CHAPTER 1

Foundations of Vision

FRANK TONG

THE PURPOSE OF VISION

For people with intact vision, it would be hard to imagine what life would be like without it. Vision is the sense that we rely on most to perform everyday tasks. Imagine if instead you had to accomplish all of your daily routines while blindfolded. We depend on vision whenever we navigate to work by foot or by car, search for our favorite snack in the grocery aisle, or scan the words on a printed page trying to extract their underlying meaning. For many mammals and especially for higher primates, vision is essential for survival, allowing us to reliably identify objects, food sources, conspecifics, and the layout of the surrounding environment.

Beyond its survival value, our visual sense provides us with an intrinsic source of beauty and pleasure, a tapestry of richly detailed experiences. We may find ourselves captivated by an expansive view from a seaside cliff, a swirl of colors in an abstract oil painting, or an endearing smile from a close friend.

The power of vision lies in the dense array of information that it provides about the surrounding environment, from distances near and far, registered by the geometry of light patterns projected onto the backs of the eyes. It is commonly said that *a picture is worth a thousand words*. Consider for a moment the chirping activity of the ganglion cells in your retinae right now, and their outgoing bundle of roughly 1 million axonal fibers through each optic tract. Following each glance or microsaccade, a new pattern of activity is registered by the photoreceptors, then processed by the bipolar neurons and the ganglion cells, after which these high-bandwidth signals are relayed to the lateral geniculate nucleus and ultimately to the visual cortex for in-depth analysis.

Psychologists and neuroscientists have made remarkable advances in understanding the functional organization of the visual system, uncovering important clues about its perceptual mechanisms and underlying neural codes. Computational neuroscientist David Marr (1982) once quipped that the function of vision is "to know what is where by looking." As Marr well appreciated, the problem underlying vision is far easier to summarize than it is to solve. Our visual system does a remarkably good job of solving this problem, getting things pretty much right about 99.9% of the time. On those rare occasions where the visual system seems to come up with "the wrong answer," as in the case of visual illusions, scientists can gain insight into the powerful computations that underlie the automatic inferences made by the visual system.

Perception, Introspection, and Psychophysics

Most fields of natural science rely exclusively on third-person observation and experimentation. In contrast, vision scientists can learn a great deal from introspecting on their personal visual experiences and by directly testing their own eyes and brains. The seminal contributions of vision research to the emergence of psychology as a field can be explained by the fact that scientists could so readily test and analyze their own perceptions.

Some early discoveries were made by fortuitous observation, such as when Addams (1834) noticed after staring at a waterfall that his subsequent gaze at the neighboring rocky cliff led to an unexpected impression of upward motion. His description of the *motion aftereffect*, or waterfall illusion, helped set the path toward the eventual development of ideas of neuronal adaptation and opponent-based coding to account for visual aftereffects. Other discoveries involved more purposeful observations and simple experiments to characterize a perceptual mechanism. Sir Charles Wheatstone devised an optical apparatus to present different pictures to the two eyes, and then drew simple pictures to capture how a 3D object would appear slightly differently from the vantage point of each eye. By presenting these image pairs in his stereoscope, he discovered that it was possible to re-create an impression of stereo-depth from flat pictures. He also found that distinct patterns presented to the two eyes could induce periodic alternations in perception, or form-based binocular rivalry. His optical invention grew so popular (akin to the current-day popularity of 3D TV and 3D movies) that the Wheatstone *stereoscope* could be found in many parlor rooms in England in the 1800s.

As the process of characterizing perception became more formalized, a scientific methodology evolved. *Psychophysics* refers to experimental methods for quantifying the relationship between the psychological world and the physical world, which usually involves systematic manipulations of a stimulus and measuring its perceptual consequences. For instance, Weber reported that the ability to detect a *just noticeable difference* (JND) between two stimuli depended on their relative difference (or ratio) rather than the absolute difference. Expanding upon this idea, Fechner (1860) proposed that the perceived intensity of a sensory stimulus should increase in a predictable manner proportional to the logarithm of its physical intensity. Specifically, $S = \log(I)$, where S refers to the intensity of the sensation and I refers to the intensity of the physical stimulus. By describing this simple lawful relationship between physical intensity and psychological experience, the field of visual psychophysics was born. A central tenet of visual psychophysics is that perceptual states can be quantified and formally characterized, to help reveal the underlying mechanisms.

Signal Detection Theory

A fundamental advance in visual psychophysics was the application of *signal detection theory* to quantify the sensitivity of the human visual system. This statistical theory was originally developed to address the problem of detecting faint radar signals reflected by a target in the presence of background noise (Marcum, 1947). In visual psychophysics, this same logic and approach can be applied to both *visual detection* and *visual discrimination* paradigms (Tanner & Swets, 1954). These concepts are central to vision research, so we will spend a good amount of time reviewing them, but if they are already very familiar to you, consider moving on to the section "Why Vision Is a Hard Computational Problem."

A common design for a visual detection task is as follows. There is a 50/50 chance that a very faint target stimulus will be presented on each trial, and the observer's task is to make a binary decision regarding whether the target was present or absent. Let us assume that the stimulus is extremely weak and that

the visual system has some inherent level of noise, so perfect performance is impossible. There are four possible stimulus-response outcomes, as shown in Figure 1.1A. If the target stimulus is present and the observer correctly reports "target present" this would constitute a *hit*, but if the observer incorrectly reports "target absent" this would constitute a *miss*. Now, consider trials where the target is absent and the observer correctly reports "target absent"; this would be a *correct rejection*. But if the observer incorrectly reports "target present," this would be considered a *false alarm*.

Now, imagine that a set of neurons in the brain is selectively activated by the target, but these neurons exhibit some degree of intrinsic noise even when the target is absent. For example, the baseline firing rate of these neurons may vary somewhat from trial to trial. If a device was used to read out the activity of these neurons, how would it decide whether the target was presented or not on a given trial?

In Figure 1.1B, you can find hypothetical probability density functions that illustrate how active these neurons will be under two scenarios: when the target is absent and the response arises from noise only, and when the target is present and the response arises from noise plus signal. (If a stronger neural response occurred on a given trial, it would correspond to an observation further to the right on the abscissa. For mathematical convenience, the noise distribution is plotted with a mean value of zero, even though in reality, a neuron's mean baseline firing rate must be greater than zero and cannot produce negative values.) Note how the two distributions partially overlap such that perfect discrimination is impossible. Both distributions are Gaussian normal with a common standard deviation of σ, corresponding to the level of intrinsic noise, whereas the distance D between their central means corresponds to the magnitude of the signal-induced activity. According to signal detection theory, sensitivity at this detection task is mathematically specified by the *signal-to-noise ratio* or what

(A)

Response	Target: Present	Target: Absent
"*Present*"	Hit	False Alarm
"*Absent*"	Miss	Correct Rejection

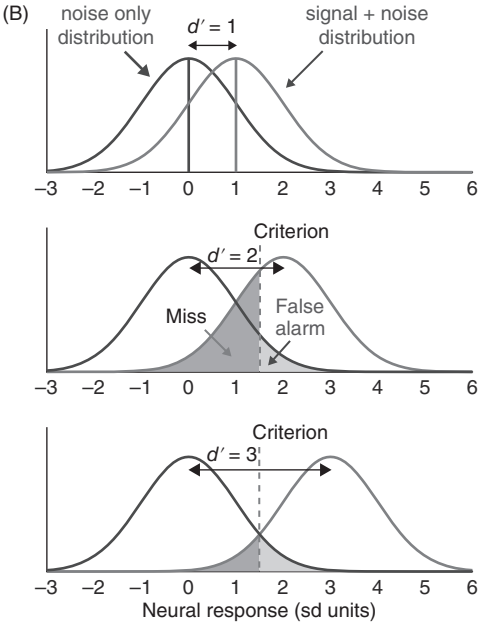

Figure 1.1 Overview of signal detection theory. (A) Table showing classification of an observer's responses to a target stimulus, regarding its presence or absence. (B) Signal detection theory proposes that the signal + noise distribution is separated from the noise only distribution by distance D. Assuming that both distributions share a common standard deviation, σ, then visual sensitivity or d' in this task will be determined by D/σ. As the signal becomes stronger, the signal + noise distribution shifts rightward, leading to larger d' and allowing for better detection performance. Examples of $d' = 1$, 2, and 3 are shown. The vertical dashed line indicates the criterion (β) that the observer uses for deciding whether the target is present or absent. If the criterion lies midway between the two distributions, the observer is unbiased and the proportion of misses and false alarms will be equal (bottom panel). Relative to the midway point, leftward shifts lead to a more liberal criterion for reporting target present, while rightward shifts lead to a more conservative criterion. The middle panel depicts a conservative criterion, where the proportion of false alarm responses would be reduced, but at the cost of a greatly inflated proportion of miss responses. Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.
SOURCE: Figure created by Frank Tong; used with permission of the author.

is commonly called *d-prime* or d', where $d' = D/\sigma$.

Greater visual sensitivity and larger d' values will arise when the noise-only distribution and noise-plus-signal distribution are more separated, sharing less overlap. For d' values of 1, 2, or 3, the *nonoverlapping* portions of the two distributions would comprise about 69%, 84%, and 93% of the total area under the two curves. This percentage of nonoverlap corresponds to the maximum accuracy that one might attain in a detection task if the observer were unbiased. If the two distributions overlapped entirely, d' would equal zero and performance would be at chance level.

Performance at this task also depends on the *criterion* that the observer adopts for deciding whether the target is present or absent. If the threshold is set to where these two probability density functions intersect (Figure 1.1B, bottom panel with $d' = 3$), then responses will be unbiased. That is, an equal proportion of miss responses and false alarm responses will be made. If instead, the observer adopts a *conservative criterion* by setting a threshold that lies to the right of the midway point between the two distributions (see Figure 1.1B, middle panel with $d' = 2$), then a higher level of activity will be required to respond "target present." As a consequence of this conservative criterion, the proportion of false alarm responses will be lower, but the proportion of hit responses will also be lower, resulting in a greater proportion of miss responses (hit rate = 1 − miss rate). Conversely, if the observer adopts a *liberal criterion* by shifting the threshold to the left, so that lower levels of activity are needed to report "target present," then the proportion of misses will decrease (i.e., more hits) but the proportion of false alarms will increase. Larger biases that lead to a greater imbalance between the frequency of these two types of errors—misses and false alarms—result in a higher overall error rate. Despite this inherent cost of bias, there are certain situations where a bias might be preferable. For example, one might favor a liberal criterion for a diagnostic medical test to minimize the likelihood of reporting false negatives.

Vision scientists are usually more interested in characterizing the visual sensitivity of the observer rather than decisional bias. A strategy for measuring sensitivity more efficiently and eliminating bias is to adopt a *two-alternative forced-choice* (2AFC) paradigm, by presenting a target to detect on every trial at say one of two spatial locations or during one of two temporal intervals. By requiring the observer to report which location/interval contained the target, a target present response is obtained on every trial, thereby eliminating the possibility of bias. Researchers have found that people's performance on 2AFC tasks can be modeled by assuming that the observer can determine the difference in the strength of the signal/noise received in each of the two intervals, and then base their decision on that difference signal.

Characterizing Visual Sensitivity

Signal detection theory provides the theoretical foundation for modern day psychophysics and a powerful approach for characterizing human visual sensitivity across a range of stimulus conditions. To get an idea of this approach in action, consider Figure 1.2A, which shows the detection accuracy as a function of stimulus contrast for gratings presented at two different spatial frequencies. Performance at the fovea is much better at spatial frequencies of 1.0 cycles per degree (cpd) than at extremely higher frequencies of 32 cpd. By fitting a psychometric function to these data, one can identify the

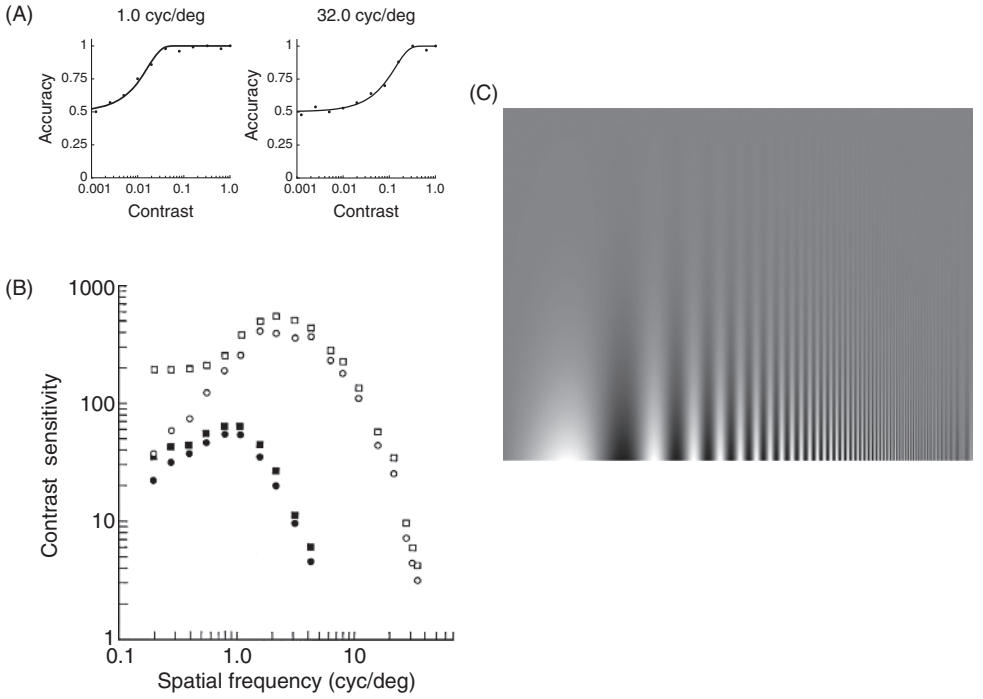

Figure 1.2 Contrast sensitivity as a function of spatial frequency. (A) Examples of psychometric functions showing detection accuracy plotted as a function of stimulus contrast. (B) Contrast sensitivity plotted as a function of spatial frequency for sine-wave gratings (circles) and square-wave gratings (squares) under brightly lit (500 cd/m^2) viewing conditions (open symbols) and dimly lit (0.05 cd/m^2) scotopic viewing conditions. Square-wave gratings are easier to detect at very low spatial frequencies, because they contain higher spatial frequency components that exceed the observer's contrast threshold. With scotopic viewing, rod photoreceptors are sensitive to much lower range of spatial frequencies. (C) Visual demonstration of how contrast sensitivity varies with spatial frequency.
Each row of pixels shows a common range of luminance modulation, with the highest contrast appearing at the bottom of the figure and progressively lower contrasts appearing above. Lower spatial frequencies appear to the left in the figure and higher spatial frequencies appear to the right. Perception of a hill-shaped bump of contrast modulation, akin to the open circles plotted in (B), is due to superior sensitivity at moderately high spatial frequencies.
SOURCE: (A) Example figures of performance accuracy as a function of contrast created by Frank Tong; used with permission from the author. (B) From Campbell and Robson (1968).

contrast level at which performance reaches 76% correct in this 2AFC task (corresponding to $d' = 1$) to characterize the observer's sensitivity at each spatial frequency. Figure 1.2B shows contrast sensitivity as a function of spatial frequency, and the shape of the full contrast sensitivity curve (open circles). The dependence of visual sensitivity on spatial frequency can be directly experienced by viewing the Campbell-Robson contrast sensitivity chart (Figure 1.2C), where each row of pixels depicts a common range of luminance variation at progressively higher spatial frequencies (from left to right). Sensitivity is highest at intermediate spatial frequencies, where one can perceive the stripes extending farther upward along the chart.

This ability to quantify visual sensitivity across a range of stimulus conditions is remarkably powerful. For example, Campbell and Robson (1968) could accurately predict the differences in contrast sensitivity to sine-wave and square-wave gratings, based on signal detection theory and Fourier analysis of the spatial frequency content of the gratings. Likewise, this approach has been used to characterize the differences in spatial resolution under bright and dimly lit conditions (see Figure 1.2B), as well as the differences in temporal sensitivity under these two regimes. Such approaches have also been used to estimate the spectral absorption properties of cone receptors, by using psychophysical methods to quantify visual sensitivity to different wavelengths following selective color adaptation (Smith & Pokorny, 1975; Stockman, MacLeod, & Johnson, 1993). Studies have further revealed the exquisite sensitivity of the visual system following dark adaptation. Indeed, human observers are so sensitive that their detection performance is modulated by quantum level fluctuations in light emission and absorption (Hecht, Shlaer, & Pirenne, 1941; Tinsley et al., 2016).

Signal detection theory can also be used to quantify how well observers can discriminate among variations of a stimulus. For example, if one were to judge whether a grating was subtly tilted to the left or right of vertical, the two distributions shown in Figure 1.1B can instead be conceptualized as the neuronal responses evoked by a leftward tilted stimulus and a rightward tilted stimulus. Studies such as these have shown that orientation thresholds remain remarkably stable across a wide range of contrast levels, leading to the notion that orientation-selective neural processing is largely contrast invariant (Skottun et al., 1987). Studies have also revealed that visual sensitivity is not perfectly uniform across orientations. People are more sensitive at discriminating orientations that are close to horizontal or vertical (i.e., *cardinal orientations*) as compared to orientations that are oblique. Later in this chapter, we will also see how signal detection theory has been used to characterize how top-down attention can improve visual performance at detection and discrimination tasks.

From what we have just learned, it should be clear that the psychophysical approach is essential for characterizing the sensitivity of the human visual system. Although neuroscience data can be highly informative, many critical factors are grossly underspecified, such as how the brain combines and pools signals from multiple neurons or what information the observer will rely on when making a perceptual decision. A case in point is that of *visual hyperacuity*: People can distinguish relational shifts between two-point stimuli, even when they are spatially shifted by just fractions of a photoreceptor unit (Westheimer & McKee, 1977). Without psychophysical testing, this empirical finding would have been very difficult to predict in advance. Psychophysical measures of visual performance provide the benchmark of the visual system's sensitivity, by directly testing the limits of what a person can or cannot perceive.

Why Vision Is a Hard Computational Problem

The initial encoding and processing of local visual features, such as luminance, color, orientation, and spatial frequency, provides an essential front end for visual perception. After these early processing stages, however, the visual system faces even greater challenges it must solve. Indeed, much of what the visual system must do is interpretive and inferential in nature. Following each eye movement, this system is presented with a distinct pattern of light on the retina, akin to a new megabyte puzzle that must be solved.

