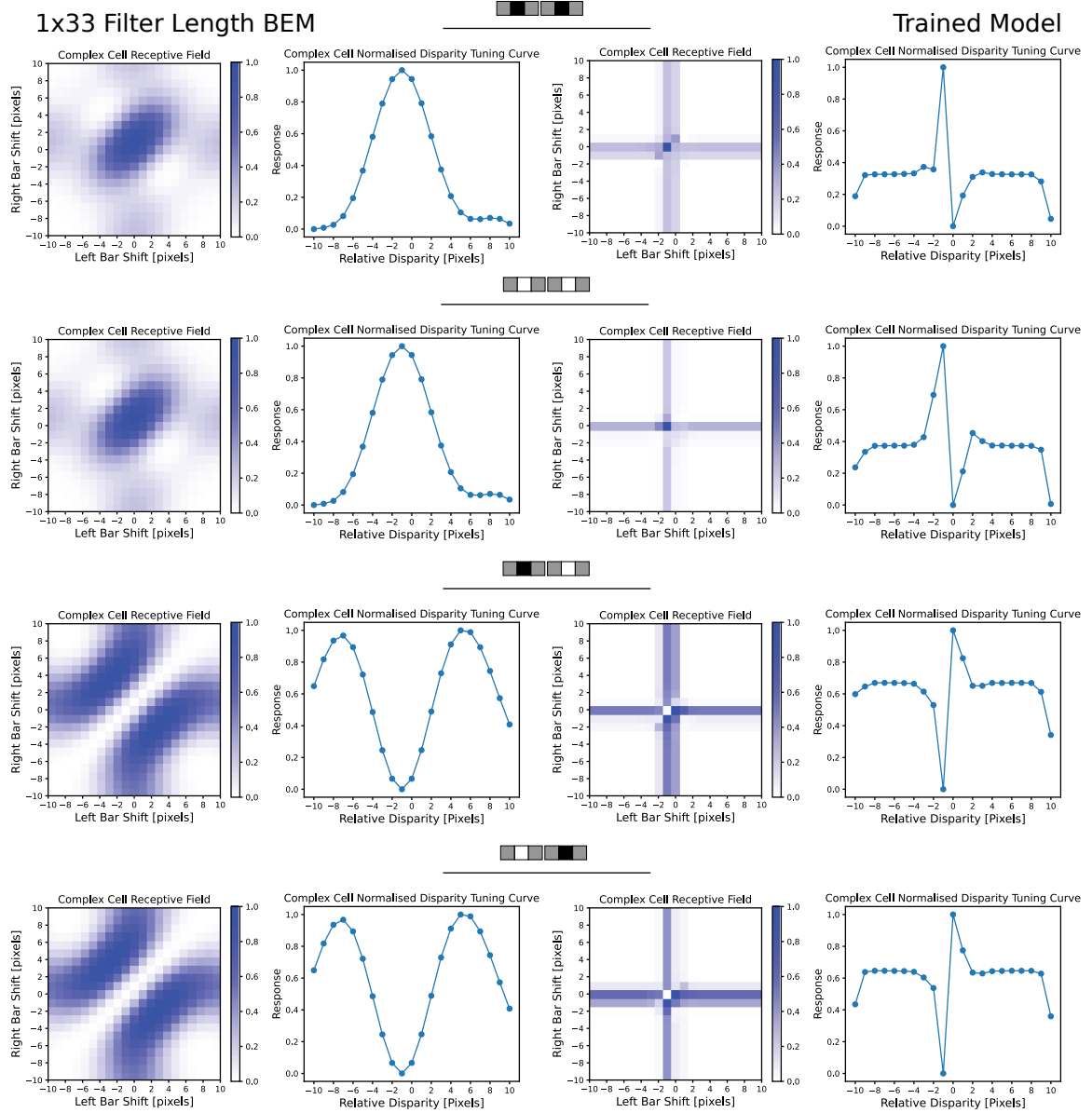


Figure 17 Depiction of RFs and the corresponding disparity tuning curves for both the BEM and the trained 4-subunit model, detailed for each bar permutation.

