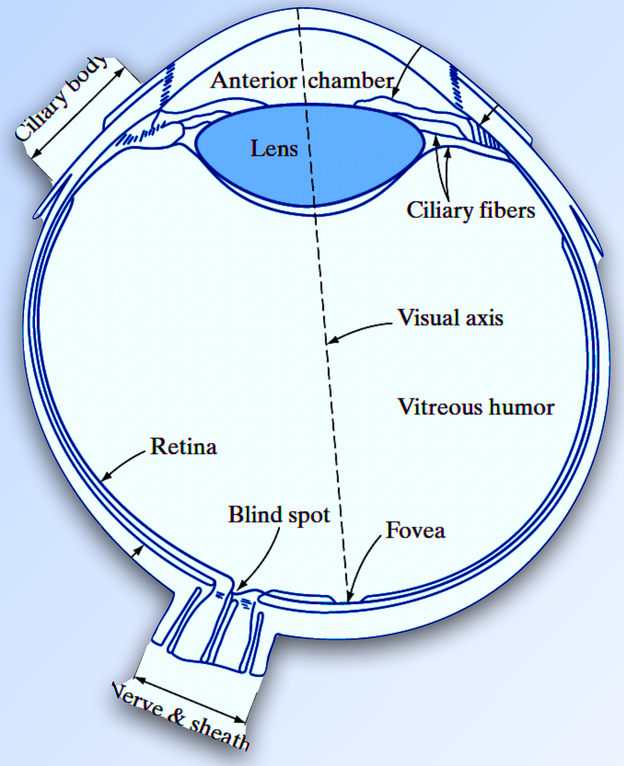
AI61201: Visual Computing With AI/ML

Module 2: The Human Visual System and Modeling Visual Perception

Dr. Somdyuti Paul

Structure of the Human Eye

- Humans are the most common consumers of digital visual data.
- Understanding of the human visual system is thus essential for the efficient design of visual computing frameworks.
- Processing of the visual stimulus starts at the human eyes and ends at the primary visual cortex (a region in the temporal occipital lobe of the brain).
- The human eye is a complex optical system that is functionally similar to a camera
  - Pupil - controls the amount of light entering the eye.
  - Lens and Cornea: focuses the light
  - Retina - captures the image using photoreceptors that convert light into an electrical signal.



Cross section of the human eye

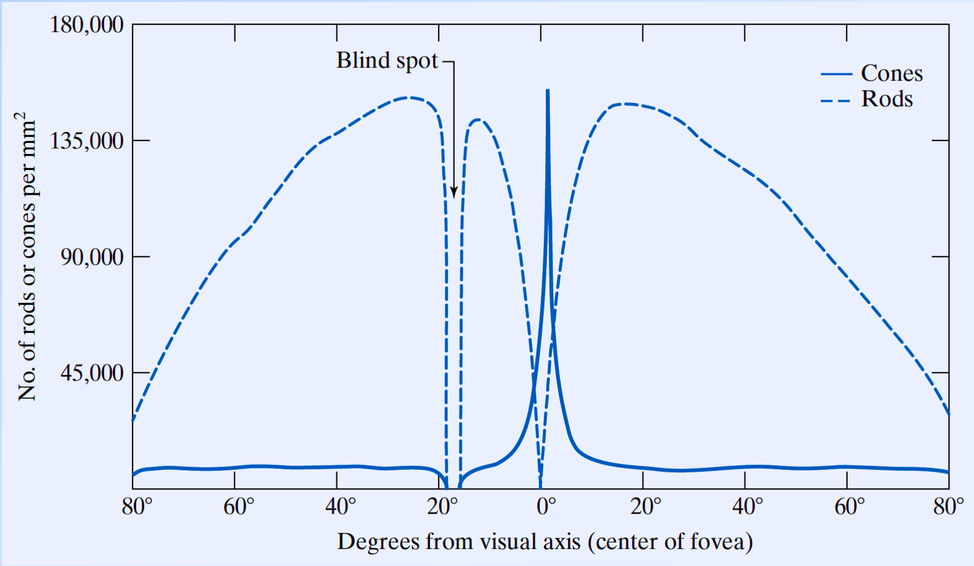
Types of Photoreceptors

- The retina consists of two distinct types of photoreceptors:
  - Rods: responsible for scotopic (dim-light) vision, and are not involved in the color perception
  - Cones: responsible for photopic (bright light) vision, and are primarily responsible for color perception

- The distribution of the rods and cones on the surface of the retina is highly non-uniform

- There are 6-7 million cones, primarily concentrated around the fovea, where the image is perceived at the finest detail.

- There are 75-150 million rods distributed over the surface of the retina, which encompasses a larger field of view, with some loss of details.



Distribution of rods and cones in the retina

Tristimulus Theory and Color Sensitivity

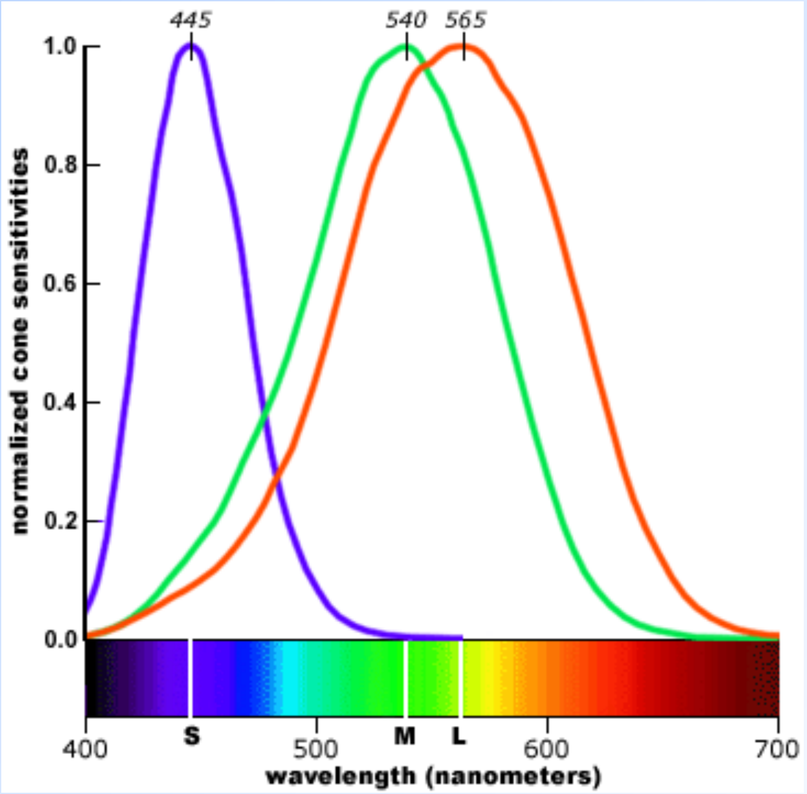
- There are three different types of cones which respond to three different ranges of wavelengths:
  - L: peak response around 564 nm, comprises 65% of all cone receptors
  - M: peak response occurs around 534 nm; comprises 33% of cone receptors.
  - S: peak response occurs around 420 nm, comprises just 2% of all cone receptors.