3	4	4	2	1	1	1	1	1	2	5	4	2	1	1	3	3	1	1	3	2	1	1	1	1	1	1	1	3
3	6	6	3	1	1	1	1	1	1	1	1	0	0	1	3	4	1	1	3	3	2	1	1	1	1	0	1	4	6
4	6	7	5	1	1	1	1	0	0	0	0	0	0	0	0	1	1	2	4	4	3	1	1	1	1	1	4	6	7
4	7	8	6	3	1	1	0	0	1	1	1	2	3	2	2	2	1	1	3	4	3	1	1	1	1	3	6	7	7
5	7	8	7	5	1	0	1	2	4	4	4	4	4	4	3	3	3	1	1	2	2	1	1	1	3	6	7	7	7
5	7	8	8	4	1	3	4	5	6	6	5	4	4	4	3	3	3	2	1	0	1	1	1	2	4	6	7	7	7
6	8	8	7	1	3	5	5	6	6	6	6	5	4	4	4	3	3	2	1	1	0	1	1	2	3	5	7	7	7
6	7	8	4	2	5	6	7	7	6	6	6	5	4	4	4	3	3	3	2	1	0	1	2	2	2	4	6	8	8
6	7	7	2	3	6	6	7	7	6	6	6	6	6	5	4	3	3	2	3	2	1	1	2	2	2	3	5	7	8
6	7	5	1	4	5	6	6	7	7	7	6	5	5	3	1	1	1	1	1	2	2	0	2	2	2	2	4	6	7
6	7	3	2	4	5	6	7	7	6	5	5	5	4	2	1	1	1	1	1	1	2	1	1	2	2	2	2	5	7
5	7	3	2	4	5	6	6	5	2	1	2	5	5	2	2	2	1	1	1	1	2	1	1	2	2	2	2	3	6
4	6	3	2	4	5	5	3	1	1	1	2	5	6	2	2	3	3	2	2	2	3	2	1	1	2	2	2	2	4
3	5	3	2	4	5	4	2	3	1	1	2	5	6	4	2	3	3	3	3	3	3	2	1	1	1	2	2	2	3
2	3	2	1	4	5	3	3	2	2	3	4	6	6	5	3	2	4	3	4	3	2	2	1	1	2	2	2	2	2
2	2	2	1	4	5	4	3	4	5	5	6	6	6	5	3	2	3	4	3	3	2	2	2	1	2	2	2	2	2
2	2	2	1	3	5	5	6	6	7	6	7	6	6	4	2	1	2	3	4	3	2	2	2	1	1	2	2	2	2
2	2	2	1	1	5	6	7	7	7	6	6	5	2	2	1	1	2	3	4	3	2	2	2	1	1	2	2	2	2
2	2	2	3	3	4	5	6	7	6	6	6	4	3	4	4	3	3	3	2	2	2	2	2	1	2	2	2	2	2
2	2	2	4	6	5	5	5	6	6	6	5	6	6	5	4	2	1	1	1	2	2	2	1	3	2	2	2	2	2
2	2	2	5	7	5	4	5	5	6	6	5	6	5	2	1	1	1	1	1	1	2	1	1	3	4	2	2	2	2
2	2	2	4	6	4	4	5	5	5	5	5	5	3	1	3	3	2	2	2	2	1	1	1	1	3	3	2	2	2
2	2	2	3	5	5	5	4	5	5	5	4	3	2	4	3	2	2	2	2	1	1	1	1	1	0	2	2	2	2
2	2	2	2	4	6	5	4	5	5	5	5	4	5	5	3	3	2	3	2	2	1	1	1	1	0	0	0	0	0
3	2	2	2	2	4	5	4	4	5	5	4	4	4	4	4	4	4	4	3	2	1	0	1	1	0	0	0	0	0
5	3	2	2	2	2	2	3	5	4	4	4	4	5	5	5	5	4	3	2	1	1	0	1	2	0	0	0	0	0
6	4	2	2	2	2	2	3	5	5	4	4	4	4	5	5	4	2	2	1	0	0	0	2	2	0	0	0	0	0
7	5	3	2	2	2	2	2	4	7	5	3	3	3	3	3	2	1	1	1	0	0	1	3	1	0	0	0	0	0
7	6	4	2	2	2	2	2	3	5	5	4	3	2	2	1	1	1	1	1	0	1	3	3	0	0	0	0	0	0
7	6	5	3	2	2	2	2	1	1	0	1	4	3	2	2	1	1	1	1	0	2	4	2	0	0	0	0	0	0

Figure 1.3 How to recognize an array of numbers depicting an image. An image of a recognizable object becomes impossible to perceive when it is presented as a matrix of numbers rather than as light intensity values. This figure conveys a sense of the challenge faced by our visual system when interpreting patterns of light. The grayscale version of this image is shown in Figure 1.4.

Look at the two-dimensional array of numbers shown in Figure 1.3. Can you tell what object is embedded in this matrix of numbers? Larger numbers correspond to brighter pixels of an image. This is the kind of input a computer vision algorithm would receive if it were tasked with identifying objects in digital images. When faced with a real-world image in this paint-by-numbers format, it becomes apparent that our visual system must solve a very challenging computational problem indeed. You probably have no idea what this image depicts. Yet if the numbers were converted into an array of light intensities, the answer would be obvious (see Figure 1.4).

Figure 1.4 Digitized image of the array shown in Figure 1.3. Grayscale image with intensity values specified by the matrix in Figure 1.3, showing a coarse-scale digitized image of President Barack Obama.
SOURCE: Image adapted by the author.

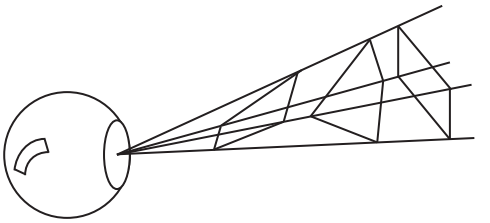

Figure 1.5 The inverse optics problem. The inverse optics problem refers to underconstrained nature of visual inference. For example, any number of quadrilateral shapes in the environment that join together the four lines of sight (drawn in blue) would create the same rectangular pattern on the retina. How then does the visual system infer the shape of an object from the 2D pattern observed on the retina?
SOURCE: Figure created by Frank Tong; used with permission of the author.

This problem is challenging for several reasons. First and foremost, the visual input that we get from the retina is underspecified and ambiguous. People tend to think of *seeing as believing*, but in reality, the visual system rarely has access to ground truth. Instead, it must make its best guess as to what physical stimulus out there in the world might have given rise to the 2D pattern of light intensities that appear on the retina at this moment. This is known as the *inverse optics problem* (Figure 1.5). Given the *proximal stimulus* (i.e., the retinal image), what is the *distal stimulus* that could have given rise to it?

Consider the scene depicted in Figure 1.6A and the square patches marked with the letters A and B. Which square looks brighter? Actually, the two patches have the same physical luminance, yet pretty much everyone perceives B to be much brighter than A. If you cover the other portions of the image, you can see for yourself that the two squares are the same shade of gray.

This well-known brightness illusion, created by Ted Adelson, illustrates that people do not perceive the brightness of a local region in terms of the raw amount of light that is emitted from that region. Context matters: The fact that square B appears to be lying in a shadow while square A is exposed to light has a strong influence on this perceptual judgment. Some people might think of this illusion as revealing the mistakes made by the visual system. Humans can be easily swayed by contextual factors to make gross errors—a photometer would perform so much better! However, another way to think about this illusion is that our visual system is remarkably sophisticated, as it is trying to infer a more complex yet stable property of the visual world, namely, the apparent "paint color" or *reflectance* of the local surface patch. Knowing the stable reflectance of an object is far more useful than simply knowing what colors are being reflected from its surface. For example, it would be helpful to know whether a banana is greenish or ripe, regardless of whether it is viewed in broad daylight, cool fluorescent light, or in the orangey glow of sunset.

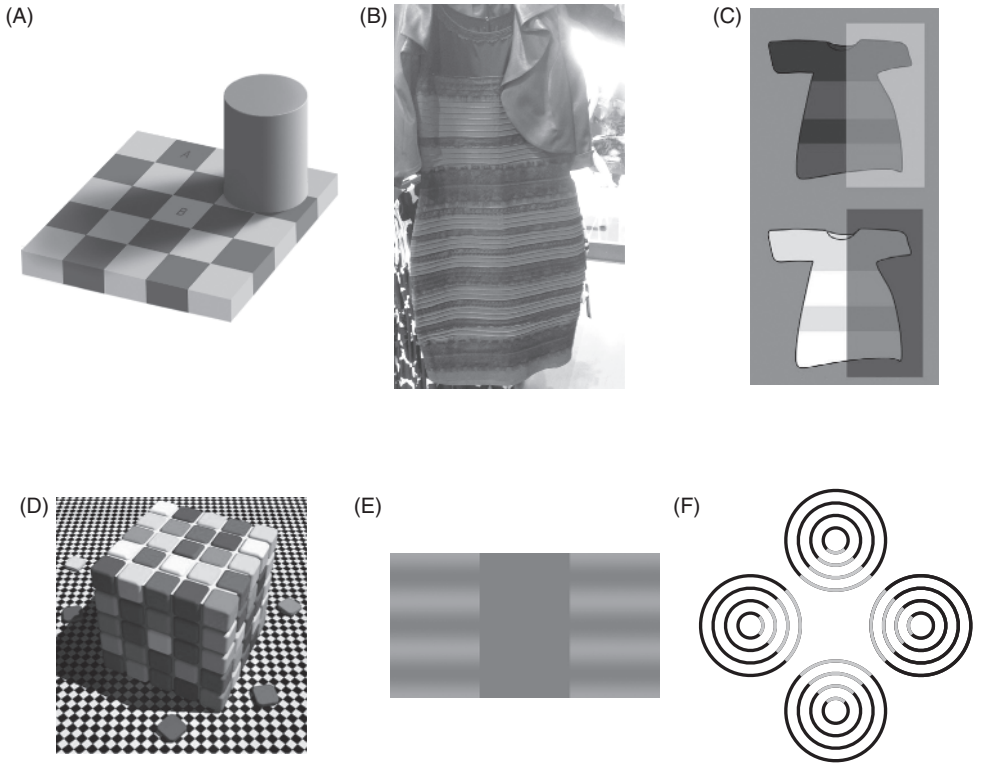

Figure 1.6 Examples of visual illusions. (A) Adelson checkboard brightness illusion. (B) *#TheDress*. (C) The right side of each dress consists of the exact same physical colors, but the apparent reflectance of each dress is very different, as the left one appears to be lit by yellowish light, and the right one appears in a bluish shadow. (D) Color perception illusion. The middle square on the top surface and the middle square on the front surface actually show the same physical color, but they are perceived very different. (E) Visual phantom illusion. The two sets of horizontal gratings are separated by a uniform gray gap, but people tend to perceive the gratings as extending through the blank gap region. (F) Subjective contour illusion, induced by the sudden color transition on the inducers. The blue inducing components can lead to the perception of an illusory transparent diamond shape hovering in front of the inducers, as well as neon color spreading. Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.
SOURCE: (A) Reproduced with permission from Edward Adelson. (B) Dress image reproduced with permission from Cecilia Bleasdale. (D) Reproduced with permission from Beau Lotto. (E), (F) Used with permission from Frank Tong.

Determining the reflectance of an object is an underspecified computational problem, one that requires some inference and guesswork. Why? Because the amount (and spectral distribution) of light that reaches our eye from an object is determined by two factors: the object's reflectance and the light source that is shining on the object. Unless we know the exact lighting conditions, we cannot know the true reflectance of the object. (This problem is akin to being told that the number 48 is the product of two numbers, X and Y, and being asked to figure out what X and Y are.) Usually, the visual system can make a pretty good inference about the nature of the light source by comparing the luminance and

color spectra of multiple objects in a scene, separately analyzing the regions that appear to receive direct light and those that are in shadow. But sometimes it can prove tricky to infer the nature of the light source.

A striking example of this comes from #TheDress (Figure 1.6B), an amateur photo taken in a clothing store that became a viral sensation on the Internet. Why? People were shocked to discover that not everyone perceives the world in the same way. Some people perceived the dress as appearing as blue and black, whereas others saw it as white and gold.

This illusion arises in large part because some people perceive the dress to be lying in direct sunlight, in which case the dress must have a dark reflectance (blue and black), whereas others perceive the dress to be in shadow, under an unseen awning (Lafer-Sousa, Hermann, & Conway, 2015; Winkler, Spillmann, Werner, & Webster, 2015). To appreciate how the inferred light source can affect the perception of brightness and color, Figure 1.6C shows a simpler illusion, similar to #TheDress. The right portion of each dress shows identical physical colors, yet they are perceived differently depending on whether they appear to lie in yellowish light or bluish shadow. So what one perceives depends on what the visual system infers about the source of illumination (see Figure 1.6D for another color/brightness illusion).

The *inverse optics problem* also occurs when we must infer the 3D structure of an object from its 2D pattern of the retina. (Binocular depth and motion parallax cues are often weak or absent.) There are thousands of common objects that we know by name, and depending on the observer's viewpoint and the lighting conditions, any single object can give rise to a multitude of 2D images. How then can one determine the 3D shape and identity of an object from the pattern of light it creates on the retina? Consider even a very simple pattern, such as a set of four lines that cast a rectangular pattern on the retina. It turns out that an infinite possible variety of quadrilaterals could have given rise to this retinal image (Figure 1.5). Indeed, even a set of four disconnected lines could lead to the same pattern on the retina, though admittedly, it would be surprising to stumble upon a set of lines that were arranged just so to be viewed from this line of sight. One strategy the visual system employs is to make the simplifying assumption that the current view is *nonaccidental*. Two lines that appear parallel on the retina are assumed likely to be parallel in the real world. Likewise, two lines that appear to terminate at a common point are assumed to form a junction in the 3D world. As we will see next, our perceptions can be well described as a form of statistical inference.

Perception as Statistical Inference

Hermann von Helmholtz described the nature of perception as one of *unconscious inference*. By *unconscious*, he meant that perceptual inferences are made rapidly and automatically, scarcely influenced by conscious or deliberative thought. When presented with a visual illusion such as the one shown in Figure 1.6A, we can be told that patches A and B actually have the same luminance. However, this cognitive information will not overcome the inferences that are automatically supplied by our visual system. When the surrounding context is particularly suggestive, as in cases of *perceptual filling-in*, the visual system may even infer the presence of a nonexistent stimulus, such as shadowy stripes (Figure 1.6E) or a hazy blue diamond (Figure 1.6F) extending through a physically blank region. Such illusions are often described as "fooling our very eyes." However, does this necessarily mean

that the visual system, and the computations that it makes, are foolish? As we will see, such a conclusion is unwarranted and far from the truth.

Although von Helmholtz did not know how to formalize the concept of unconscious inference back in the 19th century, since the 21st century there has been a growing appreciation that perception can be understood as a form of statistical or *Bayesian inference* (Ernst & Banks, 2002; Kersten, Mamassian, & Yuille, 2004; Knill & Pouget, 2004). Given the pattern of light that is striking the retinae (i.e., the sensory data or the proximal stimulus), the brain must infer what is the *most likely* distal stimulus that could have generated those sensory data. What the brain considers most likely will also depend on a person's expectations and prior experiences. For example, when judging an ambiguous stimulus such as #TheDress, some people may be predisposed to infer that the dress is lying in shadow, whereas others may consider it more likely that the dress is lying in direct sunlight, leading to drastically different perceptions of the same stimulus.

The formula for inferring the probability of a stimulus, given the sensory data, is as follows:

$$p(\text{stimulus} \mid \text{data}) = \frac{p(\text{data} \mid \text{stimulus}) \times p(\text{stimulus})}{p(\text{data})}$$

Since the denominator term, p(data), is independent of stimulus to be inferred, it can be effectively ignored with respect to determining the most likely stimulus that could have given rise to the observed sensory data. So, all that needs to be maximized to make this inference is the numerator term.

Notice that any system that seeks to determine the probability that the sensory data would result from given the stimulus, or p(data | stimulus), would require some type of memory representation of the many previous encounters with that stimulus, along with the sensory data evoked by those encounters. Likewise, the probability of encountering the stimulus, p(stimulus), is sometimes referred to as one's *prior* expectations, which also depend on a form of memory. What this implies is that vision does not simply reflect the processing of information in the here and now. Instead, it reflects the interaction between processing of the immediate sensory input and what has been learned over the course of a lifetime of visual experiences. A telltale example is that of face perception. We often see faces upright but rarely get to see them upside-down, so we have greater difficulty recognizing a face when the sensory data appears inverted on our retinae.

There is a growing body of evidence to support this Bayesian view of perception, though this theoretical framework has yet to be fully tested or validated. That said, even if the visual system does deviate from Bayesian inference in certain respects, this framework remains useful because it can help us appreciate the conditions in which visual processing deviates from statistical optimality.

FUNCTIONAL ORGANIZATION OF THE VISUAL SYSTEM

Now that we have a better grasp of the computational challenges of human vision, let's consider how the visual system actually solves them. In this section, we will review the anatomical and functional organization of the visual system, characterizing how visual information is processed and transformed across successive stages of the visual pathway from retina to cortex. With this knowledge in hand, we will consider how psychophysical and neural investigations have shed light on the mechanisms of visual perception, attentional selection, and object recognition.

The visual system processes information throughout the visual field in parallel, analyzing and transforming the array of visual signals from one processing stage to the next, through a series of hierarchically organized brain areas (see Figure 1.7A). After phototransduction and the early-stage processing of light information in the retina, the vast majority of retinal outputs project to the dorsal lateral geniculate nucleus of the thalamus (LGN). LGN relay neurons in turn have dense projections to the input layer of the *primary visual cortex*, or *area V1*, forming a myelinated stripe that can be seen in cross section by the naked eye (i.e., *stria of Gennari*). This is why V1 is also called *striate cortex*. Intensive processing and local feature analysis occurs within V1, which then sends outputs to extrastriate visual areas V2, V3, and V4 as well as the middle temporal area (MT) for further analysis (Figure 1.7B). Two major pathways can be identified in the visual cortex: a dorsal pathway that projects from the early visual cortex toward the parietal lobe and a ventral pathway that projects toward the ventral temporal cortex. While the dorsal pathway is important for spatial processing, eye movement control, and supporting visually guided actions, the ventral pathway has a critical role in visual object recognition.

The patterns of activity that are evoked by a stimulus at each level of this network can be considered a neural representation of that stimulus, and the changes in stimulus's representation across successive stages can be understood as a series of nonlinear transformations that are applied to that initial stimulus input. That said, feedback projections are just as prominent as the feedforward connections between most any two visual areas, so visual processing is not strictly feedforward or hierarchical, but rather bidirectional and subject to top-down influences from higher cortical areas.

Retina

The retina can be thought of as a multilayered sheet that lies on the rear interior surface of the eye (G. D. Field & Chichilnisky, 2007; Masland, 2012). *Photoreceptors* form the outer layer of the retina, which, curiously, lies farthest from the light source (Figure 1.8). Each photoreceptor signals the amount of light (or dark) it is receiving by modulating the amount of glutamate that is released onto *bipolar cells* in the middle layer of the retina. Bipolar cells, in turn, project to *retinal ganglion cells* that form the inner layer of the retina. These ganglion cells provide the output signal from the retina, with a large axonal bundle that exits the *optic disk* (i.e., blind spot) and projects to the lateral geniculate nucleus.

Embedded among the photoreceptors and bipolar neurons are *horizontal cells*, which provide a form of lateral inhibition to enhance the contrast sensitivity of retinal processing. *Amacrine cells* are interspersed among the bipolar neurons and ganglion cells and strongly contribute to the center-surround receptive field organization of the ganglion cells.