4.4.1 Comparative Performance

Analysing the models by their response to a correlated and anti-correlated 128x128 RDS with a 32-pixel-wide figure, as depicted in Figure 18, revealed differences in their ability to extract depth and attenuate response across all subunits. While all models demonstrated the capability to extract depth from correlated RDSs and attenuate anti-correlated disparities, the efficacy and precision of these operations varied. The 3-subunit model exhibited a response closely mirroring that of the 4-subunit model, suggesting redundancy in subunit 3 in the original model (reflected in its unresponsive RFs in Figure 14 across all bar permutations). The 2-subunit model, despite having half the subunits of the original, managed to produce a disparity region average magnitude of 89.2, compared to an average magnitude of 107.8 in the 4-subunit model, representing only a 16.8% decrease in responsiveness. Conversely, the 1-subunit model yielded a significantly lower response average magnitude of 44.3, marking a substantial 58.8% decrease compared to the 4-subunit model. Attenuated responses in the no-disparity region averaged a magnitude of approximately 35 for the 4-2 subunit models, while the 1-subunit model exhibited a markedly lower average magnitude response of 20.

Nevertheless, all models successfully reduced the response to anti-correlated disparities.

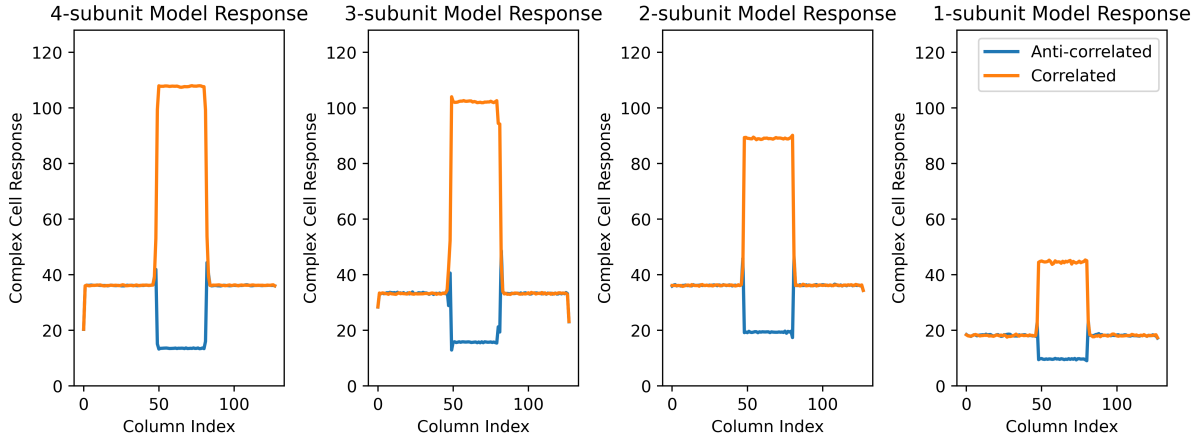


Figure 18 Response of models with a range of subunit counts (from 1 to 4) to identically correlated and anti-correlated RDS. To enhance the visualisation of attenuation effects, the response is displayed as the sum of all pixels in each column.

4.4.2 Efficacy and Limitations

The 1-subunit model, while able to detect disparity at the correlated disparity region and exhibit signs of attenuation, was markedly inferior in performance compared to its multi-subunit counterparts. Analysing the filters of the 1 and 2 subunit models in Figure 19 reveals that there are identical filters between the 1-subunit model and subunit 2 of the 2-subunit model, indicating the 1-subunit model has a sensitivity restricted to black pixels. This is reflected in Figure 14 and Figure 13 whereby subunit 4 has a strong response to the BB permutation and similar filters. This reveals an inability to extract depth from an RDS with no black pixels. In Figure 18 it can be noted that the 1-subunit model's response is half that of the 2-subunit model, which reflects that the 1-subunit model is only triggering for BB. On the other hand, the 2-subunit model demonstrates the ability to extract depth and attenuate anti-correlated disparities with a reduced subunit count, highlighting its potential as a computationally economical yet effective alternative.