- The L, M and S cones have overlapping bandpass responses.

- The response of the cone receptors are the tristimulus values given by:

$$c_k = \int I(\lambda) m_k(\lambda) d\lambda \quad (k = L, M, S)$$

where $I(\lambda)$ is the spectral power as a function of wavelength (λ) and $m_k(\lambda)$ is the spectral sensitivity of the k^{th} cone sensor.



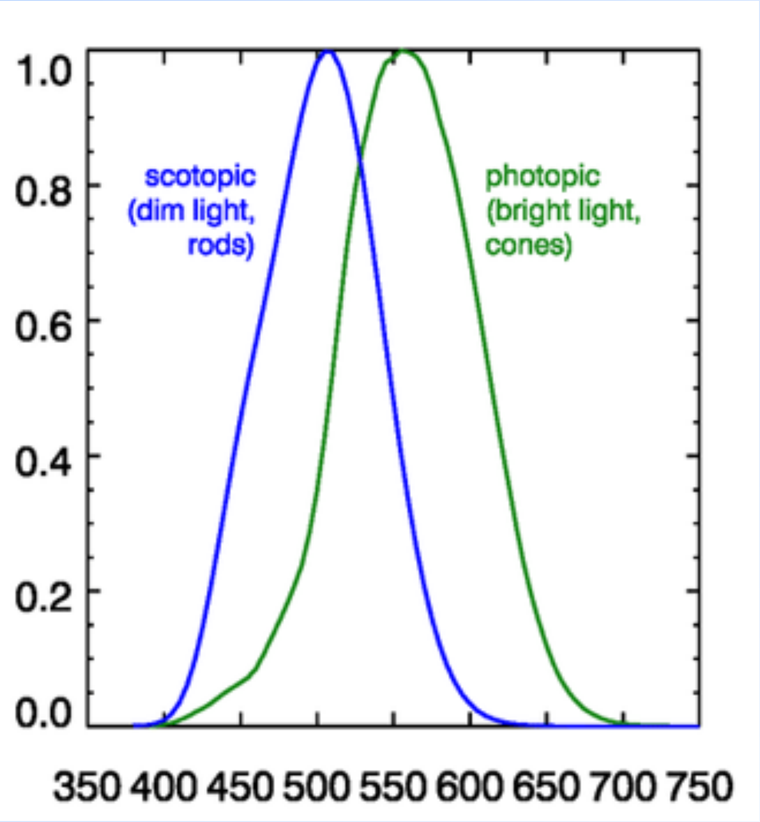
Cone sensitivity as a function of wavelength

Luminance Sensitivity

- For the same radiance, different wavelengths evoke different sensations of brightness in the HVS.
- Luminosity functions describe the perceived brightness at different wavelengths.
- The photopic and scotopic luminosity functions reveal that the perceived brightness is highest around 555 nm for photopic vision and around 507 nm for scotopic vision.
- The perceived luminance is given by:

$$y = k \int I(\lambda) l(\lambda) d\lambda$$

where $I(\lambda)$ is he spectral power as a function of wavelength (λ) and $l(\lambda)$ is the luminosity function.



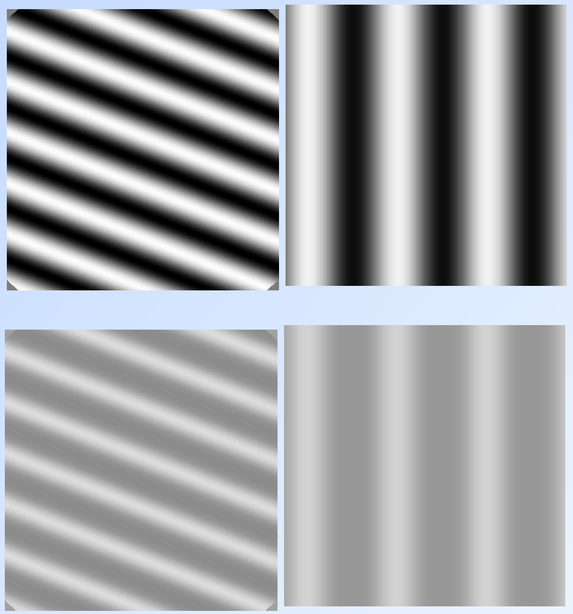
Luminosity functions for photopic and scotopic vision

Contrast Sensitivity and Spatial Frequency Response

- The HVS is more sensitive to contrast (i.e. change of luminance) rather than absolute luminance.
- The Michelson's contrast of a visual stimulus is given as follows:

  $C = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}}$, where I_{max} and I_{min} are the maximum and minimum luminance of the visual stimulus.
- Contrast sensitivity refers to the ability of the HVS to discern different luminance levels in a visual stimulus.
- The dependence of contrast sensitivity on spatial frequency was measured using sine wave gratings of varying spatial frequency and contrast (Campbell and Robson) as given by:

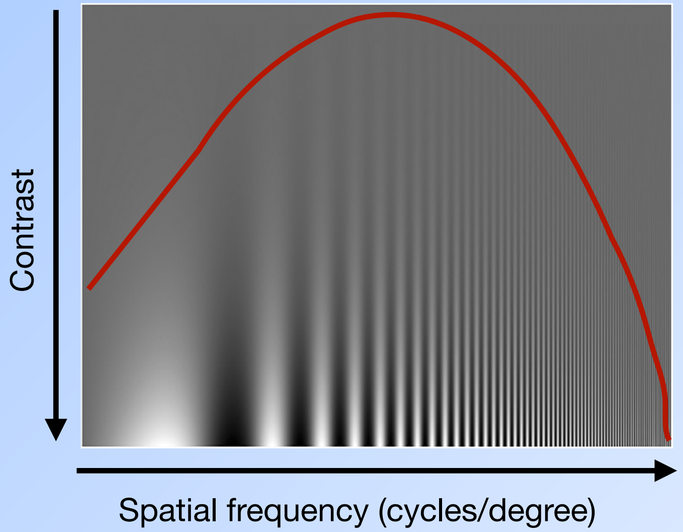
  $A\sin(Ux + Vy) + A_0 \ (0 < A < 1)$



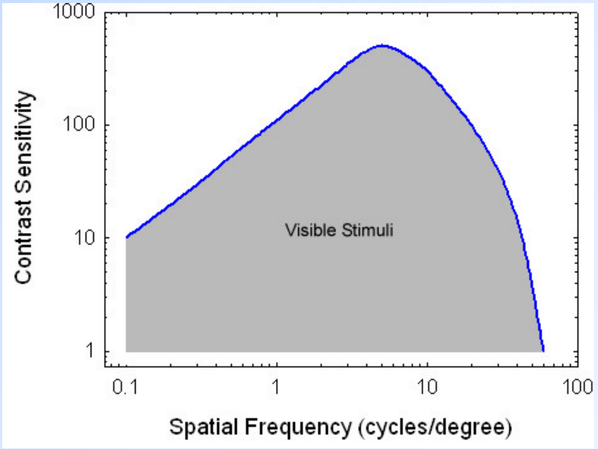
Sine wave gratings at different frequencies and contrasts

Contrast Sensitivity and Spatial Frequency Response

- For each spatial frequency, the minimum contrast at which the sine wave grating pattern becomes distinguishable was recorded as the contrast threshold for that frequency.
- The contrast sensitivity function (CSF) is the inverse of the contrast threshold as a function of frequency.
- The peak of the CSF occurs at a spatial frequency of about 4 cycles/degree
- Perceived contrast also depends on viewing distance.
- The CSF reveals that the HVS has a bandpass spatial frequency response.



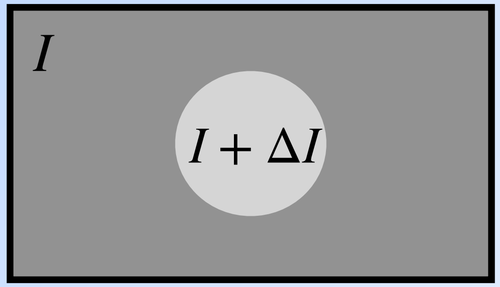
Grating at different contrasts and spatial frequency



Contrast sensitivity function

Luminance Masking

- Luminance masking refers to the effect whereby the perceived change in the intensity of an image patch over the background luminance is diminished as the background intensity increases.
- Consider an foreground area of intensity $I + \Delta I$ superimposed on an uniform background of area I, where ΔI is the minimum intensity difference between the foreground and the background at which the foreground becomes distinguishable from the background.
- According to Weber's law, the $\frac{\Delta I}{I} = K_w$, where K_w is a constant called the Weber ratio.
- For visual stimuli, Weber's law does not hold at very low and very high intensity levels.



Just Noticeable Difference

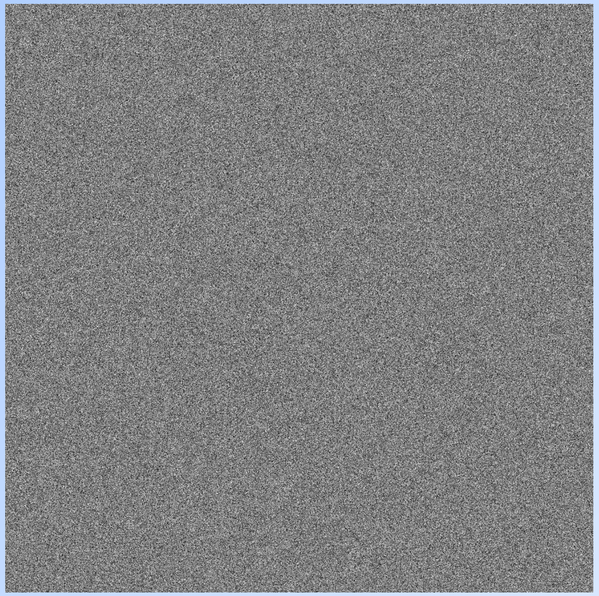
- The minimum difference between two (visual) stimuli necessary to perceptually differentiate between the two stimuli is called the Just Noticeable Difference or JND.
- In the definition of Weber's law, the JND is ΔI.
- JND measurements are often used to report the results of psycho-visual studies, and provides a means for quantifying the masking effects of the HVS.
- Since minimum visibility thresholds vary from subject to subject, multiple trials must be conducted to reliably measure JND values.
- Usually JND is recorded as the minimum change in the stimulus for which a certain fraction of the subjects (50% - 75%) can detect the change.

Contrast Masking

- The contrast masking property of the HVS refers to the reduced visibility of spatial patterns (such as distortions) when superimposed on regions that are spatially busy, i.e. rich in complex texture details.
- The visibility of image components is masked due by the presence of similar spatial frequencies in the vicinity.
- This contrast masking property is practically useful in the design of visual data compression and enhancement algorithms.

