

The Perception Pipeline

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 4 of "The NeuroAI Handbook"

How the brain turns photons into meaning, and how CNNs copied the blueprint.

One-sentence thesis. Vision is not a camera but a hierarchy: the retina, the visual cortex, and every convolutional network solve the same problem by stacking simple local detectors into progressively richer features.

PART 1: From Photons to Perception

The Problem Vision Has to Solve

A photograph is just a grid of brightness numbers. Nothing in that grid says "face" or "ball" or "curb ahead." Yet you recognize a friend across a crowded room in a fraction of a second. The gap between raw light and meaning is the entire problem of vision, and the brain does not close it by seeing harder. It closes it by *transforming* the signal, stage by stage, until an object falls out the far end.

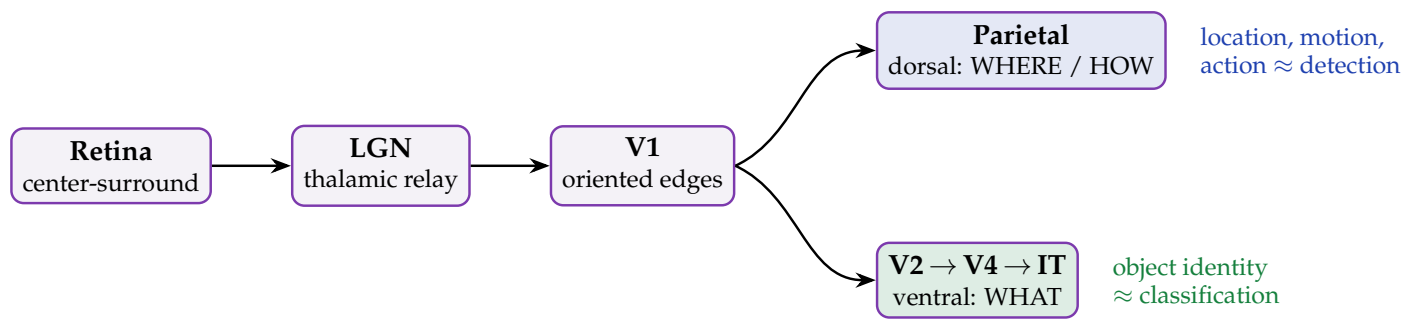
The pipeline starts before the brain. The **retina** is not a passive sensor; it is a pre-processor. Photoreceptors (rods and cones) capture light, but a network of downstream cells immediately enhances contrast and detects edges, so the signal that leaves the eye is already cleaned up. From there it passes to the **LGN** (lateral geniculate nucleus), a relay station in the thalamus that sorts and routes, and then to **V1**, the primary visual cortex, where the scene is first broken into oriented edges, lines, and motion.

Two Streams: What vs. Where

From V1 the pathway forks, and the split is one of the most important facts in vision science. The **ventral stream** ("what") runs down toward the temporal lobe through V2, V4, and IT cortex; its job is *object recognition*, and neurons along it respond to ever more complex shapes, from curves to whole faces. The **dorsal stream** ("where/how") runs up into the parietal lobe; its job is *spatial awareness and action*, tracking where things are and how to reach them. Goodale and Milner (1992) named this division of labor, and it maps almost exactly onto two AI tasks: the ventral stream is image *classification*, the dorsal stream is object *detection* and motor control.

ANALOGY

Picture the visual system as a corporation processing incoming mail. The retina is the mailroom (sorting, bundling raw letters into packages), the LGN is central dispatch, V1 is the junior analysts scanning for simple patterns, and higher cortex is the executives assembling the full picture. No single clerk sees the whole company. That is the secret: local processing plus hierarchical abstraction.



The pipeline: photons enter at the retina, get relayed through the LGN to V1, then fork into a dorsal "where" stream and a ventral "what" stream.

A Neuron's Window on the World

The single idea that unlocks both biological and artificial vision is the **receptive field**: the small patch of the visual world a given neuron actually responds to. A neuron does not see everything; it sees a keyhole, and it fires only when the right pattern lands in that keyhole.

The shape of the keyhole changes as you climb the pipeline. In the **retina**, receptive fields are small circles with a *center-surround* structure, excited by light in the center and inhibited around it, which makes them superb contrast and edge detectors. In V1, Hubel and Wiesel's Nobel Prize-winning experiments showed the keyholes are tuned to *orientation*: a "vertical" neuron fires vigorously at a vertical edge and ignores a horizontal one.

From Simple to Complex

V1 already builds a small hierarchy. **Simple cells** respond to an oriented edge at a *specific* location. **Complex cells** respond to that same edge *anywhere* in their receptive field; they achieve this by pooling the

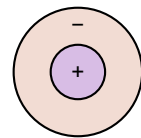
outputs of several simple cells, which buys **translation invariance**, the ability to recognize a feature regardless of exactly where it sits. **Hypercomplex (end-stopped) cells** go further, responding to corners, curves, and line endings.

The logic is compositional. Just as letters combine into words and words into sentences, the brain combines points of light into oriented lines, lines into corners, corners into parts, and parts into objects. **The point.** Complexity is never wired in directly; it is *built up* from a vocabulary of simple, reusable detectors.