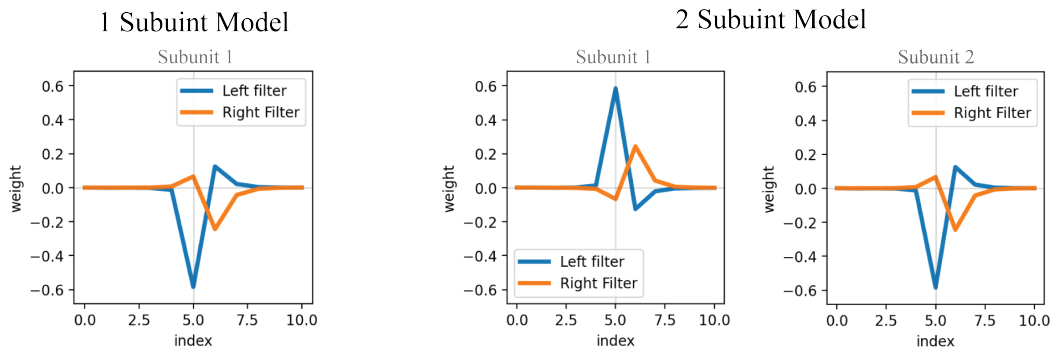


Figure 19 Visualisation of convolution filters for the 1 and 2-subunit models, demonstrating the 1-subunit model trained to have identical filters to one of the subunits in the 2-subunit model.

4.4.3 Characteristics and Functionality of Filters

Figure 20 reveals similarities in the filter characteristics between the 2-subunit model and the original 4-subunit model, even without explicit training to mimic such features. The filters within the subunits exhibited the same sign major extrema, sensitising them to specific pixel

permutations and enhancing their sensitivity to correlated disparities. The minor extremum of one filter aligns with the major extremum of its counterpart, attenuating the response to the no-disparity region to approximately half the magnitude of the correlated response. Notably, the major extrema of each subunit in the 2-subunit models is larger compared to the 4-subunit model (0.6 and 0.4 magnitude, respectively), necessitating a larger response in the bar experiment to achieve comparable results to the 4-subunit model, indicating a compensatory adaptation to the reduced subunit count. Finally, the major extrema of one of the filters in each of the subunits of the 2-subunit model has a distinctively larger magnitude than the other filter, allowing for attenuation of anti-correlated RDSs, whilst maintaining a strong response to disparities.

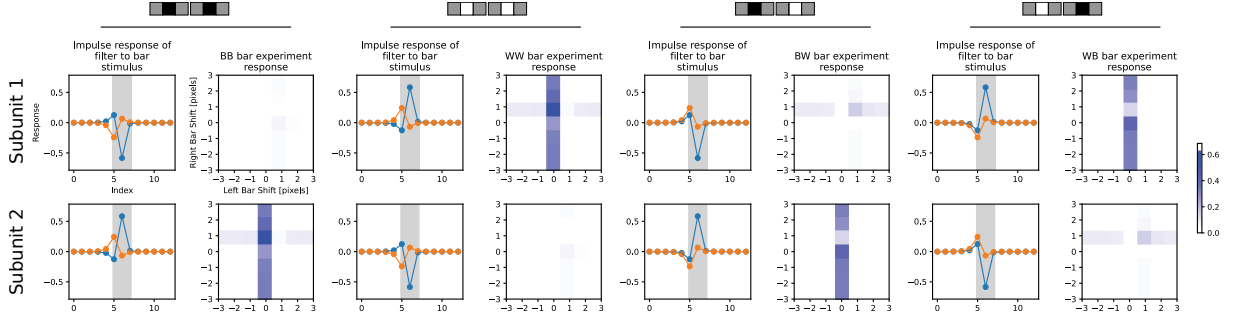


Figure 20 Plots of the 4 filters within the 2-subunit model, showcasing the impulse response for each bar combination without disparity and the output from each subunit during the bar experiment, underscoring the receptive fields of each filter.