Image having low and high contrast regions

+



Gaussian noise

=

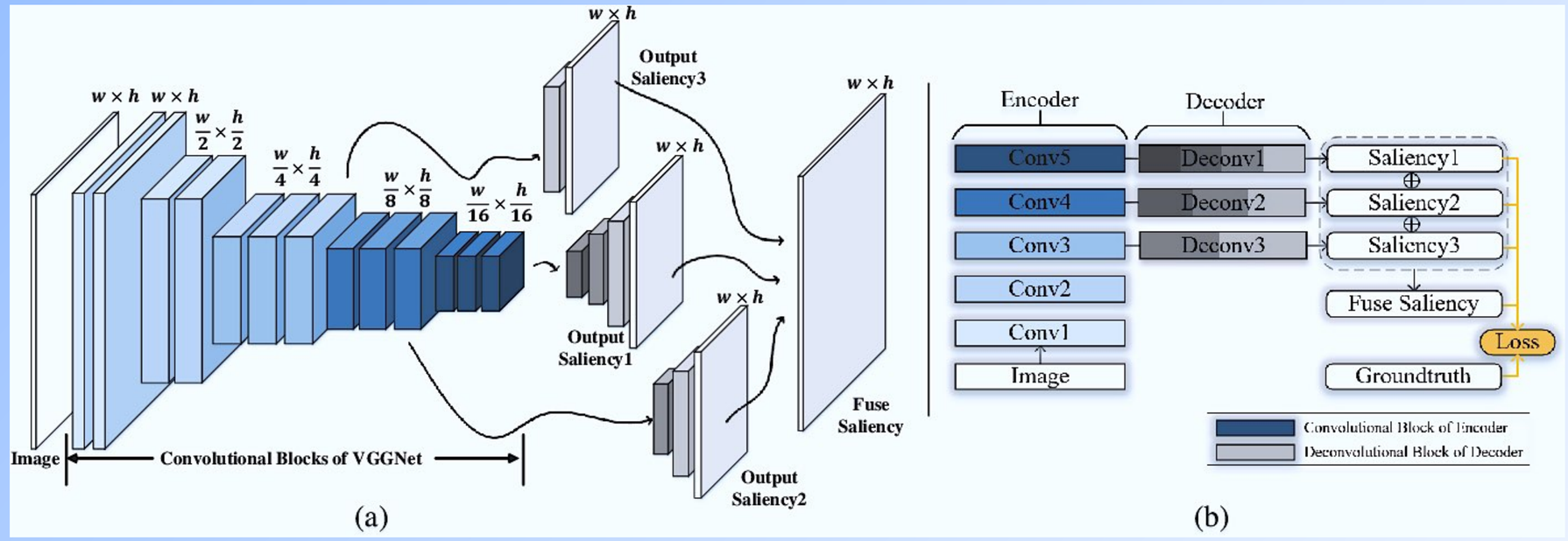
Noisy image

Visual Attention

- Visual attention refers to the process of focusing gaze on the most interesting or informative parts of a visual scene.
- Visual attention is enabled by eye movements
  - Saccades - rapid eye movements for fixation of gaze (spatial attention)
  - Pursuits - smooth eye movements to track a moving object (temporal attention)
- Visual attention is a dynamic research topic:
  - Machine learning models have been getting better and better at predicting visual attention.
  - Visual attention mechanisms usually enhance the performance of computer vision tasks.

Visual Attention Models

- Deep learning models have been successfully used to predict visual attention maps:
- *Deep Visual Attention Prediction (2018)* - Multi-level saliency was captured using shallow and deep features, and corresponding saliency maps were fused to obtain the saliency prediction.



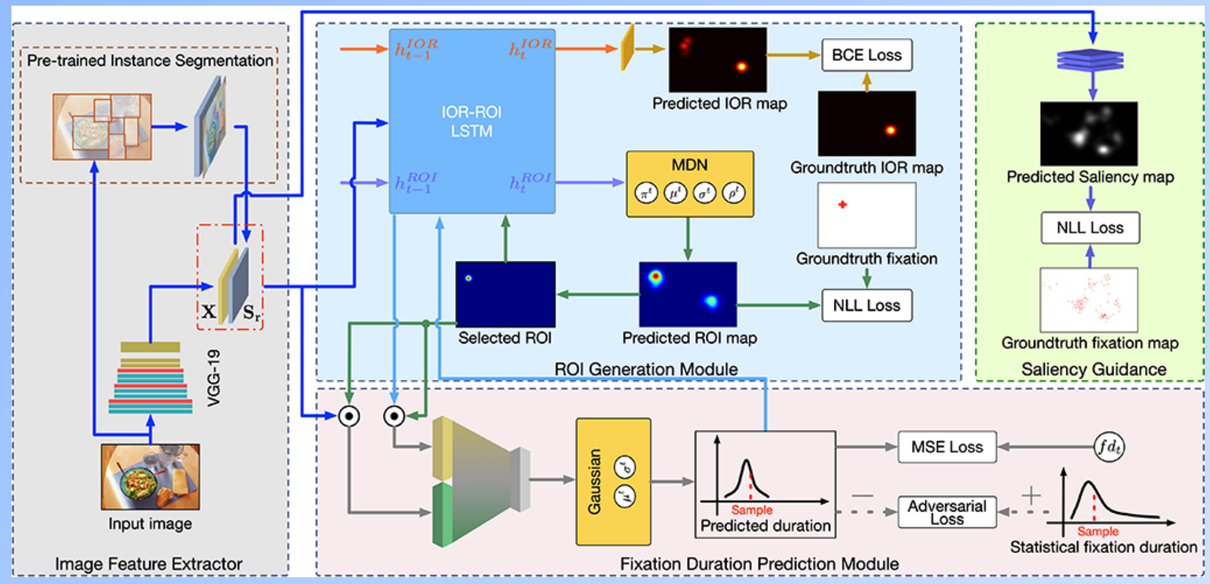
Deep Visual Attention Prediction

Visual Attention Models

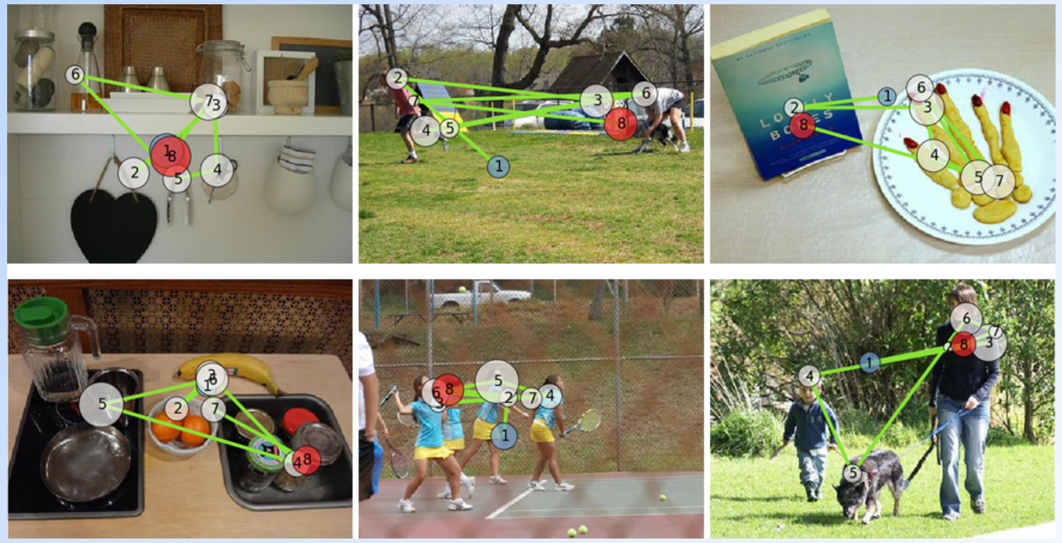
Deep Visual Attention Prediction Results

Visual Attention Models

- Deep learning has also been used to predict visual scan paths.
- *Visual Scanpath Prediction Using IOR-ROI Recurrent Mixture Density Network (2019)* - uses LSTMs to model the concepts of inhibition of return (IOR) and region of interest (ROI), while a mixture density network is used to model the next fixation distribution as a Gaussian mixture.