A **Gabor filter**, a sine wave multiplied by a Gaussian envelope, is the standard mathematical model of a V1 simple cell. In symbols, the receptive field is a wave under a soft window:

$$g(x, y) = e^{-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}} \cos\left(\frac{2\pi x'}{\lambda}\right)$$

Tune the orientation and you get a bank of edge detectors, one per angle, exactly the tiling of orientations found in V1.



retina: contrast



V1: orientation

Two receptive-field shapes. Left: a center-surround field detects contrast. Right: an oriented bar detector fires for a vertical edge and ignores others.

The Uncanny Parallel

Once you have the cortical hierarchy in mind, a **Convolutional Neural Network** stops being mysterious. A CNN is, in essence, a computational model of the ventral visual stream, and this was no accident: Fukushima's Neocognitron (1980) was explicitly modeled on simple and complex cells, LeCun and colleagues (1998) made it trainable end to end, and Krizhevsky's AlexNet (2012) scaled it up to win at object recognition.

Four correspondences carry almost the entire analogy. **Convolutional filters are receptive fields:** a kernel is a small window sliding over the image, looking for one pattern, just like a V1 neuron. **Weight sharing mirrors a universal detector:** the network applies the *same* filter everywhere, exactly as the brain reuses one "vertical edge" template across the whole visual field, and it is what makes the whole scheme efficient. **Stacked layers rebuild the hierarchy:** early layers

learn edges (like V1), middle layers learn textures and shapes (like V2 and V4), deep layers learn whole objects (like IT). **Pooling layers are complex cells:** a max-pool takes the strongest response in a small region, keeping *what* is present while discarding *where*, which is precisely how complex cells manufacture translation invariance.

What the Filter Computes

Formally, a convolution slides a small kernel K over the input and sums the overlap at every position, producing a feature map that lights up wherever the pattern occurs:

$$(I * K)_{x,y} = \sum_{m,n} I_{x+m, y+n} K_{m,n}$$

One kernel, swept across the image, is one detector deployed everywhere at once, the engineering echo of weight sharing in cortex.

Visual cortex	CNN component	What it computes
Retina (center-surround)	Input + edge preprocessing	Contrast, local edges
V1 simple cells	Convolutional filter (kernel)	Oriented edges at a position
V1/V2 complex cells	Max-pooling layer	Translation invariance
V2 / V4	Mid convolutional layers	Textures, corners, parts
IT cortex	Deep layers + classifier head	Whole objects, categories
One template, whole field	Weight sharing (shared kernel)	Efficiency: one detector, everywhere

Where the Analogy Strains

The clean story above is feedforward: signals march from retina to IT and out pops a label. Real brains are not like that. They are saturated with *feedback*: connections running backward down the hierarchy vastly outnumber the forward ones, and they carry context, expectation, and attention. Early CNNs threw all of that away, which is exactly why the frontier of AI vision now runs toward recurrence and attention.

Vision Transformers (ViTs). Rather than sliding local filters, a ViT (Dosovitskiy et al., 2021) cuts the image into patches and uses *self-attention* to weigh how every patch relates to every other. That captures global context and long-range relationships, closer to the brain's top-down, whole-scene understanding than a stack of purely local convolutions.

Predictive coding and generative models. A leading theory holds the brain is constantly *predicting* its sen-

sory input and only propagating the error. Generative models (GANs, diffusion) that learn to *create* images rather than merely label them are the AI reflection of this idea.

A Two-Way Street

The relationship is a genuine dialogue, not one-way inspiration. **Biology gave the blueprint:** hierarchical layers, local receptive fields, and the what/where split were discovered by neuroscientists first. **AI returns a model:** when a CNN trained only on images spontaneously grows V1-like edge detectors, it is evidence our theory of the brain is on the right track. Brain-Score work (Schrimpf et al., 2018; Yamins and DiCarlo, 2016; Cadieu et al., 2014) found deep networks can *rival* primate IT cortex at predicting neural responses. Faced with the same visual world, brains and machines converged on strikingly similar solutions, which hints these principles may be universal.

Quick Self-Test

1. Trace a signal from photon to object label, naming the four main stages and what each contributes.
2. What is a receptive field, and how does its shape differ between the retina and V1?
3. How do complex cells achieve translation invariance, and which CNN component copies this trick?
4. Weight sharing in a CNN corresponds to what property of biological vision, and why does it matter for efficiency?
5. Name one way real vision departs from a feedforward CNN, and one AI architecture that tries to close the gap.

If You Remember One Thing

Both the visual cortex and a CNN win by refusing to solve vision all at once. They stack a vocabulary of simple, reusable, local detectors into a hierarchy, edges to textures to parts to objects, and pool along the way to stop caring about exact position. The architecture is the insight; everything else is detail.

Richard J. Young, Ph.D. · richard@deepneuro.ai · Senior AI Research Scientist, UnitedHealth Group · UNLV Lee Business School · ORCID 0000-0002-1109-7552