Although curved in structure, the retina is better understood in terms of its two-dimensional organization. In essence, the retina forms a 2D map that registers patterns of light from the environment, preserving their spatial geometry as light passes through the pupil. High-acuity vision depends on *cone photoreceptors*, which are most densely packed at the center of the visual field, or *fovea*. The concentration of cones steadily declines as a function of *eccentricity*, or distance from the fovea. When considering the retina's 2D layout, it is more useful to consider its *retinotopic organization* in terms of *eccentricity* and *polar angle* (Figure 1.7B) instead of Cartesian (x, y) coordinates.

Cone photoreceptors support our ability to perceive color and fine spatial detail under

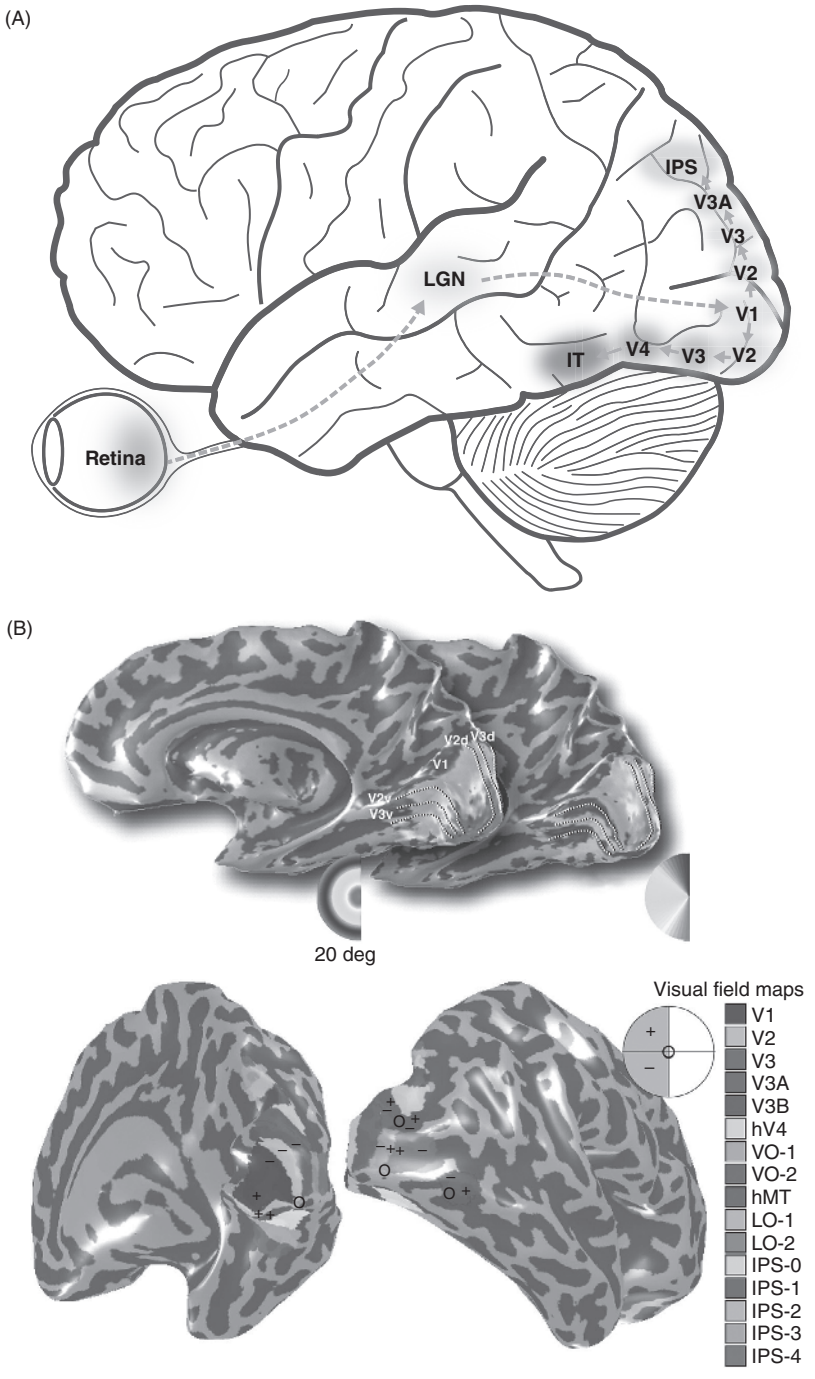

Figure 1.7 Hierarchical organization of the visual system. (A) Schematic illustration of the human visual system, with projections from retina to the LGN to primary visual cortex. From V1, projections along the ventral visual pathway ultimately lead to the inferotemporal cortex (IT), while the dorsal pathway projects toward the parietal lobe and regions in the intraparietal sulcus (IPS). (B) Retinotopic organization of the human visual system. Colors show cortical responses to changes in eccentricity and polar angle across the visual field. Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.

Source: (A) Figure created by Frank Tong; used with permission from the author. (B) From Wandell et al. (2007, pp. 368, 371). Reproduced with permission of Elsevier.

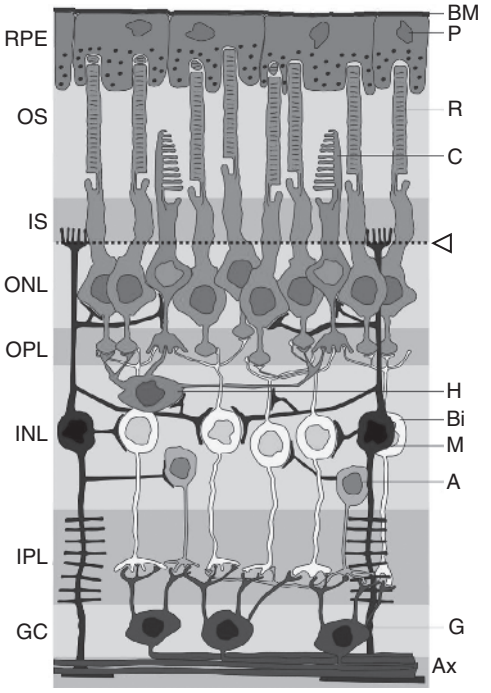

Figure 1.8 Diagram illustrating a cross section of the retina. This illustration depicts rod (R) and cone (C) photoreceptors, bipolar neurons (Bi), horizontal cells (H), amacrine cells (A), and retinal ganglion cells (RGC) with axons projecting ultimately toward the optic disk. Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.
SOURCE: From Wikimedia commons. Retrieved from https://commons.wikimedia.org/wiki/File:Retina_layers.svg

well-lit or *phototopic* viewing conditions, when reliable high-resolution spatial processing won't be limited by the amount of available light. Individual cones can genetically express one of three types of *photopsins*, which have different spectral sensitivities for long (L-cone), medium (M-cone), and short (S-cone) wavelengths of light. These roughly correspond to our ability to perceive the red, green, and blue portions of the visible color spectrum (see Chapter 3 in this volume for more on color vision).

Rod photoreceptors support low-resolution monochromatic vision in *scotopic* viewing conditions (i.e., when cones are no longer active), because of their exquisite sensitivity to very low levels of light. A single photon of light is capable of modifying the configuration of *rhodopsin*, the light-sensitive molecule contained in rods. This, in turn, leads to a cascade of molecular events that can affect hundreds of downstream molecules through a process of amplification, ultimately modifying the electrical current of the rod photoreceptor. While there are no rods in the fovea, in the periphery, rods greatly outnumber the cones.

Both rods and cones provide a continuous analog signal of the local level of light. In fact, photoreceptors remain continually active in the dark (sometimes called *dark current*), releasing glutamate steadily, and are hyperpolarized by the presentation of light. What functional advantage might this serve? This counterintuitive coding scheme ensures that rod photoreceptors can register the appearance of even very low levels of light by decreasing their rate of glutamate release. Recall that following dark adaptation, human observers appear sensitive to even single-photon events. This coding scheme also means that daylight conditions will effectively bleach the rods, so they remain in a continuous state of hyperpolarization. This is helpful and efficient, since the downstream activity of bipolar and ganglion cells will be exclusively dominated by cone activity.

Individual bipolar neurons are either excited or inhibited by the glutamate released from innervating cone photoreceptors, resulting in a preference for either dark or light in the center of their receptive field. In the fovea, it is common for bipolar cells to receive driving input from just a single cone, and to project to just a single ganglion cell. Thus, the number of cone photoreceptors that ultimately converge upon a ganglion cell's receptive field center can be as low as 1:1. Such a low *convergence ratio* from

photoreceptor to ganglion cell provides the foundation for high acuity vision in the fovea. This can be contrasted with an estimated convergence ratio of 1500:1 from rod photoreceptors to ganglion cells.

The receptive fields of ganglion cells are roughly circular in shape, with a central region that prefers light and a surround that prefers dark (i.e., *on-center off-surround receptive field*) or a central region that prefers dark and surround that prefers light (i.e., *off-center on-surround receptive field*). The receptive field structure of ganglion cells can be well described by a *difference of Gaussians (DoG) model*, as illustrated in Figure 1.9. A ganglion cell with an on-center off-surround can be characterized by the linear sum of a sharply tuned excitatory center and a broadly tuned inhibitory surround. The DoG model provides an excellent quantitative fit of the spatial frequency tuning properties of retinal ganglion cells, such as the X-cells of the cat retina as was described in the pioneering work of Enroth-Cugell and Robson (1966).

That said, the standard textbook portrayal of retinal ganglion cells tends to oversimplify their receptive field structure as being perfectly circular and nonoriented. A large number of retinal ganglion cells have elongated visual receptive fields that exhibit some degree of orientation bias, which can arise from their elongated dendritic fields. These elongations or deviations from circularity tend to be more prominent for orientations that radiate outward from the fovea (Schall, Perry, & Leventhal, 1986). These modest orientation biases, found in retinal ganglion cells, are strongly predictive of the orientation bias found in downstream LGN neurons (Suematsu, Naito, Miyoshi, Sawai, & Sato, 2013). At present, we do not know whether this heterogeneity and bias in the retina and LGN represent nuisance variables that must simply be ignored, or whether they directly contribute to development of orientation selectivity in V1.

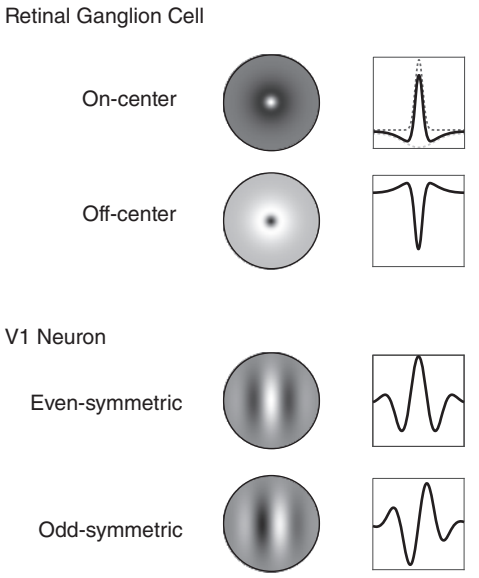

Figure 1.9 Examples of visual receptive fields in the retina and V1. This illustration shows the idealized receptive field structure of retinal ganglion cells (RGC) with either on-center or off-center organization. The 1D response profile of the on-center RGC arises from the linear sum of an excitatory center (red) and an inhibitory surround (blue). The receptive field tuning of V1 neurons can be modeled using even- and odd-symmetric Gabor functions, with their 1D profile shown to the right. Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.
SOURCE: Figure created by Frank Tong; used with permission of the author.

Magnocellular, Parvocellular, and Koniocellular Pathways

Recent studies suggest that there are about 20 different kinds of ganglion cells that tile the retina. The connectivity and function of many of these cell types remain to be determined (Masland, 2012). Arguably, each of these ganglion cell outputs could be described as its own specialized signal or channel. For our purposes, we will emphasize three major

pathways of the early visual system: the magnocellular (M), parvocellular (P), and koniocellular (K) pathways. These pathways are relayed through distinct layers of the LGN, and their anatomical segregation in the LGN has greatly facilitated their study (Casagrande & Xu, 2004).

The *magnocellular* (M) pathway supports the rapid temporal processing of transient visual events and motion but with coarser spatial resolution, whereas the *parvocellular* (P) pathway supports slower sustained processing of fine spatial detail and color information. This trade-off between temporal and spatial resolution suggests that the visual system evolved two parallel pathways for optimizing sensitivity. If the magnocellular system evolved to process rapidly changing light levels, then there will be minimal opportunity to pool visual signals over time, so integrating signals over larger regions of space is needed to improve the signal-to-noise ratio of visual processing. Higher resolution processing of static stimuli can likewise be achieved by pooling signals over time.

Magnocellular neurons in the LGN have large cell bodies and receive inputs from large, fast-conducting retinal ganglion cells, called *parasol cells*. Each parasol cell receives converging input from a fairly large number of L and M cones, leading to coarser spatial tuning and poor chromatic sensitivity. Assuming that individual parasol cells sample from local L and M cones in a fairly random way, then most of these neurons would be expected to lack strong chromatic bias.

Parvocellular LGN neurons receive their inputs from *midget cells* in the retina, which have smaller cell bodies and much smaller dendritic fields than parasol cells. In the fovea, the excitatory center of a midget cell may receive input from only a single L- or M-cone photoreceptor, allowing for both high spatial acuity and strong chromatic preference. Like all ganglion cells, midget cells become progressively larger in the periphery, integrating information from a larger number of cone photoreceptors. Although midget cells have modest tendency to sample preferentially from either L cones or M cones (G. D. Field et al., 2010), this nonrandom bias is quite weak, which may help explain why color perception is less precise in the periphery.

The *koniocellular* (K) pathway is anatomically distinct from the M and P pathways and has a specialized functional role in processing signals originating from S-cone photoreceptors. S cones comprise only ∼10% of the cones in the human retina, and project to their own specialized classes of bipolar cells and ganglion cells. These, in turn, project to the interstitial layers of the LGN.

Lateral Geniculate Nucleus

The LGN consists of multiple functional layers that each contain a complete retinotopic map of the contralateral hemifield. Layers 1 and 2 of the LGN consist of magnocellular neurons that receive their respective input from the contralateral eye and ipsilateral eye, whereas layers 3–6 consist of parvocellular neurons that receive contralateral or ipsilateral input. Between each of these M/P layers lies an interstitial layer of koniocellular neurons, whose very small cell bodies led to difficulties in detection in early anatomical studies.

These ganglion cell inputs synapse onto LGN relay neurons, which primarily project to area V1 in primates. Although the LGN has traditionally been considered just a simple relay nucleus, there is growing evidence of its role in aspects of perceptual processing as well as attentional modulation. LGN neurons show evidence of adaptation to high levels of stimulus contrast over time, and also exhibit a considerable degree of surround suppression.

Some researchers have emphasized that such modulatory effects are due to retinal mechanisms, whereas others have proposed the importance of feedback from V1 to LGN (Sillito, Cudeiro, & Jones, 2006; Alitto & Usrey, 2008; Jones et al., 2012; Usrey & Alitto, 2015).

Just as some orientation bias can be observed in retinal ganglion cells, LGN neurons can exhibit a modest but reliable orientation bias. Moreover, this bias tends to be correlated with the orientation preference of innervating retinal ganglion cells (Suematsu et al., 2013). Intriguingly, feedback projections from V1 to LGN have an oriented spatial structure that matches the tuning preference of the V1 neurons providing feedback (W. Wang, Jones, Andolina, Salt, & Sillito, 2006), suggesting that feedback from V1 to LGN may serve to modulate the efficacy of the orientation signals that V1 ultimately receives (Andolina, Jones, Wang, & Sillito, 2007). Modest orientation selectivity has also been demonstrated in neuroimaging studies of the human LGN (Ling, Pratte, & Tong, 2015). It remains to be seen whether the orientation bias of LGN neurons directly contributes to the orientation selectivity of V1 neurons. Advances in two-photon calcium imaging in rodent models will help inform our understanding of the basis of V1 orientation selectivity, as the activity of hundreds or thousands of synaptic boutons can be concurrently monitored (Kondo & Ohki, 2016; Lien & Scanziani, 2013; Sun, Tan, Mensh, & Ji, 2016). That said, direct characterization of orientation mechanisms in primates will still be essential.

There is considerable top-down feedback from V1 to the LGN, both directly and via the thalamic reticular nucleus, which may modify both the gain and the timing of spiking activity in the LGN. Shifts of covert attention can modulate LGN responses in both monkeys and humans. Single-unit studies in monkeys have found that spatial attention can boost the responsiveness of LGN neurons (McAlonan, Cavanaugh, & Wurtz, 2008) and enhance the synaptic efficacy of spikes transmitted from LGN to V1 (Briggs, Mangun, & Usrey, 2013). Human neuroimaging studies have likewise found spatially specific influences of attention in the LGN (Schneider & Kastner, 2009), as well as modulations of orientation-selective responses (Ling et al., 2015).

Primary Visual Cortex (V1)

The primary visual cortex provides a detailed analysis of the local features in the visual scene. Visual signals travel from the retina to the LGN, which in turn projects to V1 via what is known as the *retinogeniculostriate pathway*. This pathway is far more prominent in primates than in lower mammals, which is why V1 lesions in humans lead to much more severe deficits. Patients with V1 damage typically report a lack of visual awareness in the damaged part of their visual field. Some patients show some residual visual function despite this lack of reported awareness, a neuropsychological impairment that is called *blindsight* (Stoerig, 2006).

From the LGN, parvocellular and magnocellular neurons project to different sublayers of layer 4 of V1, whereas koniocellular neurons have a strong direct projection to layers 1 and 3. Feedforward inputs to V1 are also highly structured in terms of their topography. At the most global level, V1 is retinotopically organized according to eccentricity and polar angle (see Figure 1.7B), with the foveal representation near the occipital pole and more eccentric regions lying more anteriorly. Projections from LGN to V1 are also organized by eye of origin, leading to the formation of *ocular dominance columns*. These alternating monocular columns, each about 1 mm thick in humans, give rise to a striped pattern across the cortical sheet.

Such columns have been successfully mapped in humans using high-resolution fMRI (functional magnetic resonance imaging; Figure 1.10). At finer spatial scales, orientation columns and pinwheel structures can also be observed in the primary visual cortex of human (Figure 1.10C) and nonhuman primates (Obermayer & Blasdel, 1993; Yacoub, Harel, & Ugurbil, 2008). Orientation domains have also been successfully mapped in the extrastriate visual areas of monkeys using invasive imaging methods. Some have suggested that ocular dominance columns may provide the necessary scaffolding for the functional organization of binocular processing of disparity information. Curiously, however, not all monkeys show evidence of ocular dominance columns (Adams & Horton, 2003).