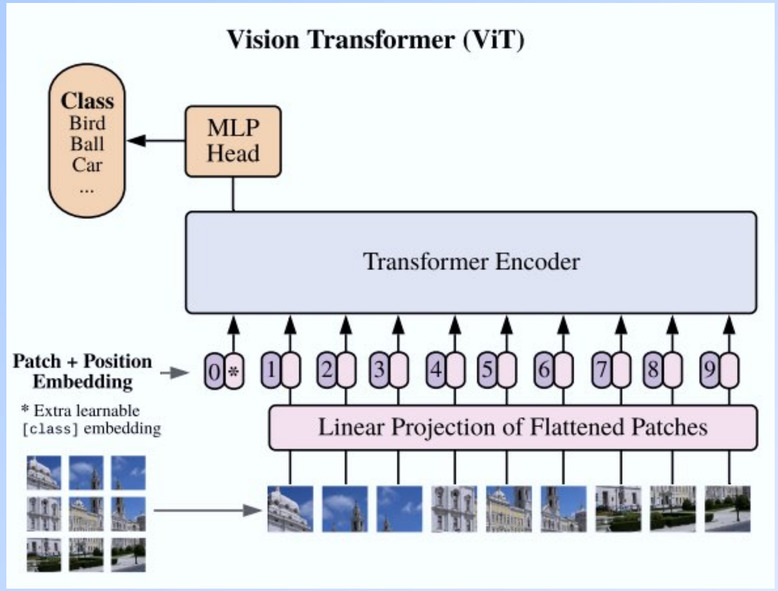
IOR-ROI Recurrent Mixture Density Network Architecture



Predicted visual scan paths

Visual Attention in Vision Tasks

- By augmenting neural networks with in-built visual attention mechanisms, notable performance gains have been reported in a range of vision tasks such as image classification, object detection, semantic segmentation etc.
- *Vision Transformer (ViT) (2020)* - simulates visual attention using self-attention and multi-head attention mechanisms.

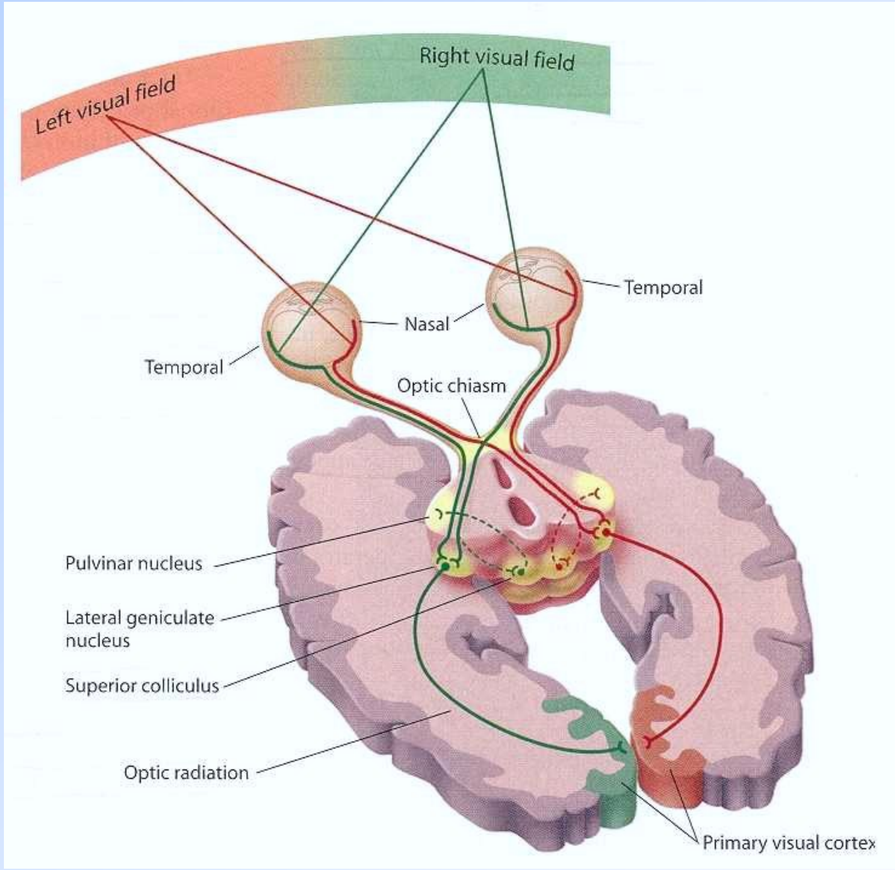


ViT Architecture

ViT Attention Maps

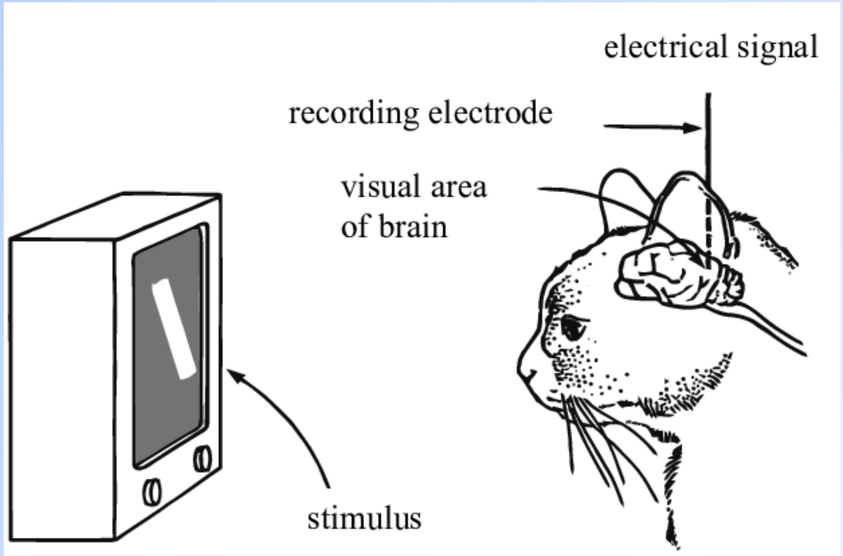
Visual Signal Processing in the Primary Visual Cortex

- The optic nerves from the two eyes cross over at the optic chiasm, where the information from both eyes are combined to reconstruct the left and right visual fields.
- The signal from both visual fields are transmitted to the opposite sides of the brain via the lateral geniculate nucleus (LGN), which acts as the relay station between the retina and the primary visual cortex
- The primary visual cortex or Area V1 is a region of the temporal occipital lobe where the the first stage of visual processing of the incoming visual signal is performed.