5. Discussion

5.1 Model Behaviour and Disparity Encoding

Although the model was forced to attenuate through the reduced no disparity region of anti-correlated RDSs while minimising MSE, the filters were free to learn entirely by themselves, with the only restriction being the filter dimensions. It autonomously shaped its filters and demonstrated unexpected behaviours in subunit utilisation and disparity encoding. While subunits 1 and 4 emerged as essential, subunits 2 and 3 were less influential, diverging from expectations based on the BEM guidelines. This not only signals an opportunity for optimisation in computational resource allocation but also beckons a deeper dive into the biological implications. The model’s divergence from the BEM might hint at alternative neural encoding strategies or hierarchical organisations within the V1,

Disparity encoding, particularly for a value of 1, exhibited variability between configurations of different subunit numbers, revealing a sensitivity to spatial disparities that was dependent on the specific training data. For instance, the 4-subunit model detected disparity predominantly through a -1 left bar shift and a 0 right bar shift. Whereas the 2-subunit model detected disparity through a 0 left bar shift and 1 right bar shift. Although there was a marginal change in encoding mechanisms and spatial sensitivities, the models were both trained to detect a convergent disparity, whereby the figure appears in front on the no disparity region. This occurs when the shifted figure, irrespective of side, moves away from the centre, always creating the same relative disparity. So a disparity at (-1, 0) and (0, 1) is inherently the same. Thus, the variation is trivial, maintaining the focus on the originally detailed model and highlighting its convolutional nature, which prioritises the relative magnitude of disparity over its location.

Re-training the 2-subunit model on new training data led to the model detecting disparity through a -1 left bar shift and 0 right bar shift. While this reflected the 4-subunit model, the filter’s shapes remained the same; the only difference was a shift in the filter’s indexes. As explained earlier, this variation in training data, while interesting, provided no new insights and was therefore not explored in detail.

5.2 Biological Implications and Filter Characteristics

Navigating through three-dimensional spaces necessitates effective depth perception, a function significantly attributed to disparity-sensitive complex cells in the V1. The model, autonomously shaping its filters and demonstrating unexpected behaviours in subunit utilisation and disparity encoding, provides a compelling argument for a hierarchical organisation of these cells within the visual system. The necessity for complex cells to receive inputs from earlier stages, such as simple cells, underscores the specificity with which V1 neurons are attuned to particular disparities, not a broad range, which is vital for effective depth extraction and visual perception.

The model’s construction, particularly its utilisation of subunits, offers a potentially biologically plausible representation of the V1, albeit challenging the necessity of employing four subunits as suggested by the BEM. It effectively extracts depth from RDSs using only two subunits, each discerning either black or white pixels, hinting at a potentially simplified depth extraction mechanism within the V1, especially for tasks like the RDS, characterised by their binary pixel composition. This streamlined model might serve foundational roles in subsequent stages of the human visual system, such as the V2 and V3, continually refining depth features to formulate a more comprehensive depth map, building upon the foundational depth cues extracted from the V1.

While the trained model’s filters partially resemble the Gabor filters seen in the BEM, the mechanisms for disparity detection primarily rely on positional shifts, thereby supporting the concept of positional shift disparity encoding in the V1. Notably, phase shifts and a magnitude variation between the major extrema within a subunit are observed, introducing the possibility of a hybrid model in the V1. This model, integrating both positional and phase shifts between filters, might illuminate the model’s distinct ability to achieve attenuation without necessitating suppressive subunits, providing a unique approach to understanding the disparity detection and encoding mechanisms of V1 neurons. Moreover, the filters showcase smaller RFs and more selective disparity tuning curves, suggesting that V1 simple cells might possess similar features, which are key in addressing the correspondence problem and elucidating attenuation, especially since the major and minor extrema from corresponding filters need to align, facilitated by smaller RFs and more selective disparity tuning curves. This nuanced approach not only provides a step forward in comprehending depth perception and the role of V1 neurons but also opens avenues for further research into the intricate mechanisms underpinning visual perception and depth extraction in biological systems.

6. Conclusions

Navigating the complex mechanisms of the HVS and its adept depth perception capabilities, the model sheds light on potential biologically plausible mechanisms, highlighting the value of using CNNs to explore the HVS. The findings emphasise a probable hierarchical structure within the visual cortex, aligning with the BEM and its subunit proposition, while suggesting a potentially simplified depth mechanism within the HVS that operates with a minimum

of two subunits, as opposed to the previously suggested four. The model introduces the concept of a hybrid BEM within the V1, combining both positional and phase shifts to encode disparity and assist in attenuating responses to anti-correlated RDSs. Additionally, it proposes the existence of a magnitude difference between the major extrema of the filters within a subunit, which could potentially alter the ratios of the correlated and anti-correlated responses and assist in attenuation.