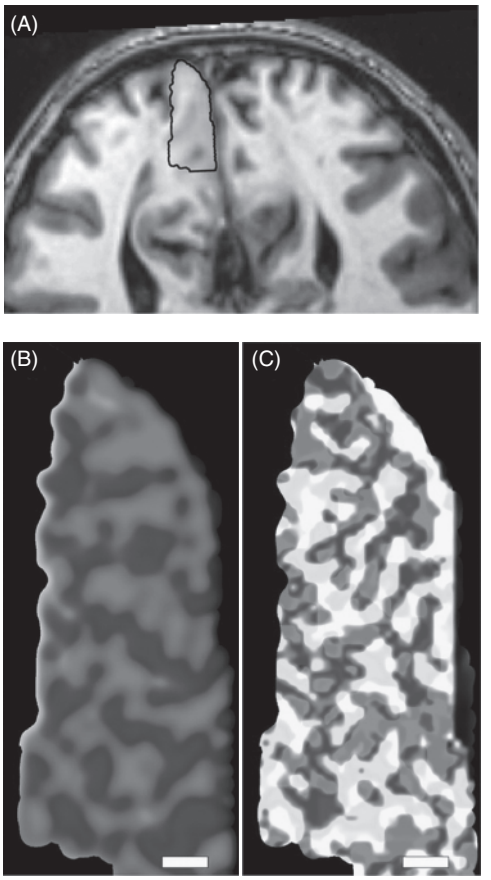

Figure 1.10 Ocular dominance and orientation columns in human V1. High-resolution fMRI of the human primary visual cortex (A) reveals the presence of ocular dominance columns (B) and evidence of columnar orientation structures (C). Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.
SOURCE: From Yacoub, Harel, and Ugurbil (2008, p. 10608). Copyright 2008 National Academy of Sciences, USA. Reproduced with permission of PNAS.

Efficient Coding Hypothesis

Much of our current understanding of neural coding can be traced back to early advances in vision research, including the seminal contributions of Horace Barlow, David Hubel, and Torsten Wiesel. When Hubel and Wiesel first planted their electrodes in area V1 of the cat, it was akin to entering *terra incognita*. V1 neurons were far more quiet—almost eerily silent—in comparison to earlier attempts to record spiking activity from the LGN or from retinal ganglion cells (Kuffler, 1953).

Why was the case? According to Barlow's (1961) *efficient coding hypothesis*, the goal of the visual system is to reduce any redundancies that exist in the natural sensory input by learning a *sparse efficient neural code*. A sparse code would require fewer spikes to *encode* the information contained in natural images commonly encountered in the environment, thereby improving the efficiency of information transmission. If natural images contain regular predictable structure (i.e., redundancy), then a more efficient code is achievable. One example of redundancy is the fact that neighboring photoreceptors usually receive similar levels of light, so their activity level is highly correlated. The center-surround organization of retinal ganglion cells serves to reduce this local redundancy to some extent.

The response tuning of V1 neurons is even more sparse and efficient. Compared to the number of retinal ganglion cells (~1 million per eye), there are far more neurons in V1 (~140 million), leading to gross oversampling of the retinal array. However, the percentage of V1 neurons that respond to any given natural image, selected at random, is much smaller than the percentage of active ganglion cells in the retina. Both computational and neurophysiological studies provide support for the proposal that V1 neurons provide a sparse efficient code for processing natural images (D. J. Field, 1987; Olshausen & Field, 1996; Vinje & Gallant, 2000).

Orientation Selectivity and the Excitatory Convergence Model

It is now part of neuroscience lore that orientation selectivity was discovered when Hubel and Wiesel accidentally triggered a V1 neuron to fire. After weeks of trying to evoke neuronal responses using projected slide images of simple round dots, a shadowy line cast by the edge of the glass slide happened to drift across the cell's receptive field at just the right orientation (Hubel, 1982). By carefully mapping the receptive-field properties of that cell and many others, they discovered the sparse feature tuning of V1 neurons as well as evidence of a hierarchical organization (Hubel & Wiesel, 1962).

One class of neurons, called *simple cells*, have a simple elongated receptive field, with *on*-regions that responded positively to the presentation of light and flanking *off*-regions that were inhibited by light. (Off-regions would also respond positively to a dark bar presented against a gray background.) Hubel and Wiesel proposed an excitatory convergence model to explain the phase-specific orientation selectivity of these neurons, which have clearly demarcated on- and off-regions. This model assumes that each simple cell pools the excitatory input from multiple LGN neurons whose circular receptive fields form an elongated receptive field (Figure 1.11).

In contrast, *complex cells* exhibit positive responses to a preferred orientation presented anywhere within their excitatory receptive field. The positional invariance of this selectivity was noteworthy because it provided novel evidence that neurons are capable of some form of abstraction. The researchers went on to speculate that this process of generalization could be important for form perception. If many of these complex cells projected to a common cell of higher order, that neuron might tolerate even greater transformations of an image while

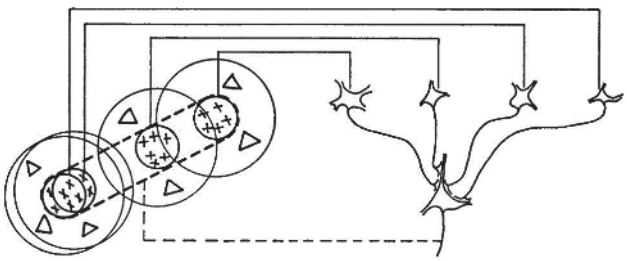

Figure 1.11 Hubel and Wiesel's proposed model of a V1 simple cell. Hubel and Wiesel proposed excitatory feedforward convergence model to account for the orientation selectivity of V1 simple cells. This cell has an on-center and off-surround, based on the summation of inputs from a series of LGN neurons with collinearly organized on-center receptive fields.
SOURCE: From Hubel and Wiesel (1968).

maintaining its selectivity. The response of a complex cell can be modeled by assuming that it receives excitatory input from multiple orientation-tuned simple cells with slightly shifted receptive fields, such that excitation from any one of these simple cells will evoke an action potential. As we will later see, this proposed architecture for simple cells and complex cells has helped to inform the design of neural networks for object processing.

Although the orientation-selective properties of V1 were discovered over 50 years ago, scientists are still striving to determine the precise nature of the neuronal circuit that gives rise to this sharp visual tuning. Individual simple cells in the feedforward input layer of V1 retain their strong orientation selectivity after V1 is cooled or silenced, implying that the sum of excitatory inputs from LGN to V1 is enough to create this oriented receptive-field structure (Ferster, Chung, & Wheat, 1996; Priebe & Ferster, 2008). Recent studies have also discovered that the neuronal projections from the LGN to layer 4 of V1 are highly structured in terms of spatial phase (Y. Wang et al., 2015), leading to a consistent overlap of on- and off-regions among neighboring simple cells in layer 4. The orientation of these elongated on- and off-regions reliably predicts the orientation preference of neurons in other layers of the same cortical column, suggesting that it determines the broader organization of the cortical orientation map.

Such findings are consistent with the predictions of the excitatory convergence model, in which multiple LGN neurons that prefer a common polarity in their center (i.e., light or dark) form an elongated region (Figure 1.11). However, an alternative theory is that each V1 simple cell receives excitatory inputs from a pair of LGN neurons with different polarities, one with an on-center receptive field that partially overlaps with an off-center receptive field (Paik & Ringach, 2011). The combination of these two LGN inputs would lead to an oriented band that prefers light and an adjacent band that prefers dark, a prediction that has received support in recent patch-clamp recordings from neurons in the input layer of the mouse visual cortex (Lien & Scanziani, 2013). However, the presence of coarse orientation selectivity in the retina and LGN presents a more complex picture of how sharp orientation selectivity is achieved in V1. With recent advances in adaptive optics, researchers can concurrently image the calcium-based activity of thousands of thalamic boutons arriving in layer 4 of mouse V1. A large percentage of thalamic boutons show some degree of orientation selectivity (Kondo & Ohki, 2016; Sun et al., 2016), raising the possibility that these coarsely tuned inputs may also contribute to the oriented structure of V1 receptive fields (Vidyasagar & Eysel, 2015).

Extrastriate Visual Areas

At higher levels of the visual pathway, neurons have larger receptive fields and more complex tuning properties. The progressive increase in receptive field size can be understood if each V2 neuron receives inputs from a local distribution of retinotopically organized V1 neurons, leading to a broader spread of retinotopic inputs to that V2 neuron. Likewise, if a V3 neuron samples from a local distribution of V2 neurons, then that V3 neuron will also have a larger receptive field than the V2 neurons from which it samples.

This progressive increase in receptive field size, when ascending the visual hierarchy, is accompanied by an increase in neuronal tuning complexity. Consider what will happen if a V2 neuron happens to integrate signals from a small pool of V1 neurons that prefer more than one orientation, spatial location, or spatial frequency. Single-unit recordings

have found that some V2 neurons prefer combinations of orientation elements, such as curved spiral or hyperbolic gratings, or sharp angles formed by abutting line orientations (Hegde & Van Essen, 2000). In this respect, V2 neurons may attain some sensitivity to the higher order relationship between orientations, an important step toward the encoding of shape.

Neurons in V4 exhibit even more heterogeneous response preferences to variations in orientation and spatial frequency, and are less well activated by simple lines or gratings. There is strong evidence that V4 neurons are sensitive to local aspects of visual shape. By presenting a variety of 2D shapes to a V4 neuron, it is possible to map the neuron's tuning preference for curvature at different positions in the receptive field (Pasupathy & Connor, 2002). A neuron might prefer a sharp convexity at one location or a moderate degree of concavity in another location. The combination of many curvature computations across an object would provide a useful code to define the shape of that object.

Area V4, or the cortical region just anterior to it, has also been implicated in color perception and color constancy. Damage around this cortical region is strongly associated with achromatopsia, that is, severe deficits in visual color perception (Bouvier & Engel, 2006). Such deficits are often restricted to just a hemifield or quarter visual field. Human neuroimaging studies have implicated the role of V4 as well as the more anterior region VO1 (Figure 1.7B) in color perception and the perception of color aftereffects (Brouwer & Heeger, 2009; Hadjikhani, Liu, Dale, Cavanagh, & Tootell, 1998).

Higher Order Visual Areas

Beyond the early visual areas (V1–V4), a series of higher order visual areas extend along the dorsal pathway toward the parietal lobe and along the ventral pathway toward the ventral temporal lobe. The effects of brain injury to these separate pathways have revealed striking dissociations of function (Farah, 2004; Goodale & Westwood, 2004). Damage to the posterior parietal lobe can sometimes lead to impairments in the ability to perform visually guided actions, what is known as *optic ataxia*. In other cases, it can disrupt the patient's ability to attentionally orient to stimuli in the contralesional region of visual space, what is known as *visual neglect*. This can be contrasted with damage to higher visual areas along the ventral pathway, which can lead to impairments in shape perception and object recognition. *Apperceptive agnosia* or impairments in shape perception often results from damage to the lateral occipital cortex, whereas damage to the ventral temporal cortex can lead to *object agnosia*, in which object recognition is impaired while the perception of basic shape information remains intact.

Retinotopic mapping has revealed several distinct visual areas in the parietal and occipitotemporal regions of the human visual system (Figure 1.7B). Distinct visual areas have also been identified in the parietal and temporal lobes of the macaque monkey, but in most cases it remains unclear as to which areas are directly homologous with those found in humans (Orban, Van Essen, & Vanduffel, 2004). The human parietal lobe contains multiple visual areas in the intraparietal sulcus, including areas IPS1, IPS2, IPS3, and IPS4 (Swisher, Halko, Merabet, McMains, & Somers, 2007). These parietal areas are sensitive to visual stimulation, shifts of attention, and planned eye movements to target locations (M. A. Silver & Kastner, 2009). In the ventral occipitotemporal cortex, multiple category-selective regions have been identified (Op de Beeck, Haushofer, & Kanwisher, 2008), as well as large expansive

regions that are generally sensitive to a variety of object stimuli (Grill-Spector & Weiner, 2014; Kriegeskorte et al., 2008). Retinotopic mapping has helped identify areas *LO1* and *LO2*, which lie in the lateral occipital cortex posterior to area MT (Larsson & Heeger, 2006). These regions are involved in earlier stages of object processing and respond preferentially to intact objects as compared to scrambled stimuli.

In the ventral temporal cortex, several category-selective areas can be found. These include the *fusiform face area* (FFA), which responds preferentially to face stimuli, and the *parahippocampal place area* (PPA), which responds preferentially to buildings, landmarks, and indoor and outdoor scenes (Epstein & Kanwisher, 1998; Kanwisher, McDermott, & Chun, 1997). On the lateral occipital surface, identified regions include the *occipital face area* and an adjacent region called the *extrastriate body area* (Downing, Jiang, Shuman, & Kanwisher, 2001). An ongoing point of discussion concerns whether the response selectivity of these brain regions, and ultimately their underlying function, can be best understood as category-selective or continuous representations of the visual-semantic properties of objects (Haxby et al., 2001; Huth, Nishimoto, Vu, & Gallant, 2012; Kriegeskorte et al., 2008; Op de Beeck et al., 2008; Weiner & Grill-Spector, 2012). With respect to this debate, it is intriguing that transcranial magnetic stimulation applied to different regions of the lateral occipital cortex can selectively impair people's ability to discriminate faces, human bodies, and 3D rendered objects (Pitcher, Charles, Devlin, Walsh, & Duchaine, 2009). That said, selective effects of disruption cannot fully establish whether the underlying representations of these stimuli are categorical or continuous in nature. (For further discussion see Chapter 8 on visual object recognition in this volume.)

MECHANISMS UNDERLYING VISUAL PERCEPTION

How does the human brain perceive basic visual properties, such as the orientation, color, or motion of a stimulus? What types of processes and neural computations are required to transform the incoming patterns of light signals into the basic qualities of our perceptions? Vision scientists have brought to bear a variety of techniques and approaches to address these challenges, including visual psychophysics, neurophysiological recordings, human neuroimaging, and computational modeling. From this work, we will see how the perception of basic visual properties is strongly linked to information processing at early stages of the visual pathway.

Visual Feature Perception

An important advance in vision science was the realization that the early stages of perceptual processing could be described by mathematical concepts such as Fourier analysis and spatial-temporal filters. Our ability to detect and discriminate simple visual patterns depends on the spectral contents of the stimulus and its match to the tuning properties of our visual system. For example, visual sensitivity at detecting a square-wave grating can be predicted by one's sensitivity to the Fourier components that comprise that grating (Campbell & Robson, 1968). Likewise, perception of motion can be described in terms of *spatiotemporal energy detectors*, or "oriented" filters in space-time (Adelson & Bergen, 1985). Once conceptualized in this way, one can quantify the motion energy that would result from any succession of images or from simple two-frame apparent motion displays. (See Chapter 5 in this volume for more on motion perception.)

Since neurons in the retina and the early visual cortex have small receptive fields, the analysis they perform is better understood as a local analysis rather than a spatially unrestricted Fourier analysis. Indeed, a 2D Gabor filter, which provides a good approximation of the tuning properties of V1 simple cells, is mathematically equivalent to a sine-wave function that is spatially restricted within a Gaussian window (Figure 1.9). This partly explains how V1 neurons provide a sparse efficient visual code for the natural images we commonly encounter in the environment (D. J. Field, 1987; Olshausen & Field, 1996).

Although spatial-temporal filter models are effective and widely applicable, it is important to keep in mind that they rely on simplifying assumptions that may not fully capture the complexities of human visual processing. For instance, visual sensitivity to oriented patterns is not uniform; people are better at detecting and discriminating orientations that are near cardinal as compared to those that are oblique (Appelle, 1972; Westheimer, 2003). Also, our perception of a stimulus does not arise from strictly local visual processing—the surrounding visual context can have a strong influence. For example, a central vertical grating surrounded by a tilted grating will appear somewhat tilted in the opposite direction, a phenomenon known as the tilt illusion (Wenderoth & Johnstone, 1988). Likewise, the perception of a compound stimulus may not necessarily be explained by the linear sum of its parts. The perception of a moving plaid, consisting of two superimposed gratings drifting in different directions, can deviate greatly from vector average of their individual motions (Adelson & Movshon, 1982).

Finally, it should be emphasized that perceptual sensitivity is not determined by the information encoded in individual neurons, but rather, by the information that can be extracted or "decoded" from a population of visual neurons, to support a perceptual decision. This concept is described as *population coding*. Indeed, computational models have been developed to characterize how a small population of feature-tuned neurons can jointly encode information about a particular stimulus (Pouget, Dayan, & Zemel, 2003). Some models, for example, rely on Poisson-process neurons tuned to different feature values, by specifying how strongly each neuron will respond on average to any given orientation (Figure 1.12). One can then apply a Bayesian estimation approach

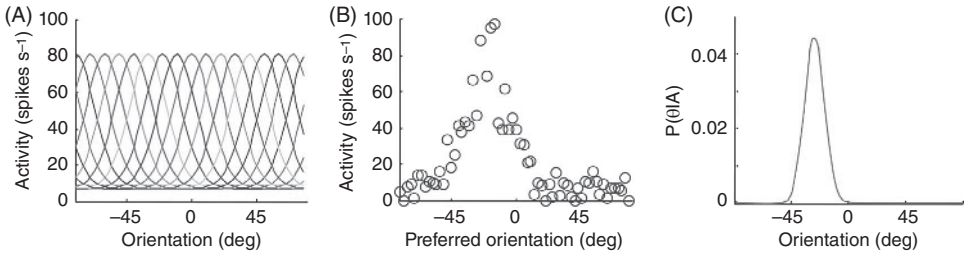

Figure 1.12 Example of a population-coding model with multiple orientation-tuned units. (A) Tuning curves show the average firing rate of each unit to a given orientation. (B) The number of spikes emitted by each tuned neuron is somewhat variable, due to presumed Poisson-process spiking activity. (C) Bayesian estimation can then be used to decode what is the most likely stimulus to have occurred given the observed number of spikes (i.e., the data). Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.
SOURCE: From Knill and Pouet (2004). Reproduced with permission of Elsevier.

to decode the stimulus orientation. Given the number of spikes emitted by each of the neurons in the population, it is possible to determine what is the most likely stimulus that could have evoked the observed response.

Neural Bases of Visual Feature Perception

Multiple lines of evidence suggest that our ability to detect and to discriminate basic visual features depends on processing that takes place in the early visual areas. From psychophysical studies, we know that prolonged monocular adaptation to an oriented grating or to drifting random dots will produce a stronger visual aftereffect (e.g., tilt or motion aftereffect) if the subsequent test stimulus is presented to the same eye; a reduced aftereffect is observed if the test stimulus is presented to the corresponding location of the fellow eye. This implies that the activity of monocular neurons, presumably those in V1, contributes to these visual aftereffects to some extent (Blake, Overton, & Lema-Stern, 1981).

Human fMRI studies also support the notion that the perception of basic features is strongly associated with visual processing in early visual areas. For example, an early study found greater sustained activity in motion-sensitive area MT+ when observers experienced a motion aftereffect while viewing a static test pattern (Tootell et al., 1995). Sensitivity to changes in visual contrast has also been linked to fMRI measures of the contrast-response function in area V1 (Boynton, Demb, Glover, & Heeger, 1999). Neuroimaging studies of binocular rivalry provided some of the first evidence to link the activity of cortical visual areas, including V1, to spontaneous fluctuations of conscious perception (Polonsky, Blake, Braun, & Heeger, 2000; Tong & Engel, 2001; Tong, Nakayama, Vaughan, & Kanwisher, 1998).