Visual Signal Processing in the Primary Visual Cortex

- Signals recorded from the primary visual cortex of cats in response to different visual stimuli were analyzed by Hubel and Wiesel (in 1959).
- Specific neurons in the primary visual cortex are sensitive to specific regions or the visual field.
- Neurons are sensitive to the orientation of objects in the visual field.
- Neurons are organized in layers:
  - Simple cells: responds to localized, oriented edges, Complex cell: aggregates responses of several simple cells.
- The primary visual cortex was an early inspiration for convolutional neural networks.



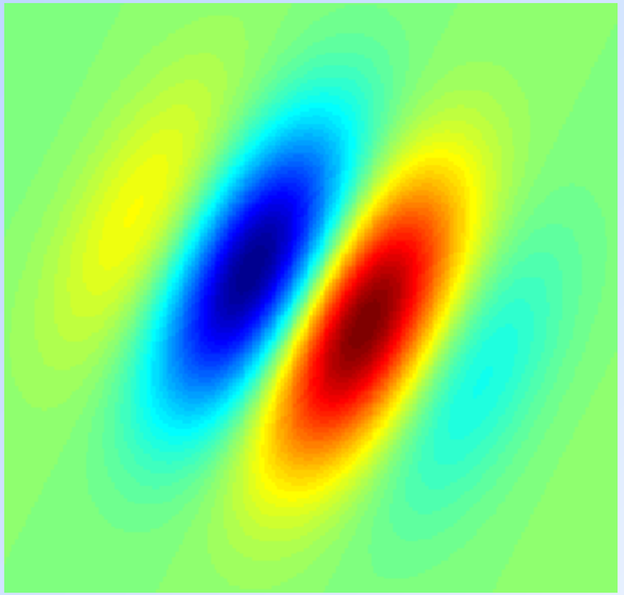
Hubel and Wiesel's experimental setup

Gabor Model For Simple Cells

- Simple cells act as oriented edge detectors and bar detectors.
- Simple cell responses have distinct excitatory and inhibitory regions, which could be modeled by Gabor filter responses.
- A Gabor filter is a linear filter, consisting of a sine wave modulated by a Gaussian envelope:

$$g(x, y) = \exp\left(-\frac{x^2 + \gamma^2 y^2}{2\sigma^2}\right) \cdot \exp\left(j\left(2\pi(ux + vy) + \phi\right)\right).$$

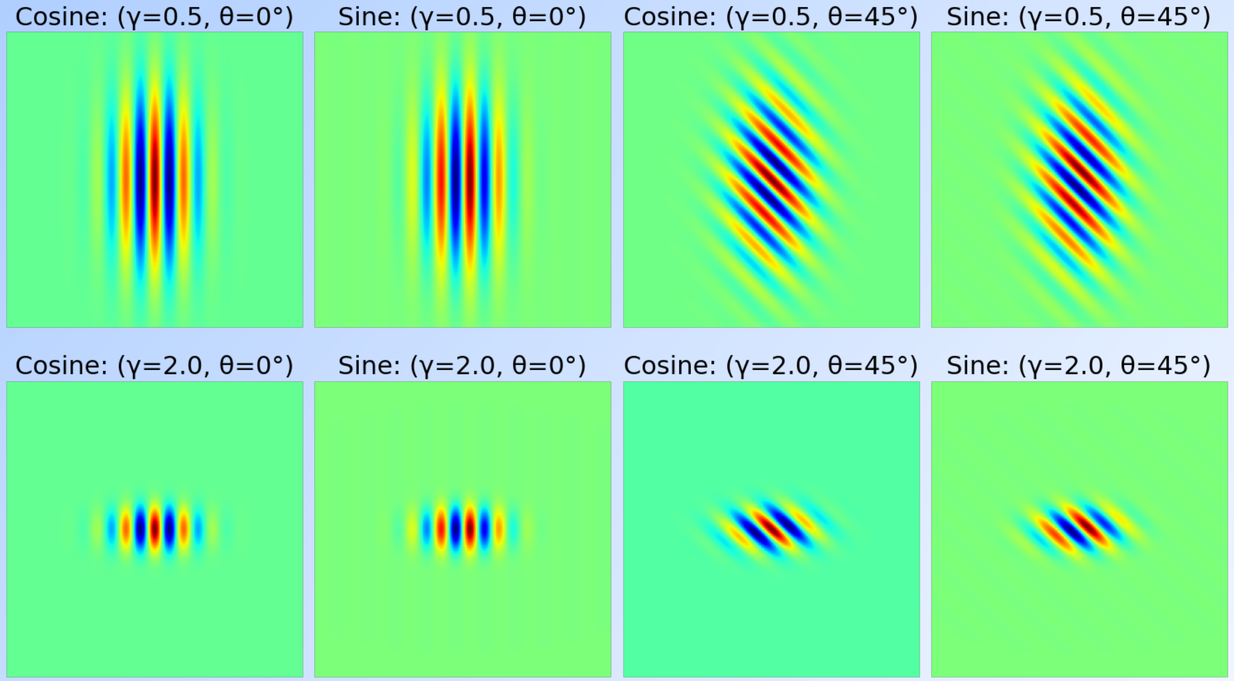
- The Gabor filter has the following parameters:
  - σ: standard deviation of the Gaussian envelope.
  - γ: aspect ratio of the Gaussian envelope.
  - u, v: spatial frequencies of the sine wave.
  - ϕ: phase offset