Furthermore, the model suggests that V1 simple cells might exhibit smaller RFs and more selective disparity tuning curves, crucial in addressing the correspondence problem and attenuating the model's output to anti-correlated RDS. While providing a foundational step towards understanding depth perception and the role of V1 neurons, the model opens avenues for further research into the complex mechanisms underpinning visual perception and depth extraction in biological systems. It not only provides a novel approach and insights into understanding the HVS but also establishes a foundation upon which future research might build, explore, and validate the biological implications of these findings, ensuring a comprehensive exploration of its architecture and functionality.

7. Future Work and Further Exploration

The progression of this project opens avenues for several extensions and investigations, as outlined below:

- **Acquisition of Psychophysical Data:** Engage in the collection of psychophysical data, utilising the correlated and anti-correlated RDSs. Employing this data through transfer learning could fine-tune convolution filters with a nuanced understanding of human perception.
- **Benchmarking Against the Human Vision System:** Implement a methodology to gather psychophysical data that mirrors the correspondence problem tests, enabling a comparison between the HVS and the model's proficiency in addressing the correspondence problem.
- **Diversifying Depth Stimuli Exploration:** Undertake training and evaluation of the model across a spectrum of depth stimuli to scrutinise the adaptability of the current filters and discern any alterations in filter behaviours amidst varied stimuli, including stimuli such as autostereograms.
- **Incorporation of Additional Visual Cues:** Investigate the model's response and adaptability to other visual cues, such as motion parallax and shading, to understand its robustness and applicability in more dynamic visual environments.
- **Adaptation to Various Visual Impairments:** Explore how the model performs or could be adapted to simulate depth perception in the presence of various visual impairments, providing insights into how these conditions impact visual processing.

References

- [1] I. Ohzawa, "Mechanisms of stereoscopic vision: the disparity energy model," Current opinion in neurobiology, vol. 8, no. 4, pp. 509–515, 1998.

- [2] J. M. Asher and P. B. Hibbard, “First- and second-order contributions to depth perception in anti-correlated random dot stereograms,” Scientific reports, vol. 8, no. 1, pp. 14 120–19, 2018.
- [3] I. Ohzawa, G. C. DeAngelis, and R. D. Freeman, “Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors,” Science (American Association for the Advancement of Science), vol. 249, no. 4972, pp. 1037–1041, 1990.
- [4] I. Ohzawa, G. C. Deangelis, and R. D. Freeman, “Encoding of binocular disparity by complex cells in the cat’s visual cortex,” Journal of neurophysiology, vol. 77, no. 6, pp. 2879–2909, 1997.
- [5] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, “Depth is encoded in the visual cortex by a specialized receptive field structure,” Nature (London), vol. 352, no. 6331, pp. 156–159, 1991.
- [6] S. Henriksen, S. Tanabe, and B. Cumming, “Disparity processing in primary visual cortex,” Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 371, no. 1697, 2016.
- [7] J. E. Sheedy, I. L. Bailey, M. Buri, and E. Bass, “Binocular vs. monocular task performance,” American journal of optometry and physiological optics, vol. 63, no. 10, pp. 839–846, 1986.
- [8] J. S. Lappin, “What is binocular disparity?” Frontiers in psychology, vol. 5, p. 870, 2014.
- [9] B. Julesz, “Binocular depth perception of computer-generated patterns,” Bell System Technical Journal, vol. 39, no. 5, pp. 1125–1162, 1960.
- [10] H. B. Barlow, C. Blakemore, and J. D. Pettigrew, “The neural mechanism of binocular depth discrimination,” The Journal of physiology, vol. 193, no. 2, pp. 327–342, 1967.
- [11] I. Ohzawa and R. D. Freeman, “The binocular organization of complex cells in the cat’s visual cortex,” Journal of neurophysiology, vol. 56, no. 1, pp. 243–259, 1986.
- [12] Y. Lian, A. Almasi, D. B. Grayden, T. Kameneva, A. N. Burkitt, and H. Meffin, “Learning receptive field properties of complex cells in v1,” PLoS computational biology, vol. 17, no. 3, 2021.
- [13] B.-E. Verhoef, R. Vogels, and P. Janssen, “Binocular depth processing in the ventral visual pathway,” Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 371, no. 1697, pp. 20 150 259–20 150 259, 2016.
- [14] D. H. Hubel and T. N. Wiesel, “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex,” The Journal of physiology, vol. 160, no. 1, pp. 106–154, 1962.
- [15] I. Ohzawa and R. D. Freeman, “The binocular organization of simple cells in the cat’s visual cortex,” Journal of neurophysiology, vol. 56, no. 1, pp. 221–242, 1986.
- [16] L. M. Martinez and J.-M. Alonso, “Complex receptive fields in primary visual cortex,” The Neuroscientist (Baltimore, Md.), vol. 9, no. 5, pp. 317–331, 2003.