A similar correspondence between cortical activity and conscious perception has been observed in threshold detection tasks. Greater activity was observed in areas V1–V3 when an observer successfully detected the presentation of a very low contrast grating as compared to when it was missed, and remarkably, activity is also greater on false alarm trials when observers mistakenly report "target present" when the grating was in fact absent (Ress & Heeger, 2003).

The development of fMRI decoding, or multivariate pattern analysis, has proven particularly useful for isolating feature-selective responses in the human visual cortex (Tong & Pratte, 2012). Kamitani and Tong discovered that activity patterns in early visual areas contain detailed information that can be used to reliably predict what stimulus orientation (Figure 13A and B) or motion direction is being viewed by the subject (Kamitani & Tong, 2005, 2006). Subsequent studies have shown how voxel-based encoding models can be used to quantify the feature-tuning preferences of individual voxels in the visual cortex, and how information from individually fitted voxels can likewise be pooled (Brouwer & Heeger, 2009; Kay, Naselaris, Prenger, & Gallant, 2008; Naselaris, Kay, Nishimoto, & Gallant, 2011; Serences, Saproo, Scolari, Ho, & Muftuler, 2009). Such approaches have been used to demonstrate compelling links between color perception and cortical responses in area V4 (Brouwer & Heeger, 2009) and to distinguish among hundreds of natural scene images (Kay et al., 2008). Researchers have also developed fMRI approaches to decode not only information about the perceived stimulus, but also, the degree of uncertainty associated with that perception. Recent work indicates that on trials in which participants exhibit greater perceptual error, greater uncertainty is evident in the cortical activity patterns of V1 (van Bergen, Ma, Pratte, & Jehee, 2015).

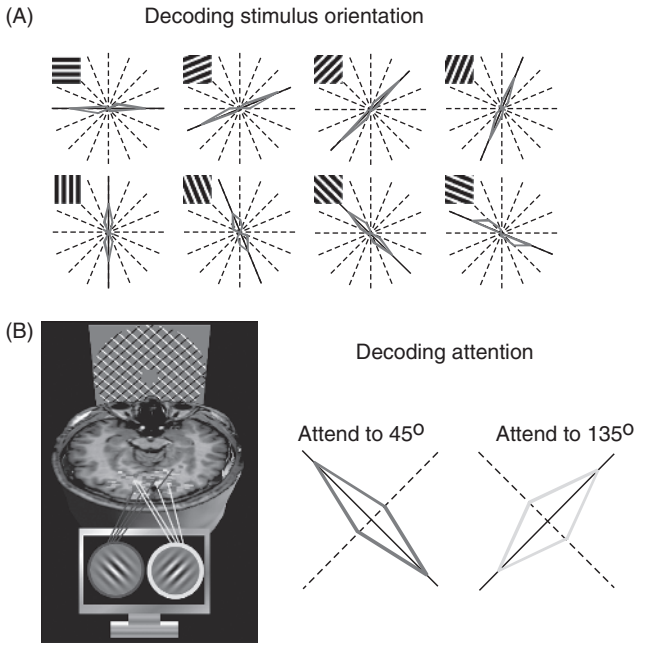

Figure 1.13 fMRI decoding of stimulus orientation and attended orientation. (A) This polar histogram shows the accuracy of decoding which of eight possible stimulus orientations is being viewed by an observer during each fMRI stimulus block. The true orientation is indicated by the thick black line, and the decoded orientation is shown in blue. (B) Orientation preferences of individual voxels in areas V1–V4 are illustrated here, and can be used to train a classifier on stimulus decoding or to decode which of two overlapping orientations is being covertly attended by the observer. (C) fMRI decoding can accurately predict the attended orientation, indicating that feature-based attention can alter orientation-selective responses in areas V1–V4. Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.
SOURCE: (A) Adapted from Kamitani and Tong (2005). (B) Figure created by Frank Tong; used with permission of the author.

To what extent might these visual representations be modified by extensive perceptual training with particular stimuli? In one study, observers underwent a month of training that required discriminating a small range of orientations in the left or right visual field (Jehee, Ling, Swisher, van Bergen, & Tong, 2012). Following this training, orientation responses in V1 were selectively enhanced for the trained orientation at the trained location. Moreover, the degree of cortical improvement was predictive of the degree of visual learning exhibited by each observer. Intriguingly, participants can even be trained, via neurofeedback, to enhance orientation-selective activity in V1 while they view a blank screen (Shibata, Watanabe, Sasaki, & Kawato, 2011). After neurofeedback training, participants showed better orientation discrimination performance for the so-called trained orientation that was never actually seen. These orientation-specific effects of learning concur with neurophysiological studies in monkeys, who after learning showed a greater preponderance of V1 neurons with tuning curves that flanked the trained orientation (Schoups, Vogels, Qian, & Orban, 2001).

Visual Segmentation and Figure-Ground Perception

Whenever we look upon a visual scene, our visual system is challenged by a continuous array of light intensity values (e.g., Figure 1.3) that must somehow be carved up into meaningful entities and objects. Parsing a visual scene engages mechanisms of *visual segmentation* and *figure-ground processing*. Differences in luminance, color, orientation, spatial frequency, and stereo-depth all provide relevant cues for distinguishing an object from its background. In fact, models of visual saliency propose that local differences in feature content are calculated throughout the visual field (Itti & Koch, 2000; Li, 2002), and that this information can then be used to determine what portions of the scene may contain potential objects of interest.

Differences in luminance or color are readily detected because they create first-order edges that are registered by enhanced levels of activity, even at the level of the retina. However, local differences in orientation or spatial frequency content are trickier to compute, because they require higher order comparisons between the feature-selective responses of different populations of neurons. This depends on more sophisticated processing in early cortical visual areas.

One mechanism that contributes to visual segmentation is *orientation-selective surround suppression*. Neurophysiological studies have shown that a V1 neuron's response to a preferred orientation in its *classical receptive field* (CRF) can be strongly modulated by stimuli presented in its surround, outside of the CRF (Cavanaugh, Bair, & Movshon, 2002a, 2002b). In general, the presentation of any stimulus in the surround will lead to some degree of response suppression, but these suppressive interactions are much stronger if the orientation in the surround matches the orientation in the center. The modulatory effects of surround suppression can be well described by computational models that incorporate *divisive normalization*, in which feedforward responses to the stimulus in the CRF are reduced in a divisive manner by the activity level of neighboring neurons corresponding to the surround (Carandini & Heeger, 2012).

Neurophysiological studies in alert monkeys suggest that additional figure-ground processes may take place in area V1. Using displays such as those shown in Figure 1.14, researchers have found two types of modulatory responses to figure-ground displays: an edge enhancement effect and a figure enhancement effect (Zipser, Lamme, & Schiller, 1996). Responses are particularly strong at the boundaries between surfaces, regardless of whether those boundaries are defined by differences in color, orientation, or stereo-depth. This is consistent with feature-tuned effects of surround suppression. However, stronger V1 responses are also observed near the center of the figure, and these emerge well after the initial onset response. Evidence suggests that this figural enhancement in V1 arises from top-down feedback, as both anesthesia and lesions of the parietal lobe eliminate this modulatory effect (Lamme, Zipser, & Spekreijse, 1998). More recently, researchers have compared the timing of these figural enhancement effects across different levels of the visual hierarchy, finding that V4 is modulated about 40 ms earlier than V1, consistent with a feedback interpretation (Poort et al., 2012).

Area V2 appears to have a more elaborate role in figure-ground processing than V1, providing a code for the apparent depth relation that occurs at visual boundaries. A large percentage of V2 neurons respond differentially to the edge of a stimulus, in a manner that depends on whether that edge comprises the left or right side of the figure. Such preferences reflect a degree of abstraction

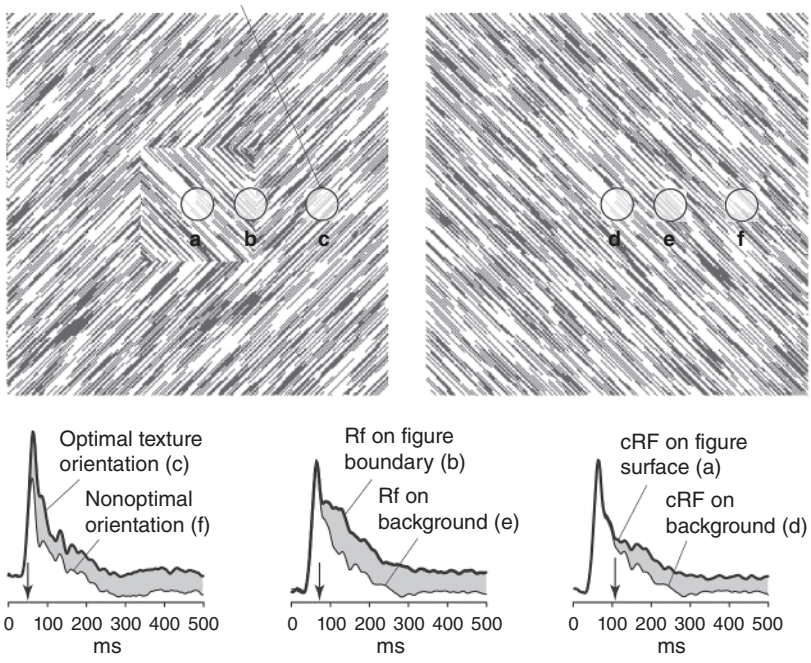

Figure 1.14 Enhanced responses to perceptual figures in V1. Effects of figural enhancement (a) and edge enhancement (b) in V1 in comparison to responses (d and e) to the same orientation in the central square region, but with a background of matching orientation. Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.
SOURCE: From Lamme and Roelfsema (2000). Reproduced with permission of Elsevier.

or invariance, as they remain consistent regardless of whether the edge is defined by pictorial or stereo-depth cues (Qiu & von der Heydt, 2005). This has led to the proposal that V2 neurons provide a visual code for *border ownership*, which serves to distinguish which portions of an encoded image belong to the figure and which portions belong to the background. Intriguingly, lesions applied to area V2 of the monkey do not affect basic visual acuity or ability to perceive the orientation of simple gratings, but severely impair the animal's ability to detect orientation-defined figures presented against a background of differing orientation (Merigan, Nealey, & Maunsell, 1993).

Although figure-ground processing has received limited attention in human studies (Scholte, Jolij, Fahrenfort, & Lamme, 2008), it has been shown that early visual areas exhibit strong effects of surround suppression (Zenger-Landolt & Heeger, 2003), including evidence of an orientation-selective component (McDonald, Seymour, Schira, Spehar, & Clifford, 2009). Salient figures defined by motion cues also lead to strong sustained responses in the figural region, corresponding to a lingering impression (or *hysteresis*) of the figural percept (Strother, Lavell, & Vilis, 2012). Recently, researchers had the opportunity to record spiking activity from area V2/V3 of a preoperative epilepsy patient, and found similar effects of figural enhancement for texture-defined figures (Self et al., 2016) as had been found in the monkey. These findings suggest a prevalent role for early visual areas in visual segmentation and figure-ground perception.

Effects of Visual Context

So far, we have learned that the processing of a stimulus does not occur in strict isolation, as neural responses are also influenced by stimuli in the surround. In some situations, the impact of visual context can be especially pronounced.

This is particularly evident in cases of perceptual filling-in, where observers are predisposed to infer the presence of a visual stimulus in regions that lack direct stimulation (Figures 1.6E and 1.6F). fMRI studies have tested for neural correlates of perceptual filling-in using displays that evoke the perception of subjective contours, neon-color spreading or visual phantoms (Meng, Remus, & Tong, 2005; Sasaki & Watanabe, 2004). These studies find evidence of enhanced activity in regions of V1 corresponding to the blank gap, even when attention is directed elsewhere, for example, by having participants perform a demanding task at fixation. High-resolution fMRI has suggested that the effects of filling-in may be more prominent in the deep layers of V1, as compared to the middle layer region that receives strong feedforward input (Kok, Bains, van Mourik, Norris, & de Lange, 2016). Neurophysiological recordings in alert monkeys have also found evidence of filling-in responses to subjective contours (Lee & Nguyen, 2001). Interestingly, these filling-in effects are observed at earlier in time in V2 than in V1. Although attention was not controlled in these monkeys, these results, in concert with the human neuroimaging studies, suggest that the top-down inferential processes of filling-in can occur in an automatic manner, without the benefit of focused attention.

What might be the origin of this top-down feedback? Although we do not know for sure, the lateral occipital complex, which has a central role in visual object processing, is a very strong candidate. This swath of cortical areas in the ventral pathway responds more strongly to intact than scrambled objects (Grill-Spector, Kourtzi, & Kanwisher, 2001), and has also been found to respond more strongly to inducers that evoke the perception of a subjective figure than to control stimuli that do not (Mendola, Dale, Fischl, Liu, & Tootell, 1999). Moreover, an fMRI study of monkeys and humans investigated what visual areas might be sensitive to collinear patterns embedded in an array of randomly oriented Gabor gratings. This study found enhanced activity in V1 and V2 to figures defined by collinearity, as well as strong enhancement in the lateral occipital complex, consistent with the potential role of the lateral occipital complex in top-down enhancement of local features represented in the early visual areas (Kourtzi, Tolias, Altmann, Augath, & Logothetis, 2003).

Researchers have also investigated the effects of visual context using more complex visual displays. When flickering checkerboard patterns are presented against a pictorial scene so that one checkerboard appears much farther away and perceptually larger than the other, the resulting illusion in perceived size is accompanied by a larger activated region in the primary visual cortex (Murray, Boyaci, & Kersten, 2006). Unlike the effects of perceptual filling-in, this size illusion effect in V1 appears to be modulated in strength by visual attention (Fang, Boyaci, Kersten, & Murray, 2008). Recordings in area V1 of monkeys have demonstrated a similar neural correlate of this size illusion, indicating the generality of these effects (Ni, Murray, & Horwitz, 2014). Presumably, processing in higher-level object areas would also be needed to interpret the complex visual scene and its portrayal of depth. This would imply that information pertaining to scene processing is fed back

to V1 representations of the target object, thereby modifying perception and associated neural responses.

VISUAL ATTENTION

Visual attention has been a longstanding area of inquiry in experimental psychology, highlighted by William James's (1890/1981) oft-quoted description of attention as "the taking possession of the mind, in clear and vivid form, of one out of what seems several simultaneously possible objects or trains of thought." Several ideas are evident here, including James's emphasis on voluntary control, selectivity, and the fact that this selection process leads to a clearer impression of the attended item. Although James proposed that attention can be focused either outwardly at an external object or inwardly at one's own thoughts, vision researchers have concentrated on the problem of how people attend to external visual stimuli.

In most social settings, we can tell where a person is attending by noting where their eyes are focused. *Overt attention* occurs when a person directly gazes at the object of interest, which leads to enhanced visual processing starting at the retina. The foveal region is overrepresented by the cones and even more so by the ganglion cells, such that any foveated stimulus will activate a much larger population of neurons in V1 and higher extrastriate areas, due to the greater *cortical magnification* of the central visual field. Moreover, a foveal stimulus will be better processed by high-level object areas such as the fusiform face area and the lateral occipital area, due to their overrepresentation of the central visual field for fine-grained object processing.

Psychologists and neuroscientists, however, are more interested in the perceptual and neural consequences of covert shifts of attention. *Covert attention* refers to attending to an object in the periphery, without moving the eyes or directly gazing at the attended item. If covert attention is capable of modifying the strength or fidelity of visual responses to a peripheral item, independent of any change in retinal stimulation, then such modulations would suggest the influence of top-down feedback that can flexibly modify the strength of feedforward responses.

Attention can also be distinguished according to whether it is guided by involuntary or voluntary factors (Posner, Snyder, & Davidson, 1980). *Exogenous attention* (or stimulus-driven attention) refers to the involuntary capture of attention. Stimuli that are bright, high contrast, colorful, or dynamic, and distinct from their surround, are more salient (Itti & Koch, 2000) and more likely to attract exogenous attention. However, our attention is not simply only governed by exogenous factors, or our attention would be forever capture by shiny salient objects like moths to the flame. *Endogenous attention* refers to the ability to shift attention in a voluntary manner, based on our top-down goals, such that we can seek out a particular target in a cluttered environment (see Chapter 6 in this volume, on visual search) or maintain attention on an object in the face of distraction.

In a typical study of exogenous attention, observers are instructed to maintain fixation while covert attention is manipulated by briefly presenting a peripheral cue to the left or right of fixation. This is shortly followed by a target, which appears at the same location as the cue on valid trials, or at a different location on invalid trials. Such experiments have revealed that exogenous attention operates quickly, transiently, and in a quite automatic manner (Nakayama & Mackeben, 1989; Posner et al., 1980).

If a valid peripheral cue appears 50–150 ms in advance of the target, participants will be

faster and more accurate at processing that target stimulus. Such facilitation occurs even if the exogenous cue does not reliably predict the target's location across trials, implying that observers tend to automatically shift attention to the exogenous peripheral cue on every trial. Consistent with this interpretation, invalid spatial cues usually lead to a behavioral cost, relative to a neutral cue (often consisting of cues at both possible target locations). These effects of exogenous cuing, though potent, are short-lived. If the target appears well over 200 ms after a valid cue, no benefit is observed, and performance may even be subtly impaired, a phenomenon sometimes described as *inhibition of return*.

With endogenous cuing, a symbolic cue, such as a letter (*L* or *R*), can be used to indicate the location to be attended. For the endogenous cue to influence performance on the task, it must be predictive of the location of the upcoming target at levels greater than chance, otherwise the observer will start to ignore these cues and focus exclusively on the target. Thus, processing of the endogenous cue is voluntary, and observers will take advantage of the cue only if it is informative. If the time between cue and target is too brief, however (i.e., less than ∼150 ms), observers will not have enough time to process the meaning of the cue and shift attention to the anticipated location of the target. Unlike exogenous attention, endogenous attention operates in a slower but sustained manner. Performance at a validly cued location is facilitated, even if the cue precedes the target by several seconds.

Psychophysical studies have revealed that covert attention reliably enhances the signal-to-noise ratio of visual processing in a manner that resembles increasing the physical contrast of the attended stimulus (Carrasco, Ling, & Read, 2004; Ling & Carrasco, 2006). Such effects can be observed with both exogenous and endogenous spatial cuing. Consistent with this idea, when covert attention is directed toward an adapting stimulus, the rate of neural adaptation is enhanced, leading to measurably stronger visual aftereffects (Alais & Blake, 1999; Chaudhuri, 1990).

Covert shifts of attention can also modify the spatial resolution of visual processing. Research suggests that exogenous cuing of attention improves the processing of high spatial frequency targets but also impairs the processing of low spatial frequency targets (Yeshurun & Carrasco, 1998). In comparison, endogenous attention tends to be more adaptive and flexible—observers are able to adopt an attentional template that matches properties of the task-relevant target (Carrasco, 2011).