Excitatory and inhibitory response modeled by Gabor filers

Gabor Model For Simple Cells

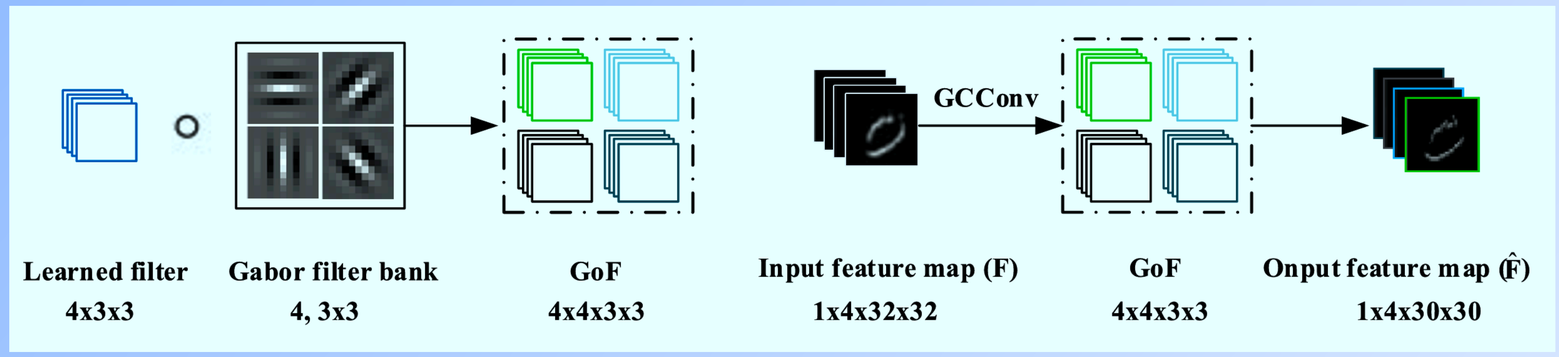
- The excitatory and inhibitory response of simple cells is modeled by the positive and negative responses of the Gabor filters.
- The spatially localized response of the simple cells is modeled by the Gaussian envelope of Gabor filters
- The orientation and frequency selectivity of the simple cells is modeled by adjusting the orientation and spatial frequency of the Gabor filters.
- The scale sensitivity of simple cells is modeled by adjusting the aspect ratio of the Gaussian envelope.



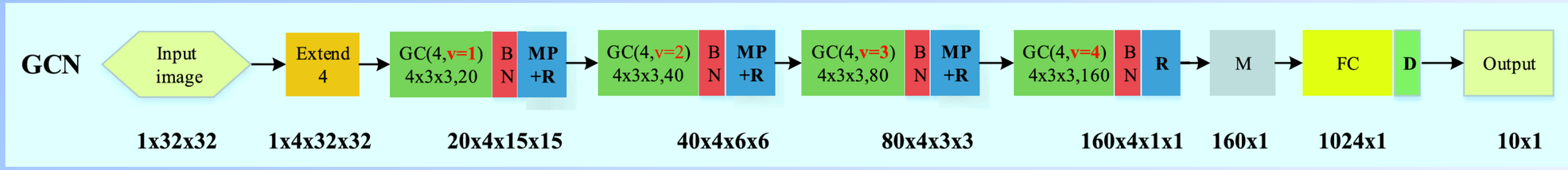
Gabor filter responses at different orientations and aspect ratios.

Gabor Neural Networks

- Gabor filters were shown to improve robustness against orientation and scale changes of the input when incorporated within a deep convolutional neural network architecture for object recognition.
- Learned CNN filters were modulated with Gabor kernels to produce Gabor-convolution layers
- The resulting https://ieeexplore.ieee.org/document/8354246 (GCN) also reduced the number of learnable