- [17] Y.-D. Zhu and N. Qian, “Binocular receptive field models, disparity tuning, and characteristic disparity,” Neural computation, vol. 8, no. 8, pp. 1611–1641, 1996.
- [18] B. G. Cumming and A. J. Parker, “Responses of primary visual cortical neurons to binocular disparity without depth perception,” Nature (London), vol. 389, no. 6648, pp. 280–283, 1997.
- [19] A. J. O’Toole and D. J. Kersten, “Learning to see random-dot stereograms,” Perception (London), vol. 21, no. 2, pp. 227–243, 1992.
- [20] B. Julesz, “Stereoscopic vision,” Vision research (Oxford), vol. 26, no. 9, pp. 1601–1612, 1986.
- [21] M. S. Terrell and R. E. Terrell, “Behind the scenes of a random dot stereogram,” The American mathematical monthly, vol. 101, no. 8, pp. 715–, 1994.
- [22] T. Doi, S. Tanabe, and I. Fujita, “Matching and correlation computations in stereoscopic depth perception,” Journal of vision (Charlottesville, Va.), vol. 11, no. 3, pp. 1–1, 2011.
- [23] S. C. Aoki, H. M. Shiozaki, and I. Fujita, “A relative frame of reference underlies reversed depth perception in anticorrelated random-dot stereograms,” Journal of vision (Charlottesville, Va.), vol. 17, no. 12, pp. 17–17, 2017.
- [24] S. Tanabe, S. Yasuoka, and I. Fujita, “Disparity-energy signals in perceived stereoscopic depth,” Journal of Vision, vol. 8, no. 3, pp. 22–22, 03 2008.
- [25] T. Doi and I. Fujita, “Cross-matching: a modified cross-correlation underlying threshold energy model and match-based depth perception,” Frontiers in computational neuroscience, vol. 8, pp. 127–127, 2014.
- [26] P. B. Hibbard, K. C. Scott-Brown, E. C. Haigh, and M. Adrain, “Depth perception not found in human observers for static or dynamic anti-correlated random dot stereograms,” PloS one, vol. 9, no. 1, 2014.
- [27] A. I. Cogan, A. J. Lomakin, and A. F. Rossi, “Depth in anticorrelated stereograms: Effects of spatial density and interocular delay,” Vision research (Oxford), vol. 33, no. 14, pp. 1959–1975, 1993.
- [28] B. G. Cumming, S. E. Shapiro, and A. J. Parker, “Disparity detection in anticorrelated stereograms,” Perception, vol. 27, no. 11, pp. 1367–1377, 1998.
- [29] D. J. Fleet, H. Wagner, and D. J. Heeger, “Neural encoding of binocular disparity: Energy models, position shifts and phase shifts,” Vision research (Oxford), vol. 36, no. 12, pp. 1839–1857, 1996.
- [30] J. C. A. Read, “A bayesian approach to the stereo correspondence problem,” Neural computation, vol. 14, no. 6, pp. 1371–1392, 2002.
- [31] S. Tanabe, R. M. Haefner, and B. G. Cumming, “Suppressive mechanisms in monkey v1 help to solve the stereo correspondence problem,” Journal of Neuroscience, vol. 31, no. 22, pp. 8295–8305, 2011.
- [32] J. C. A. Read and B. G. Cumming, “Sensors for impossible stimuli may solve the stereo correspondence problem,” Nature neuroscience, vol. 10, no. 10, pp. 1322–1328, 2007.