Attentional Modulation of Neural Responses

Once thought to be rare and elusive, it is now known that the top-down effects of spatial attention are widespread and pervasive throughout the visual system. Neurophysiological, fMRI, and electroencephalography (EEG) studies demonstrate that attentional feedback can enhance visual responses to a task-relevant stimulus, while dampening responses to task-irrelevant stimuli. According to the *biased competition model* of attention, visual stimuli that appear concurrently, especially those in close proximity, will lead to competitive inhibitory interactions across multiple levels of the visual hierarchy (Desimone & Duncan, 1995). The role of top-down attention is to bias this competition in favor of the attended stimulus, which in turn will lead to greater suppression of the unattended stimulus.

In EEG studies of attention, stimuli are usually presented concurrently in the two hemifields while the observer is cued to attend selectively to stimuli on either side.

These studies find that attended stimuli evoke a stronger *P100 component* at contralateral occipital sites, compared to stimuli that are ignored (Heinze et al., 1994; Luck, Woodman, & Vogel, 2000). The P100 is the first positive visually evoked component, associated with processing in extrastriate visual areas. Although attending to a stimulus leads to faster behavioral response times, by 20–30 ms or so, attention modulates the amplitude but not the latency of the P100 response. Presumably, this boost in response amplitude at this earlier processing stage leads to a savings in processing time at later stages.

Human fMRI studies have also demonstrated powerful and spatially specific effects of attention particularly in retinotopic visual areas V1–V4 (Gandhi, Heeger, & Boynton, 1999; Somers, Dale, Seiffert, & Tootell, 1999). In fact, decoding of the activity patterns in retinotopic visual cortex can be used to reliably predict the spatial locus of attention under conditions of constant visual stimulation (Datta & DeYoe, 2009). Modulatory effects of attention have even been detected in the lateral geniculate nucleus (Ling et al., 2015; O'Connor, Fukui, Pinsk, & Kastner, 2002). Since there are no feedback connections from the LGN to the retina, such findings indicate that attentional feedback propagates to the earliest possible stage of visual processing.

The enhancement of visual responses by attention can be modeled by implementing some type of *gain modulation*, in which top-down feedback leads to amplification of the stimulus-driven response. In some cases, attention appears to enhance the contrast sensitivity of visual neurons, leading to a leftward shift in the contrast response function. However, in other situations, attention seems to lead to a multiplicative increase in the neural response across all contrast levels. These two types of gain modulation are known as *contrast gain* and *response gain*, respectively (see Figure 1.15).

Although attention can boost the gain of visual evoked responses, both neurophysiological and neuroimaging studies have shown that attending to a blank region of space leads to enhanced activity in corresponding retinotopic visual areas (Kastner, Pinsk, De Weerd, Desimone, & Ungerleider, 1999; Luck, Chelazzi, Hillyard, & Desimone, 1997). Thus, top-down feedback is capable of boosting both synaptic and spiking activity within a local region, even in the absence of visual stimulation. This may help explain the sustained time course of endogenous spatial cuing. If a task-relevant stimulus is anticipated at a particular location, sustained attentional feedback may serve to prioritize processing at that location, whenever the stimulus should appear (Ress, Backus, & Heeger, 2000).

The nature of attentional gain modulation was once subject to considerable debate, but an emerging view is that nonlinear interactions between attentional feedback, stimulus processing, and surround suppression may account for these diverse effects. According to the normalization model of attention, shifts in contrast gain will predominate if the attentional window is large and the stimulus is much smaller (Reynolds & Heeger, 2009). This is attributed to the fact that spatially suppressive surround interactions will tend to saturate the neuron's response to the target stimulus at high contrasts and the large attentional window will contribute to this suppressive effect. However, if the attentional window is small and restricted within the stimulus, then attention is expected to boost responses to the stimulus in a multiplicative manner, by avoiding any increase in the strength of surround suppression. There is some compelling behavioral and fMRI evidence to support the predictions of the normalization model of attention

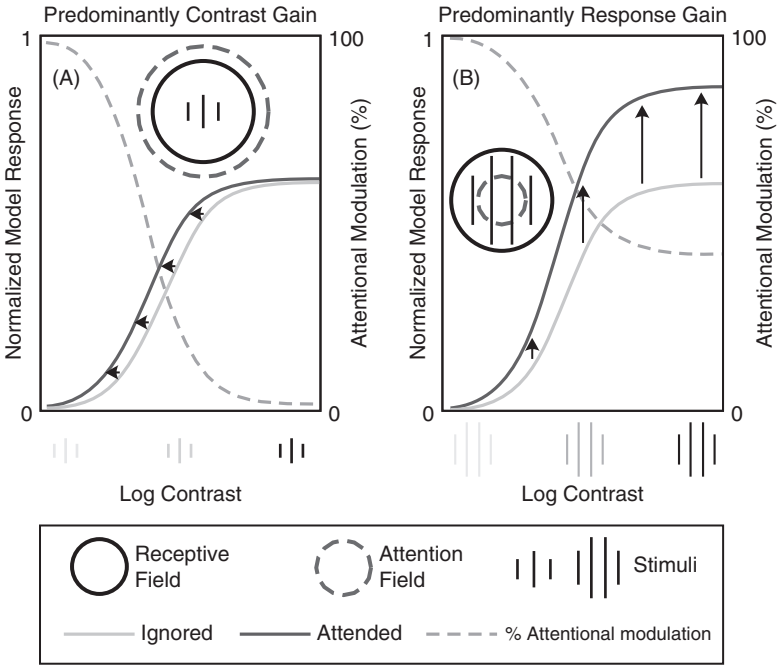

Figure 1.15 Normalization model of attention. Possible effects of contrast gain (A) and response gain (B) due to spatial attention. The normalization model of attention predicts different types of attentional modulation, depending on whether the attentional field is much larger than the stimulus or restricted within the stimulus proper. Effects of attention are plotted in comparison to an unattended condition, with attention directed to the opposite hemifield. Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.
SOURCE: From Reynolds and Heeger (2009, p. 172). Reproduced with permission of Elsevier.

(Herrmann, Heeger, & Carrasco, 2012; Herrmann, Montaser-Kouhsari, Carrasco, & Heeger, 2010), although more research will be needed to fully evaluate this model.

Attending to Multiple Spatial Locations

A spotlight metaphor is commonly used to describe people's ability to voluntarily shift covert attention from one location to another. However, attentional resources can be allocated in more flexible ways. Multi-object tracking studies have shown that people can rely on covert attention to concurrently track about three to four dynamically moving objects (Scholl, 2001), even when such objects are not directly foveated. Such skills are critical when we navigate busy crowds, drive through traffic, or watch a fast-paced basketball game. Parietal visual areas and the frontal eye fields are strongly involved in this dynamic tracking process. Activity in these attentional control centers increases as a function of the number of objects to be tracked, peaking at about three to four items, consistent with behavioral limits of attentional performance (Culham, Cavanagh, & Kanwisher, 2001). Other studies have tested people's ability to monitor rapid serial sequences of letters presented at multiple locations concurrently. These studies find enhanced activity at multiple attended locations in areas V1–V4, indicating that the attentional spotlight can indeed be divided

across multiple regions of space (McMains & Somers, 2004). Taken together, these studies imply that people can concurrently attend to multiple spatial locations and to multiple dynamic objects.

Attending to Visual Features

Although we usually think of attention as having a particular spatial locus, attention can also be flexibly directed toward specific visual features or objects. *Feature-based attention* refers to the ability to attend to a featural property, such as a particular color, orientation, or motion direction (Maunsell & Treue, 2006). Imagine you have forgotten where you parked your car in a crowded lot. If your car happens to be blue, you might find yourself attending to a series of blue cars scattered about the lot until you eventually find your own. Theories of feature-based attention propose that top-down feedback to early visual areas can selectively enhance the representation of a particular feature value, such as "blueness." Moreover, this feature-selective feedback is spatially diffuse, modulating the activity of blue-preferring neurons throughout the visual field, in a manner that can facilitate the visual search process.

Neuronal recordings in motion-selective area MT of alert monkeys provided novel evidence to support a feature-based mechanism of attention (Treue & Martinez-Trujillo, 1999). Monkeys were presented with overlapping sets of upward and downward drifting dots in one part of the visual field, and instructed to monitor for speed changes in either set of dots. In the opposite hemifield, a task-irrelevant motion stimulus was presented, consisting of either upward or downward moving dots. The researchers found that direction-selective responses to these task-irrelevant dots were boosted when they matched the motion direction being attended in the other hemifield. This implies that attending to a particular feature at a specific location led to the enhancement of that feature representation throughout the visual field.

Neuroimaging studies have used multivariate pattern analysis to isolate feature-selective responses in the visual cortex to characterize the effects of selective attention. When observers were cued to attend to one of two overlapping orientations or motion directions, activity patterns were reliably biased in favor of the attended feature (Kamitani & Tong, 2005, 2006). These effects of feature-based attention were pervasive, encompassing the primary visual cortex, extrastriate area V2–V4, as well as area MT+ in the case of attending to motion direction. In studies where task-relevant and task-irrelevant stimulus features were presented in separate parts of the visual field, spatial spreading of feature-based attention has also been observed (Serences & Boynton, 2007).

It is worthwhile to consider whether the allocation of feature-based attention might depend on whether the task requires visual detection or discrimination. When performing a fine-grained discrimination task, such as deciding whether a grating is rotated slightly clockwise or counterclockwise relative to vertical, it would be advantageous to boost the response of neurons that can best distinguish between an oriented stimulus tilted say +2° or –2°. Based on the orientation-tuning bandwidth of cortical neurons and the information they can convey, one would expect that orientation responses should be enhanced at the near flanks of the discrimination boundary, say at +10° and –10°, rather than centered on the discrimination boundary itself. Both psychophysical and human neuroimaging studies provide strong support for this prediction (Scolari & Serences, 2009, 2010), demonstrating that feature-based attention can be allocated in a flexible manner to optimize task

performance. This approach of extracting orientation-selective responses has also been successfully applied in EEG studies by measuring steady-state visually evoked potentials to flickering gratings. These studies reveal a multiplicative gain modulation of attention on the strength of orientation-tuned responses (Garcia, Srinivasan, & Serences, 2013).

Attending to Objects

Many studies find that attention can enhance the processing of particular spatial locations or visual features, but to what extent does attention act upon the representations of visual objects? Whenever we encounter an object in the environment, it has a particular spatial location and consists of a set of visual features, so it can be difficult to tease apart whether attention to that object is guided by its spatial location, its features, or its higher order object properties.

Studies of *object-based attention* have focused on some key predictions. First, if covert attention is directed to one part of an object, it should tend to spread to other parts of that same object. Exogenous cuing studies have found evidence of both spatial and object-based attention. People respond most quickly when an initial cue and subsequent target appear at the same location on a common object. However, they are also somewhat faster to respond if the cue and target appear at different locations on the common object (e.g., two ends of a rectangle) as compared to equidistant locations on different objects, suggesting that attention tends to spread throughout an attended object. fMRI studies provide support for this view, finding enhanced activity in retinotopic visual areas V1–V4 at the site of the initial cue, but also spreading effects of enhancement at distal locations corresponding to the same object (Muller & Kleinschmidt, 2003). Neurophysiological recordings in monkeys have also found evidence of spatial spreading of attention along the length of an object. When these animals perform a mental curve-tracing task, the activity of V1 neurons is enhanced if that neuron's receptive field falls along the line to be covertly traced (Roelfsema, Lamme, & Spekreijse, 1998). Moreover, the latency of this modulation corresponds well with the distance along the curve. These studies demonstrate an interaction between spatial attention and object-based mechanisms, in which spatial attention spreads more readily along a perceptually defined object.

Other studies have investigated people's ability to attend to one of two overlapping objects. When presented with simple objects, such as a tilted line that spatially overlaps a rectangle, participants are faster and more accurate at making judgments about the two visual properties if they pertain to a common object, and slower if the judgment involves both objects (Duncan, 1984). This so-called *two-object cost* was convincingly established in a follow-up study where observers had to track two overlapping gratings that dynamically and independently changed over time, in orientation, color, and spatial frequency along a randomized trajectory through this feature space (Blaser, Pylyshyn, & Holcombe, 2000). These well-tailored stimuli minimized the possibility of relying on spatial attention or attention to a static feature. When tasked with following both dynamic gratings, observers were unable to do so, yet they could effectively attend to one dynamic grating at a time, indicating a powerful object-specific capacity limit.

fMRI studies have capitalized on the category selectivity of high-level visual areas, by presenting stimuli such as overlapping face-house images to investigate object-based attention. When participants were cued to attend to the face (or the house), enhanced activity was observed in the fusiform face area (or the parahippocampal place area),

as predicted (O'Craven, Downing, & Kanwisher, 1999). Interestingly, if the face happened to be moving while the house remained static, activity in area MT+ was also greater when the attended object was moving. These results implied that object-based attention enhances multiple properties of the attended object, even those that are not immediately relevant to the task at hand.

What specific mechanisms allow for the attentional selection of a visual object? fMRI studies have found that attentional feedback signals to early visual areas serve to enhance the representation of the low-level features that comprise the attended object (Cohen & Tong, 2015). As a consequence, it is possible to decode which of two objects is being attended from the activity patterns found in early visual areas. Studies have also reported greater functional connectivity between early visual areas and higher category-selective areas when participants are attending to the object corresponding to that region's preferred category (Al-Aidroos, Said, & Turk-Browne, 2012). These studies suggest that object-based attention involves a strong interplay between higher order visual areas and early visual areas.

Sources of Top-Down Attentional Feedback

We have learned a good deal about how attentional selection is mediated by the top-down modulation of activity in the visual cortex, but where do these top-down attentional signals come from? According to the *premotor theory of attention*, common brain structures are likely involved in controlling overt shifts of the eyes and covert shifts of attention (Awh, Armstrong, & Moore, 2006), in particular the frontal eye fields and the lateral intraparietal area. First studied in nonhuman primates, these frontal-parietal areas are known to have strong reciprocal connections with extrastriate visual areas, the pulvinar, the superior colliculus, and with each other.

Neuroimaging studies have revealed retinotopically organized maps in the frontal eye fields and in multiple intraparietal areas (IPS1 through IPS4). These maps can be revealed by mapping responses throughout the visual field evoked by visual stimulation, planned eye movements, or covert shifts of attention (M. A. Silver & Kastner, 2009). Such findings provide support for the premotor theory that overt and covert shifts of attention involve a common coding scheme. Moreover, damage to the parietal lobe often leads to visuospatial neglect of the contralateral hemifield (Corbetta & Shulman, 2002). In healthy participants, transcranial magnetic stimulation applied to the IPS can cause impaired detection of stimuli presented in the contralateral hemifield, especially when a competing stimulus appears in the ipsilateral visual field (Hilgetag, Theoret, & Pascual-Leone, 2001). These studies provide causal evidence of the role of the parietal lobe in spatial attention.

Microstimulation studies performed in monkeys also demonstrate a causal role for the frontal eye fields (FEFs) in the top-down allocation of spatial attention. In these studies, researchers first applied strong stimulation to an FEF site to determine where the monkey would overtly look. Next, they presented visual stimuli at this spatial location while the animal maintained fixation, applying mild stimulation at levels too weak to evoke an eye movement. Remarkably, the monkey was much better at detecting appearances of a target at that corresponding location whenever weak stimulation was applied (Moore & Fallah, 2001). Simultaneous recordings in area V4 during FEF stimulation further revealed attention-like effects of feedback in area V4, which boosted the neuron's response to stimuli

presented at the presumably attended location (Armstrong & Moore, 2007; Moore & Armstrong, 2003).

Neuroimaging studies have also investigated the brain areas associated with the attentional control of feature-based and object-based attention. Some studies have reported greater activity in medial regions of the parietal lobe time-locked to when participants voluntarily switch their attentional focus from one feature to another, or from one object to another (T. Liu, Slotnick, Serences, & Yantis, 2003; Serences, Schwarzbach, Courtney, Golay, & Yantis, 2004). Studies employing multivariate pattern analysis have also provided evidence of feature-selective representations in intraparietal areas IPS1–IPS4 as well as the frontal eye fields (T. Liu, Hospadaruk, Zhu, & Gardner, 2011). A magnetoencephalography study investigated the relative timing of attentional modulations across the brain by presenting an overlapping face and house that flickered at different rates (Baldauf & Desimone, 2014). Not only was object-specific modulation observed in the fusiform face area and parahippocampal place area; attention also modulated their degree of synchrony with an inferior frontal region (near the FEF). These frontal modulations appeared to lead the ventral temporal modulations by about 20 ms, implying that the prefrontal region was the likely source of the top-down attentional signal. Taken together, these studies suggest that frontoparietal regions associated with the control of spatial attention may also have an important role in the controlling of nonspatial aspects of attention (Ester, Sutterer, Serences, & Awh, 2016).

OBJECT RECOGNITION

An essential function of vision is the ability to categorize and identify objects from a distance: the *what* part of *knowing what is where by looking*. It is hard to imagine what it would be like to see color, lines, and rudimentary shapes, without the ability to recognize the objects around us. However, patients with visual object agnosia demonstrate that such outcomes are possible. Following damage to the ventral temporal cortex, the perception of basic features and shapes usually remains intact; nevertheless, patients with associative agnosia have great difficulty at identifying objects by sight (Farah, 2004; Moscovitch, Winocur, & Behrmann, 1997).

Research suggests that there are different subtypes of visual agnosia, including evidence of a double dissociation between the processing of upright faces and the processing of non-face objects (as well as upside-down faces) (Farah, Wilson, Drain, & Tanaka, 1995; Moscovitch et al., 1997; Rezlescu, Barton, Pitcher, & Duchaine, 2014). *Prosopagnosia*, or severe impairments in face recognition, is strongly associated with damage to the fusiform gyrus (Meadows, 1974), whereas damage to more lateral portions of the inferior temporal cortex usually leads to general impairments in object recognition (i.e., visual object agnosia). The challenges faced by these patients indicate that critical computations for object processing take place at higher levels of the ventral visual pathway.

To identify an object, the visual system must analyze the complex pattern of retinal input and determine the corresponding identity (recall Figure 1.3), thereby allowing access to previously stored information about that type of object. This includes information about the object's visual appearance, such as its shape, color, and texture, as well as its semantic properties and associated verbal label.

The visual analysis required for successful object recognition is a very hard computational problem: The recognition system

must somehow analyze and transform the 2D retinal image into a representation that is both *selective* for that particular object and *invariant* to the image variations that can arise from variations in 3D viewpoint or lighting. This is a difficult problem to solve because most strategies that lead to greater selectivity will lead to *less*, not more, tolerance to variation. Related to this challenge is the *inverse optics problem* (recall Figure 1.5), which requires inferring what would be the most likely 3D object that could have given rise to the observed 2D image. Would a solution to this problem necessarily require solving for the full 3D structure of the observed object, or might object recognition involve matching diagnostic parts of the 2D image to a flexible but image-based representation in memory? As we will see, multiple computational approaches have been proposed for solving this critical problem of object recognition.

Early Models of Object Recognition

A variety of object recognition models have been proposed over the years, often reflecting the Zeitgeist of each period. In the following, we will consider models from the early 1980s to the present, to shed light on how scientific understanding of object recognition has evolved.