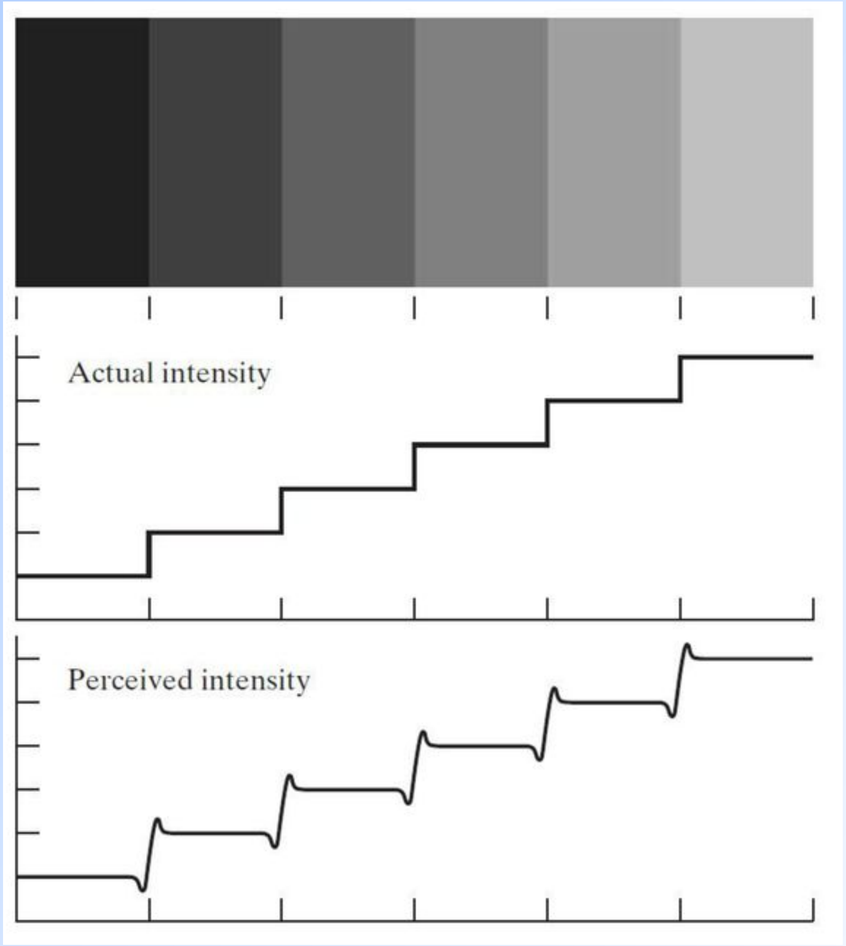
GCN convolution



GCN Network Architecture

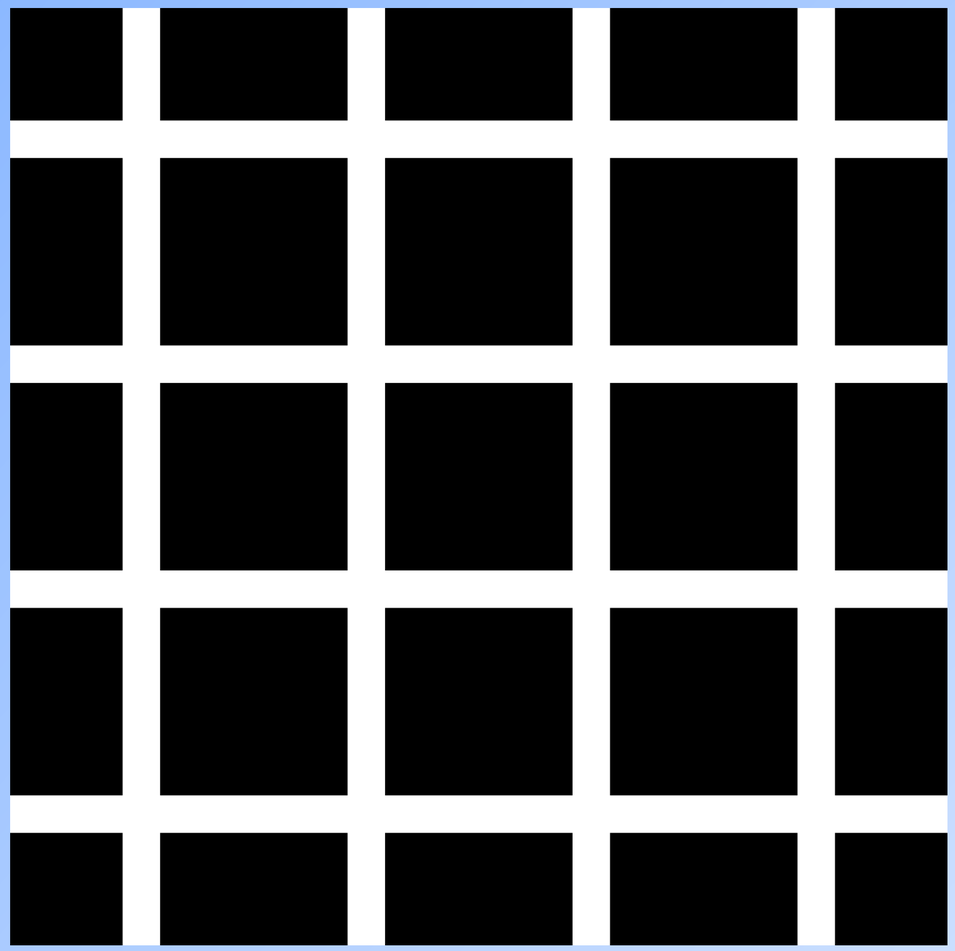
Lateral Inhibition and Contrast Adjustment in the Visual Cortex

- Lateral inhibition refers to the process whereby the activation of a neuron in the visual cortex suppresses the activation of neighboring neurons.
- The effect manifests in the form of enhancing contrast along intensity gradients.
- By sharpening the boundaries between different levels of light or color, lateral inhibition helps in detecting edges and fine details in visual scenes.
- Lateral inhibition causes optical illusions such as Mach bands, Hermann grid and scintillating grid.

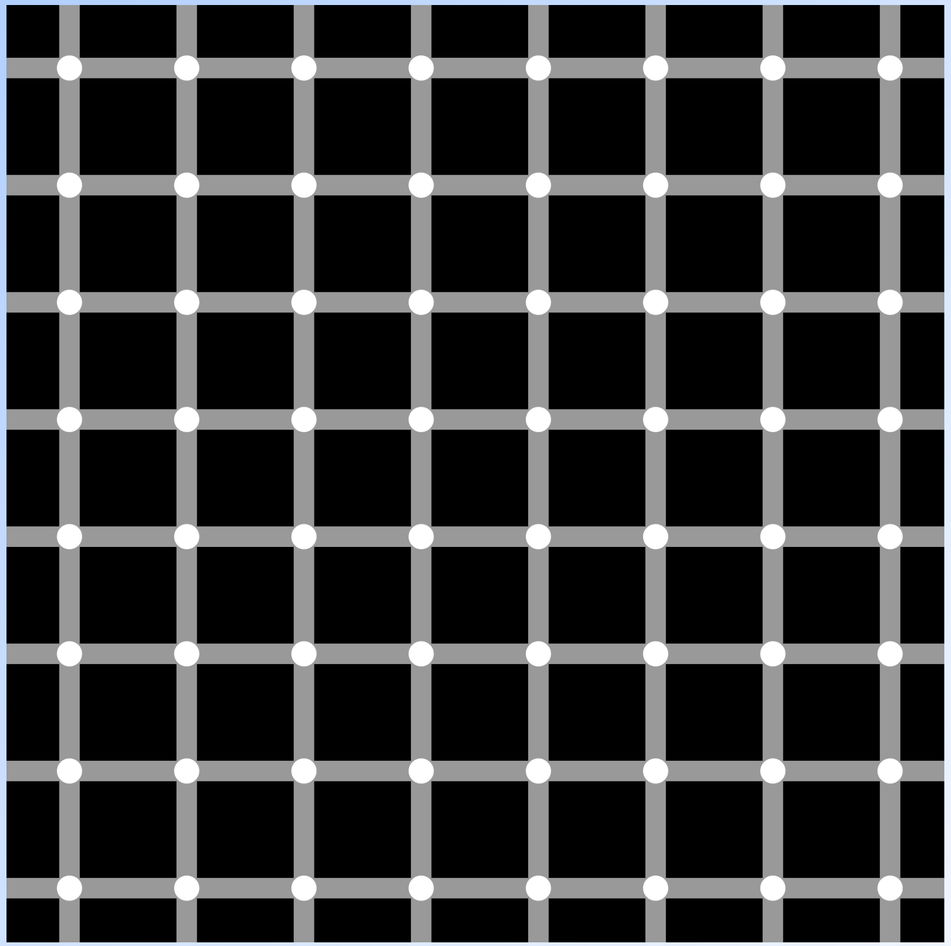


Mach Bands

Lateral Inhibition and Contrast Adjustment in the Visual Cortex



Hermann Grid



Scintillating Grid