- [33] R. M. Haefner and B. G. Cumming, “Article: Adaptation to natural binocular disparities in primate v1 explained by a generalized energy model,” Neuron (Cambridge, Mass.), vol. 57, no. 1, pp. 147–158, 2008.
- [34] J. C. READ, A. J. PARKER, and B. G. CUMMING, “A simple model accounts for the response of disparity-tuned v1 neurons to anticorrelated images,” Visual neuroscience, vol. 19, no. 6, pp. 735–753, 2002.
- [35] W. Hao, W. Yizhou, L. Yaqin, and S. Zhili, “The role of activation function in cnn,” 12 2020, pp. 429–432.
- [36] A. G. RADVÁNYI, “Spatial depth extraction using random stereograms in analogic cnn framework,” International journal of circuit theory and applications, vol. 24, no. 1, pp. 69–92, 1996.
- [37] L. Taylor and G. Nitschke, “Improving deep learning using generic data augmentation,” CoRR, vol. abs/1708.06020, 2017.
- [38] O.-C. Novac, M. C. Chirodea, C. M. Novac, N. Bizon, M. Oproescu, O. P. Stan, and C. E. Gordan, “Analysis of the application efficiency of tensorflow and pytorch in convolutional neural network,” Sensors (Basel, Switzerland), vol. 22, no. 22, pp. 8872–, 2022.
- [39] X. Ying, “An overview of overfitting and its solutions,” Journal of Physics: Conference Series, vol. 1168, no. 2, feb 2019.