In the 1980s, it was generally believed that the visual system analyzed the 2D retinal image by deriving a 3D model of the viewed object. For example, David Marr (1982) proposed that the visible surfaces of an object can be computed from the image to form a viewer-centered 2.5D sketch, based on various cues to the depth dimension including stereopsis, shape-from-shading, shape-from-texture, and so forth (Figure 1.16A). The 2.5D sketch could contain information about the distance of different points along the object and its curvature along the depth dimension, but from a viewer-centered perspective. This, in turn, could be used to determine an object-centered 3D representation of the object's structure.

Consistent with this theory of 3D coding, visual experiments have shown that presentation of an object in one viewpoint can facilitate or prime the recognition of that same object when shown from a different viewpoint. People are also good at matching pictures of unfamiliar objects across changes in viewpoint, especially if the distractor objects have different 3D parts or had a distinct spatial structure (Biederman & Gerhardstein, 1993; Cooper, Biederman, & Hummel, 1992). According to Biederman's *recognition by components theory* (Figure 1.16B), objects are represented by the visual system according to their geometric elements, or *geons*, and the spatial arrangement of those elements, which can lead to a unique *structural description* for many individual objects (Biederman, 1987). For example, a coffee mug and a pail consist of the same geons: a cylindrical geon that has an opening at the top and a curve cylinder that is connected to the base cylinder. Whether the object is a mug or pail, however, depends on whether the curved cylinder connects to the side or the top of the base cylinder.

Although the recognition by components theory provided a simple and coherent account of object recognition, several challenges for this account began to emerge. First, it is nontrivial to determine what geons are contained in an object from a 2D image; this correspondence problem could prove just as difficult as determining the identity of the object. Second, geons might provide a reasonable account of the 3D structure of man-made objects, but it is not clear how a geon-based account would generalize to the recognition of objects in the natural world, such as plants, animals, and people. To what extent do the geons that describe a dog, cat, or horse differ from one another?

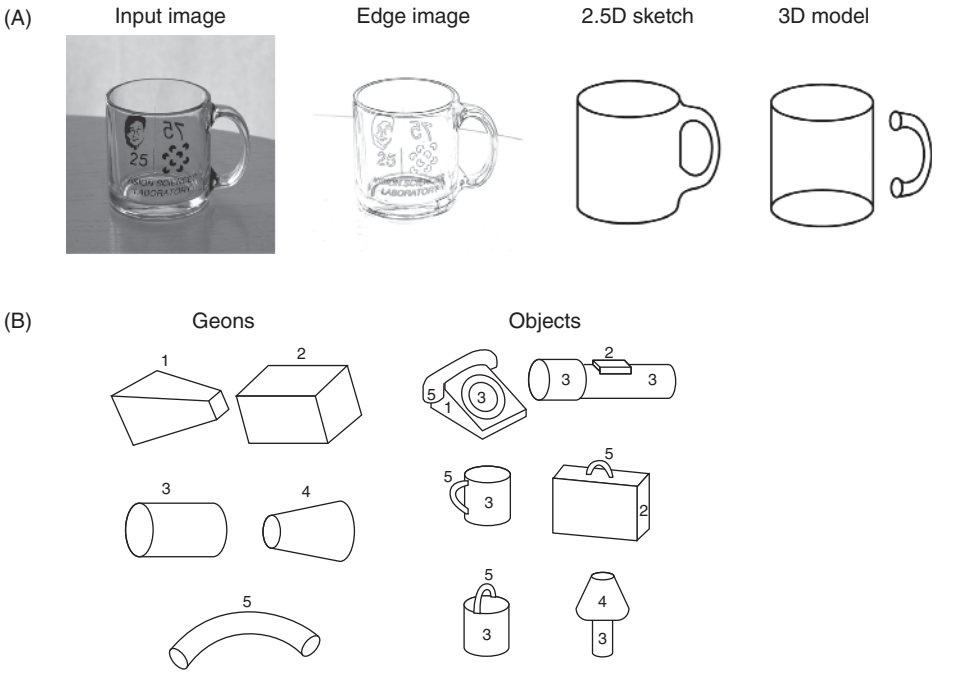

Figure 1.16 Early 3D-based models of visual object recognition. (A) Depiction of Marr's proposed stages of processing, which involved extracting edges in the image, calculating a 2.5D sketch of the object with depth and curvature information estimated from the viewer's perspective, and inferring a fully invariant 3D model. (B) Biederman's recognition-by-components theory proposed that objects are recognized based on the spatial arrangement of their geon-defined parts. Many man-made objects have a unique structural description, according to this view.
SOURCE: (A) Figure created by Frank Tong; used with permission of the author.

In the 1990s, psychophysical studies began to reveal behavioral costs in object recognition performance following changes in 3D viewpoint. For simple geometric shapes such as geons, these costs were modest (Tarr, Williams, Hayward, & Gauthier, 1998), but for structurally similar or confusable 3D stimuli, such as faces or contorted wire-clips, the costs of viewpoint change were far more severe (Bülthoff, Edelman, & Tarr, 1995; Hill, Schyns, & Akamatsu, 1997). Concurrently, recordings from the inferotemporal cortex of the monkey revealed that most neurons respond to a preferred object over a limited range of views, implying view-specific tuning for objects (Logothetis, Pauls, Bulthoff, & Poggio, 1994). These findings led to the proposal that the visual system stores a series of discrete 2D views. fMRI studies of adaptation to visual objects likewise found that lateral occipital object areas primarily show view-specific adaptation, with little evidence of view invariance (Grill-Spector et al., 1999). While most face-selective neurons in the monkey appear to be tuned in a view-specific manner, a subset have been found to be view-invariant, especially in more anterior regions of the temporal lobe (Freiwald & Tsao, 2010; Perrett et al., 1991).

One theoretical argument against view-specific representations goes as follows: It would be too costly for the visual system to encode a near-infinite number of views of every possible object. However, a viable alternative would be to encode a handful

of distinct views and to rely on an interpolation process for intermediate views. The correlational similarity between object images is usually quite high following modest depth rotations, and storing a small number of discrete views would be enough to support near-invariant performance.

Although there was growing evidence that the ventral visual system relies on view-specific object representations, it took a while for researchers to develop plausible neural models for image-based recognition. An influential model emerged in the late 1990s, inspired by the hierarchical organization of the visual cortex (Riesenhuber & Poggio, 1999). This hierarchical model, referred to as *HMAX*, relies on a multilayered architecture (cf. Fukushima & Miyake, 1982) that capitalizes on the functional architecture of V1 simple cells and complex cells (Figure 1.17).

The HMAX model expands on ideas originally proposed by Hubel and Wiesel, noting that simple cells achieve greater visual selectivity by performing an AND-like computation, whereas complex cells achieve greater invariance by performing an OR-like computation. Mathematically speaking, simple cells compute a weighted sum of inputs from the preceding layer, according to a preferred template or filter (e.g., an orientation-tuned Gabor function). This is followed by half-wave rectification so that negative responses are set to zero. In contrast, the OR-like function involves performing a maximum-pooling operation (*MAX*), so that the complex cell's response is determined by the response of the most active simple cell from which it receives input. Strong activation by any one of those units will suffice to activate the complex cell, thereby achieving an invariant preference for orientation across local changes in spatial phase.

In the HMAX model, layers 1 and 2 of this network consist of simple- and complex-cell units, respectively. These AND and MAX

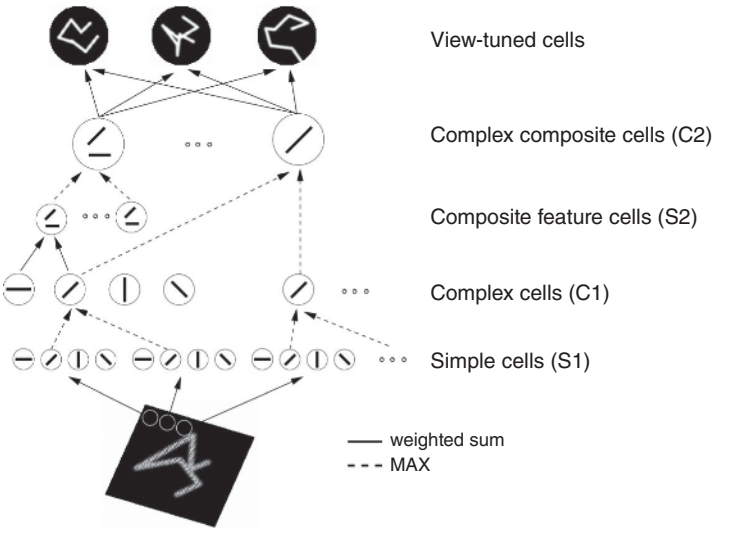

Figure 1.17 HMAX model of object recognition. Depiction of the HMAX model consisting of repeating layers of simple units and complex units, prior to the final output layer. The model was trained on different paperclip objects at one specific view, and then tested across a range of depth rotations. Despite the specificity of training, the model shows reasonably good tolerance to viewpoint change following training.
SOURCE: From Riesenhuber and Poggio (1999). Used with permission of Nature Publishing Group.

operations are then repeated in subsequent layers, such that tuning for combinations of orientations arise, as well as tuning for more complex properties related to 2D shape. For example, a curve line can be described as a combination of two orientations that meet at a junction. The final layer is then trained to learn specific stimuli. Notably, after the network is trained to discriminate 3D-rendered computer stimuli across a limited range of viewpoints, the model showed a reasonable degree of invariance to changes in 3D viewpoint. The modeling results suggested that a 2D image–based approach might prove effective for recognizing objects across changes in viewpoint. Elaboration of this work has shown that it is possible to attain greater selectivity and invariance by creating deeper networks with more layers and training on a greater number of images (Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007).

Deep Learning Models of Object Recognition

However, it was not until 2012 that a major breakthrough occurred in the computational modeling of object recognition, with the advent of *convolutional neural networks*, or CNNs (Krizhevsky, Sutskever, & Hinton, 2012). CNNs are deep neural networks that consist of much of the same architecture as the HMAX model, with repeating layers of rectified linear units followed by maximum-pooling units. The critical advance was the application of deep learning methods to train these multilayer networks on massive image datasets (LeCun, Bengio, & Hinton, 2015). Deep learning has led to major advances in multiple domains of artificial intelligence, ranging from object recognition to self-driving cars to grandmaster level performance at the exceedingly complex game of Go (D. Silver et al., 2016). Supervised deep learning relies on backpropagation to modify the weights of the network from the top layer downward, with the goal of minimizing error in classification performance. Another simplifying assumption used by CNNs is that the stacks of units in the early layers should share a common set of weights, such that they provide a common set of filters or basis functions for encoding the information in their receptive field.

In the 2012 ImageNet competition, Alex Krizhevsky and his colleagues demonstrated the power of CNNs, training a network on 1.2 million images to classify real-world images according to 1,000 different object categories. This network, now called AlexNet (Figure 1.18A), outstripped the competition, selecting the correct object category as one of its top five choices about 84% of the time on a large test dataset. Since then, multiple research groups have pursued the goal of attaining more accurate performance with CNNs (He, Zhang, Ren, & Sun, 2016; Szegedy et al., 2015), and some suggest that machine performance is approaching the accuracy of human performance (He et al., 2016; Yamins et al., 2014).

Because CNNs are exceedingly complex—AlexNet has 6 million parameters—some have argued that the computations performed by CNNs are akin to a black box. However, researchers have devised various methods to visualize the tuning preferences of individual units of the CNN (Bach et al., 2015; Zeiler & Fergus, 2014). Since the higher units have highly nonlinear receptive fields, one can only visualize the particular features of a given image that lead to the strong excitation of particular unit. Nevertheless, these studies suggest that CNNs capture some of the tuning properties of biological visual systems. Lower-level units are predominantly tuned to color or orientation, similar to neurons in V1, whereas units in the intermediate layers exhibit tuning for textured patterns or combinations of features (Figure 1.18B).

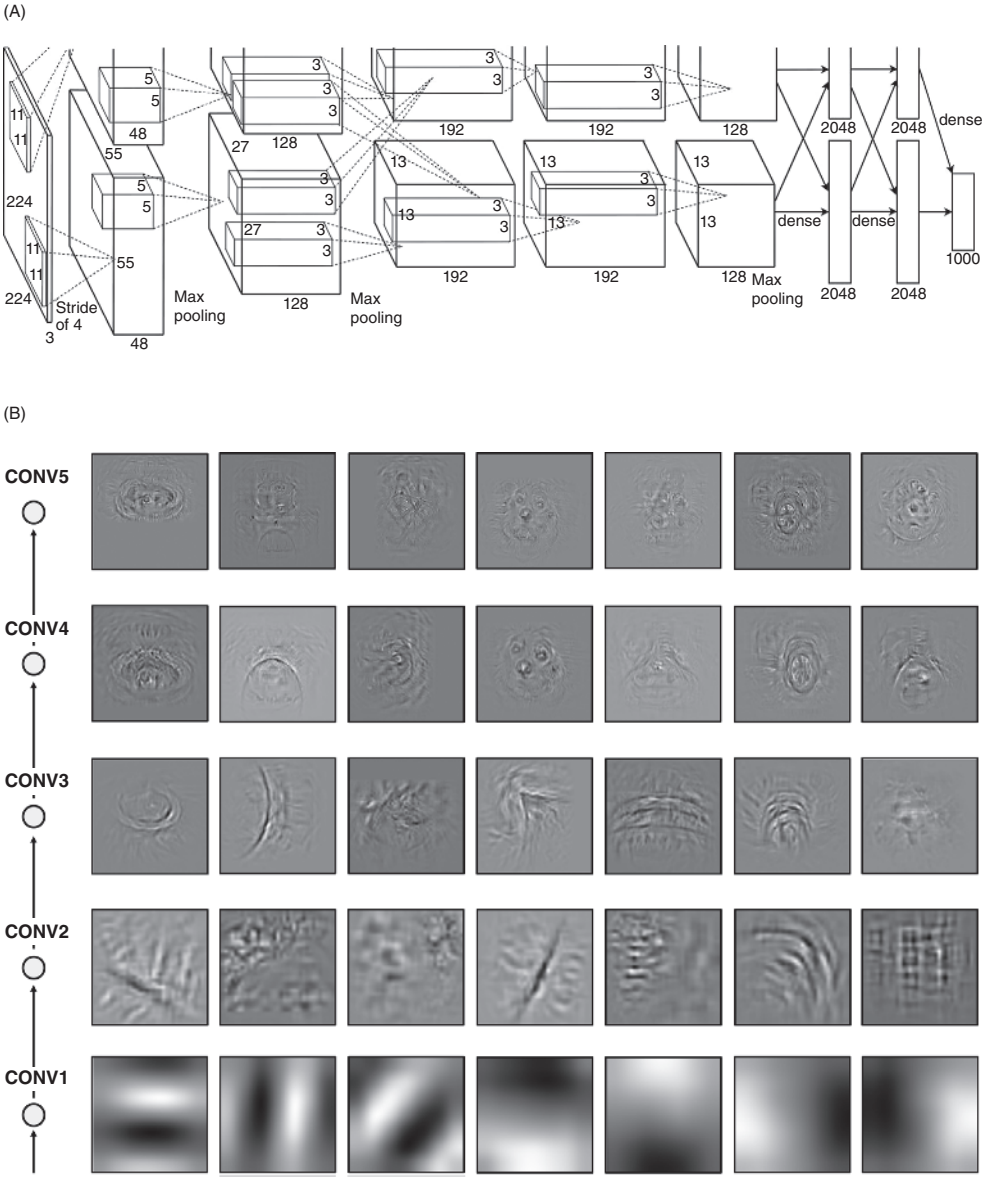

Figure 1.18 Architecture and tuning properties of a convolutional neural network. (A) Architecture of AlexNet, a convolutional neural network that outperformed all other algorithms in the 2012 ImageNet competition. The input layer is on the left, and neurons in each successive layer sample from just a local region of the preceding layer. Successive stages of filtering, nonlinear rectification, and max pooling are performed, until at the last few stages are fully convolutional. (B) Visualization of tuning preferences of individual units of a convolutional neural network, based on a deconvolution approach to depict image components that strong responses for units in convolutional layers 1 through 5. Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.

SOURCE: (A) From Krizhevsky, Sutskever, and Hinton (2012). (B) Images provided courtesy of Hojin Jang and Frank Tong (copyright).

By contrast, units at high levels of the CNN, which have large receptive fields that span the full array of units in the preceding layer, respond best to complex patterns or even to specific objects such as faces. Researchers have compared the object tuning preferences of inferotemporal neurons in the monkey with units in these CNNs, and find evidence of a strong correspondence between brain and machine (Yamins et al., 2014). In fact, the tuning preferences of individual IT neurons can be well predicted by a weighted combination of the responses of units at the highest levels of a trained CNN, whereas lower-level units provide a poorer account of IT response preferences. Neuroimaging studies have also found that individual voxel responses and cortical activity patterns in higher object-sensitive areas correspond well with the response preferences of high-level units in CNNs, whereas fMRI responses in early visual areas are better described by low-level units (Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014).

Although this research is at an early stage, CNNs provide the most powerful and plausible model of object recognition in humans to date. The ability to relate the response properties of CNNs to single-unit activity and fMRI activity suggests that this approach can lead to greater insight into the neural bases of object processing. For many, it may be surprising that training on a large set of 2D images of objects, with no explicit representation of 3D structure, can allow for the accurate recognition of objects across variations in viewpoint as well as generalization to novel exemplars. However, neuroscientists have argued that the function of inferotemporal cortex is to learn the appropriate mappings that serve to untangle the representations of different objects through a series of nonlinear transformations (DiCarlo, Zoccolan, & Rust, 2012). Perhaps the remarkable accuracy and flexibility of human object recognition is simply a product of a lifetime of visual experiences and learning opportunities.

That said, a major limitation of current CNNs is their reliance on supervised approaches for deep learning. Infants and children do not receive such frequent or explicit feedback when they encounter new objects in the world, nor do they appear to require nearly as many training examples. Unsupervised networks can extract object structure from training examples (Le et al., 2012), but have yet to achieve the performance levels comparable to supervised networks. The ability to shift from supervised to unsupervised approaches to train these networks would constitute a major advance in deep learning and may also clarify the biological bases of visual learning. Another limitation of current CNNs such as AlexNet and GoogLeNet is their reliance on strictly feedforward processing, as it is known that top-down attentional feedback can improve perception and object recognition performance. It will be interesting to see if future CNN models that incorporate higher level neural processes such as dynamic feedback might lead to even better performance while shedding light on the neural computations underlying human vision.

Face Recognition and Subordinate-Level Discrimination

When performing a recognition task, an object can be identified or labeled with varying degrees of specificity. For example, we may want to distinguish between two distinct classes of objects, such as dogs and cats, whereas in other cases, we may want to make a more fine-grained distinction, such as differentiating a pug from a bulldog. The verbal labels that people use most often to identify objects may provide clues as to how they prefer to distinguish among visual stimuli.