Appendix A Supporting Figures

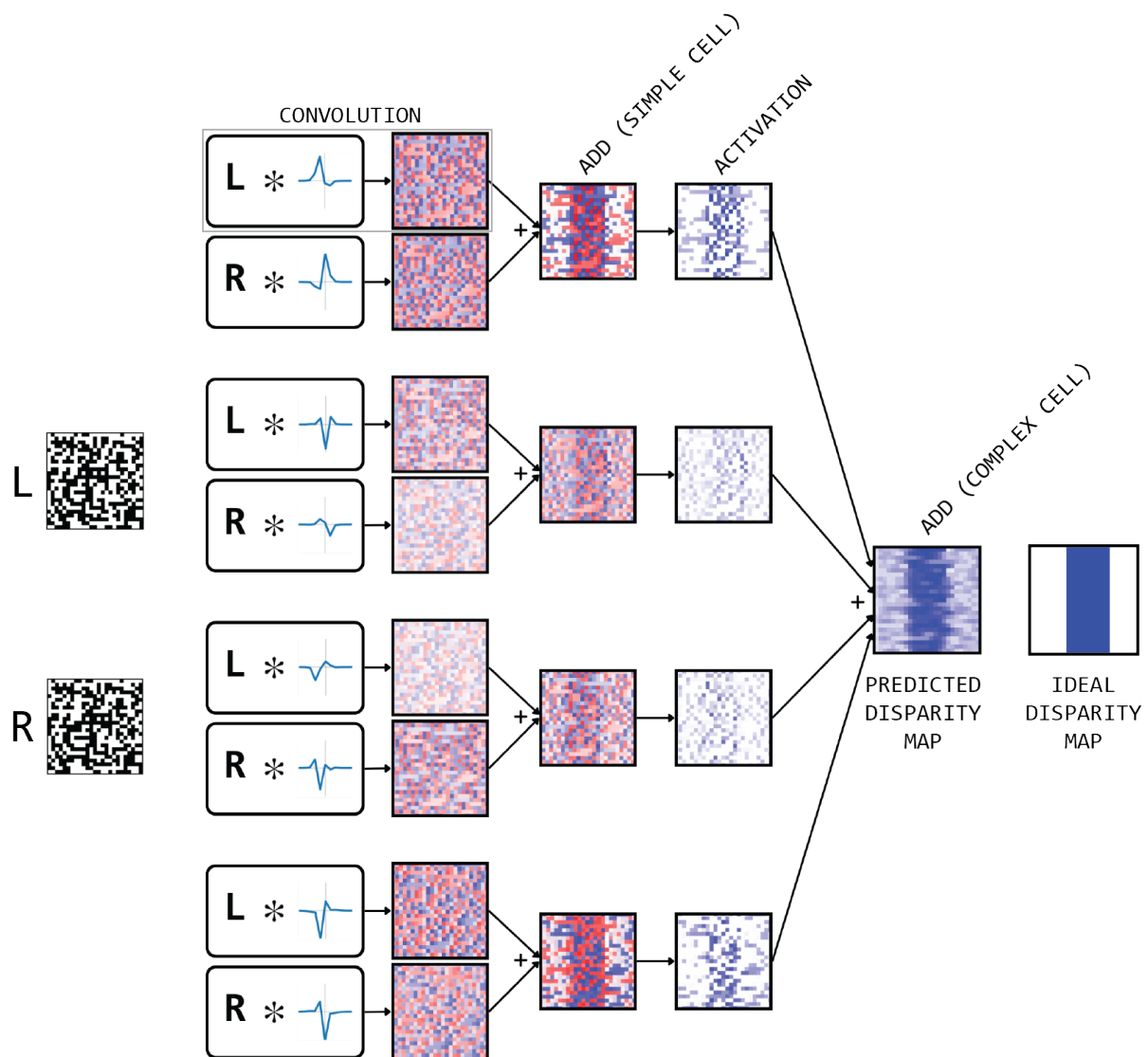


Figure A1 Visualisation of 25x25 correlated RDS with a disparity of 1 being passed through the trained model.

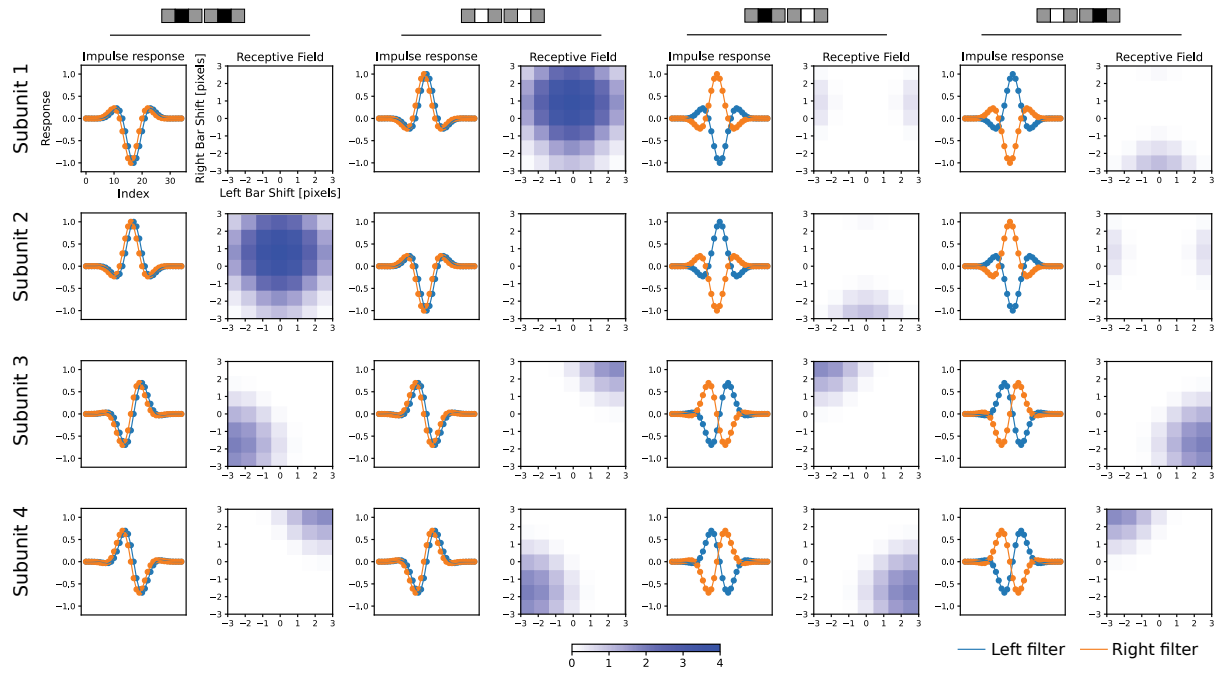


Figure A2 Full analysis of the BEM subunits.