According to theories of visual categorization, people should be faster and more accurate at naming objects according to their *basic-level category* (Mervis & Rosch, 1981; Palmeri & Gauthier, 2004). The basic level is believed to maximize the within-class similarity among exemplars within the category (e.g., different breeds of dogs) while maximizing the separation between that category and other basic-level categories (e.g., dog vs. cat). In comparison, telling apart *exemplars* from a common basic-level category requires more fine-grained discrimination and usually requires more processing time to determine the *subordinate-level category* of an object. Thus, when shown a picture of a dachshund, the first thought to come to mind might be "dog," then perhaps "short legs," before it is followed by "oh, it's a dachshund." Subordinate-level categorization occurs whenever we identify a dog by its breed, a car by its model, or a bird by its species.

While we may be predisposed to identify common objects at the basic level, human faces seem to constitute a special class of stimuli that people process at the subordinate level, with greater focus on the uniquely distinguishing properties of each individual face. The task of *face recognition* requires particularly fine-grained discrimination, as all faces share the same basic parts and a common configuration. It is the subtle variations in the local features and their relative arrangement that distinguish one face from another face, which the visual system somehow learns to tell apart. Our ability to recognize upright faces gradually improves with experience throughout childhood and early adulthood, up to at least one's mid-30s (Germine, Duchaine, & Nakayama, 2011). One consequence of this extensive training is that we are far better at perceiving, recognizing, and remembering faces when presented in a familiar upright orientation than when upside-down (Lorenc, Pratte, Angeloni, & Tong, 2014; McKone & Yovel, 2009; Valentine, 1988). This *face inversion effect* can be observed even in more basic tasks that require detecting the presence of a face in an ambiguous image or perceiving an emotional expression (Figure 1.19). In many ways, people appear to be experts at processing upright faces, and when given the opportunity to train at distinguishing exemplars from another stimulus class, such as dogs, cars, or artificially rendered objects, they tend to show a greater cost of stimulus inversion following training (Diamond & Carey, 1986; Gauthier & Tarr, 1997).

The shared similarity of faces would present a major challenge to any recognition system. The study of face processing has helped reveal how the visual system represents and distinguishes the variations that occur among exemplars from this natural stimulus class. Vision scientists have measured the 3D structure of faces, using laser range-finding methods, and applied analytic methods to reveal how faces naturally vary across individuals. For example, two of the principal components along which faces vary in 3D shape can be roughly described in terms of gender and adiposity (i.e., how wide or thin a face appears) (Leopold, O'Toole, Vetter, & Blanz, 2001). Studies of visual aftereffects suggest that the visual system encodes faces according to deviations from a prototype (or the central tendency of exemplars), such that prolonged viewing of a masculine face will cause a gender-neutral face to appear feminine, and vice versa (Webster, Kaping, Mizokami, & Duhamel, 2004). Similarly, adaptation to a thin face will cause an average face to appear much wider. Both human neuroimaging studies and neuronal recordings in monkeys provide support for the notion that faces are encoded according to how they deviate from an average face, as larger deviations or caricatured faces tend to evoke stronger responses at face-selective sites

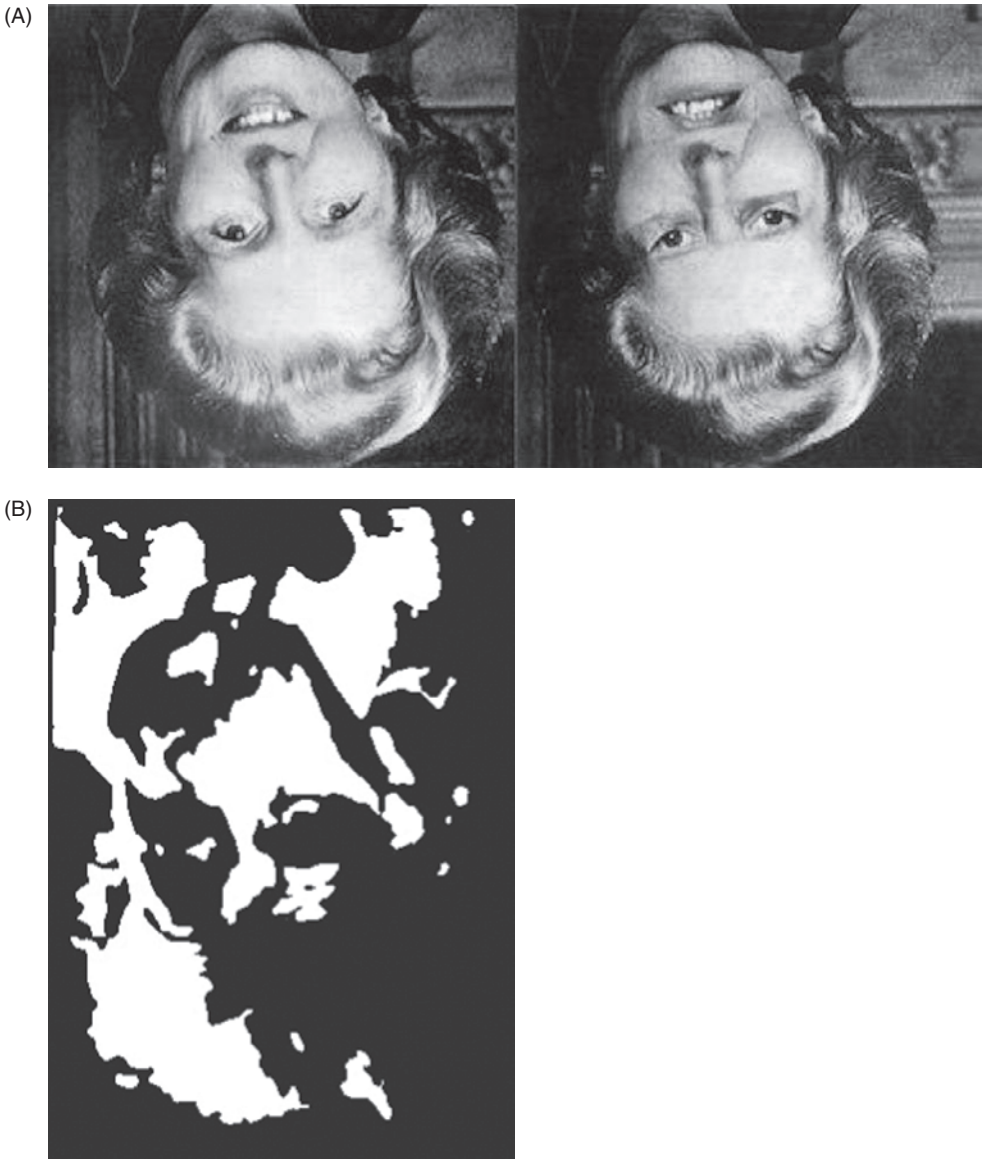

Figure 1.19 Examples of the effect of face inversion. (A) Thatcher illusion by Pete Thompson. Facial features and emotions are difficult to perceive upside down. (B) Sparse images of faces, such as two-tone Mooney images, are difficult to perceive as faces when show upside-down. Color version of this figure is available at http://onlinelibrary.wiley.com/book/10.1002/9781119170174.
SOURCE: (A) Adapted from Thompson (1980). Figure created by Frank Tong; used with permission of the author.

(Leopold, Bondar, & Giese, 2006; Loffler, Yourganov, Wilkinson, & Wilson, 2005).

The constrained nature of face stimuli has also allowed vision scientists to apply psychophysical procedures to determine what features of a face are most informative for particular tasks. One such method, called Bubbles, involves presenting randomly

selected components of a face, across multiple spatial scales, to determine what local features of a face are most needed to perform a discrimination task (Figure 1.20). This approach can be used to map what features are most informative for determining the gender or emotional expression of a face (Adolphs et al., 2005; Gosselin & Schyns, 2001).

Neural Mechanisms of Face Processing

What are the neural mechanisms that underlie our remarkable abilities at discriminating and recognizing faces? One of the first category-selective visual areas identified in humans was the fusiform face area, or FFA (Kanwisher et al., 1997). This cortical region, which lies anterior to the extrastriate visual cortex in the fusiform gyrus, responds preferentially to human faces, animal faces, and schematic cartoon faces, as compared to a variety of non-face stimuli, including hands, body parts, flowers, and a variety of inanimate objects (Kanwisher et al., 1997; McCarthy, Puce, Gore, & Allison, 1997; Tong, Nakayama, Moscovitch, Weinrib, & Kanwisher, 2000). Activity in the FFA is strongly associated with the conscious perception of faces (McKeeff & Tong, 2007; Tong et al., 1998) and more strongly engaged by holistic processing of upright faces (Kanwisher, Tong & Nakayama, 1998; Yovel & Kanwisher, 2005). Moreover, this region can be modulated by visual adaptation to individual faces, suggesting that it is sensitive to face identity (Loffler et al., 2005; Rotshtein, Henson, Treves, Driver, & Dolan, 2005). The causal role of the FFA in face perception has also been shown in electrical stimulation studies of preoperative epilepsy patients and can impair face recognition (Allison et al., 1994) and even induce perceptual distortions of viewed faces (Rangarajan et al., 2014).

A more posterior face-selective region, known as the occipital face area (OFA), responds at an earlier latency than the FFA and is associated with early face detection processes (J. Liu, Harris, & Kanwisher, 2002; Pitcher, Walsh, & Duchaine, 2011). The OFA, which lies near the surface of the skull, can be targeted by noninvasive transcranial magnetic stimulation (TMS), and TMS applied to the OFA disrupts performance on face perception tasks (Kietzmann et al., 2015; Pitcher, Walsh, Yovel, & Duchaine, 2007). Human neuroimaging studies commonly find another face-selective region in the superior temporal sulcus (STS) that responds more strongly to stimuli associated with dynamic facial motion, including both static and dynamic images of facial expressions, movements of the eyes, and movies of mouth movements during speech (Hoffman & Haxby, 2000; Puce, Allison, Bentin, Gore, & McCarthy, 1998). Such findings have led to the proposal that face processing relies on a distributed set of brain areas that include both a ventral component and a dorsal component (Haxby, Hoffman, & Gobbini, 2000). The FFA, which lies more ventrally, is presumably dedicated to processing the invariant aspects of faces needed for identification, whereas the more dorsal STS region serves to process the dynamic and variable aspects of faces, such as those that occur during facial expressions, shifts of overt attention, and speech.

About a decade after the discovery of these face-selective visual areas in humans, neuroscientists devised paradigms to perform parallel fMRI studies in alert monkeys (Tsao, Freiwald, Tootell, & Livingstone, 2006). This work has revealed a set of six face-selective patches in the macaque temporal cortex that seem to share strong homologies with the human face-processing network (Tsao, Moeller, & Freiwald, 2008). In the monkey, all six patches respond more strongly to faces than to a variety of non-face stimuli (e.g., bodies, fruits, man-made objects). Moreover, electrical stimulation applied to any one of these sites leads to activation at the

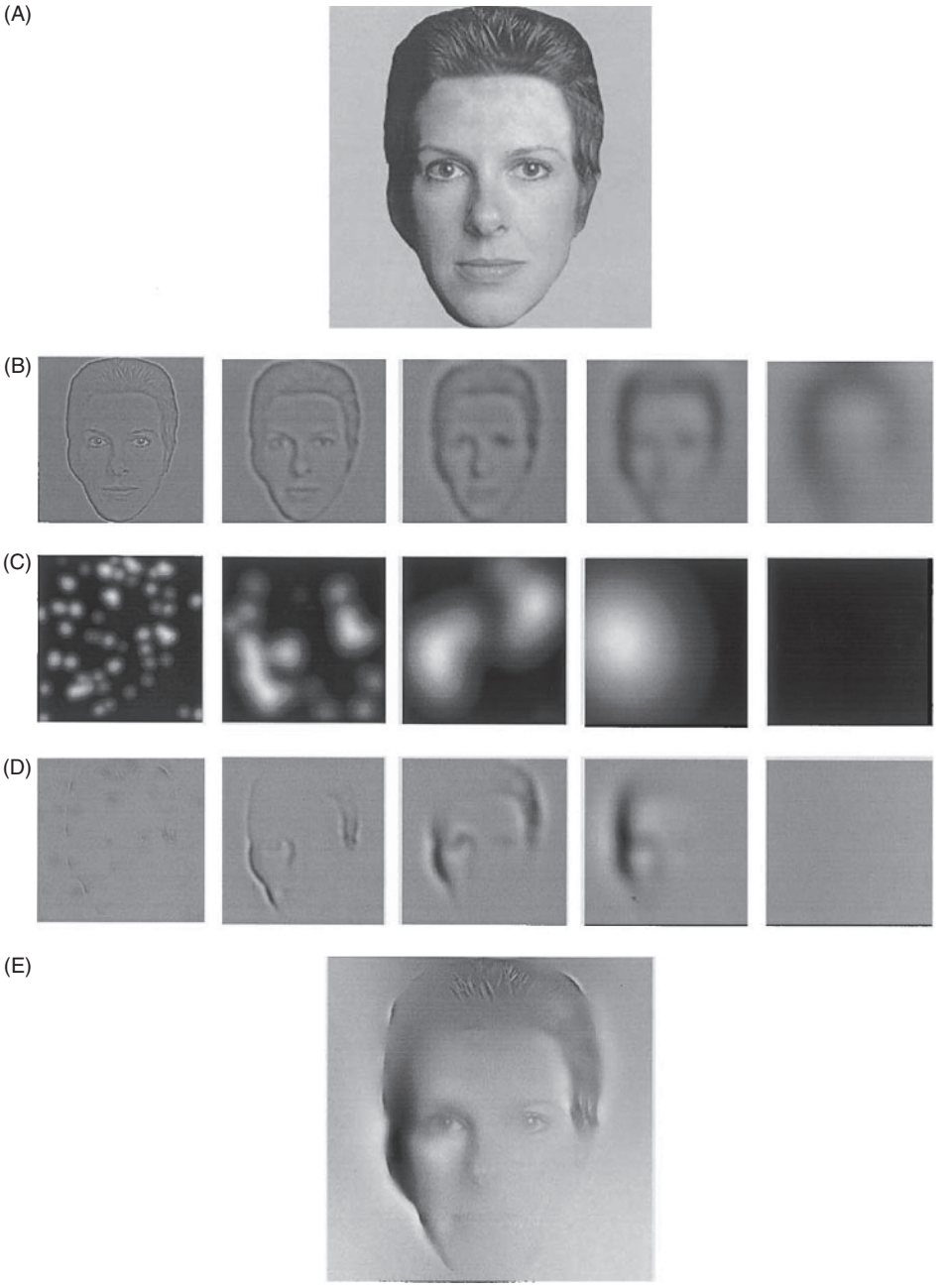

Figure 1.20 Example of the Bubbles technique. This psychophysical technique presents information about a stimulus (A) at different frequencies (B) and locations (C) to determine what local features and spatial scales are the most informative for performing a discrimination task. A randomly generated composite image is presented on every trial (D). Here, Bubbles is being used for a gender judgment task, and the reconstructed image, shown in (E), shows the most informative local features of the face in this task.
SOURCE: From Gosselin and Schyns (2001). Reproduced with permission of Elsevier.

other sites, indicating that these functionally defined face patches comprise a highly interconnected network (Moeller, Freiwald, & Tsao, 2008).

Single-unit recordings from the ventral face patches have further revealed a hierarchically organized shift from view-specific to view-invariant coding (Freiwald & Tsao, 2010). In the middle face patch, which appears to be homologous to the human FFA, individually recorded neurons show peak tuning preference for just a single viewpoint, and most respond best to front views of faces. Most of these cells are tuned to one or only a few facial features, such as the face's width or aspect ratio, the distance between the eyes, iris size, and so forth (Freiwald, Tsao, & Livingstone, 2009). The pattern of responses from many these neurons can thereby provide a code for distinguishing between individual faces. Indeed, facial identity can be reliably decoded from their patterns of activity, whereas information about exemplars from other object categories has proven unreliable (Tsao et al., 2006). Many of these neurons also show evidence of holistic processing, responding more strongly to their preferred feature when it is presented in the context of a facial outline. At the next stage of processing, in a more anterior patch called AL, many neurons exhibit viewpoint symmetric tuning (i.e., faces rotated to a similar degree to the left or right of a front-on view), suggesting a partial degree of viewpoint invariance. Finally, in the most anterior face patch called AM, many neurons respond well to the full range of possible face views. Some of these neurons even show stable preference for a specific face identity across large changes in viewpoint. Taken together, these findings suggest that viewpoint invariance is achieved by first combining view-specific inputs to achieve viewpoint symmetric tuning, followed by the integration of these signals to achieve view-invariant selectivity at the highest levels of the inferotemporal cortex.

It should be emphasized that the development of these face-selective networks depends on both nature and nurture. Cross-sectional testing of thousands of online participants has revealed that people steadily improve in their face recognition abilities over the first 30+ years of life (Germine et al., 2011). Also, people are expert at distinguishing faces from their own cultural group, but quite poor at recognizing faces from unfamiliar cultures. This cross-race recognition deficit seems largely attributable to a lack of visual training. Cross-sectional studies also suggest that after people move to a new country, they gradually improve in their ability to recognize faces of the initially unfamiliar cultural group over a prolonged period extending up to two decades (Rhodes et al., 2009). Taken together, these results suggest that face recognition gradually improves with each new face that is learned, until eventually, after decades of exposure, performance begins to asymptote. Neuroimaging studies have found potential correlates of these behavioral improvements. An fMRI study comparing children (ages 7–11) and adults found that the right FFA increases threefold in size by adulthood, whereas the left FFA is only modestly larger (Golarai et al., 2007).

Biological and genetic factors also have a strong influence on face processing. Twin studies suggest that there is a prominent heritable component to face recognition ability (Wilmer et al., 2010), whereas developmental prosopagnosia has a tendency to run in families (Duchaine, Germine, & Nakayama, 2007). Researchers are beginning to uncover differences in cortical organization that may account for individual differences in face recognition ability, including changes in the size of the FFA, differences in white matter tracts, and microstructure differences (Gomez, 2015, 2017; Pinel, 2015; Saygin, 2011).

Studies of perceptual training with non-face objects suggest that much of the inferior temporal cortex is highly plastic and capable of learning new visual forms and new visual associations. Visual expertise with a stimulus class, such as birds, cars, or radiological images, tends to lead to greater activity in the FFA as well as other regions in the ventral temporal cortex (Gauthier, Skudlarski, Gore, & Anderson, 2000; Harley et al., 2009; McGugin, Gatenby, Gore, & Gauthier, 2012). This suggests that processing in the FFA might not be exclusively dedicated to faces. Intriguingly, researchers have investigated the effects of prolonged training in monkeys with various stimuli at different ages of onset (Srihasam, Mandeville, Morocz, Sullivan, & Livingstone, 2012). Monkeys were assigned to discriminate letters, Tetris-like block patterns, or schematic cartoon faces. Early training with a particular stimulus type led to functionally distinct effects, with a spatially distinct region of selectivity emerging in the inferotemporal cortex, whereas later training did not. These results suggest that early visual experience can strongly modify the functional organization of the inferotemporal cortex, whereas training at an older age leads to more constrained effects, presumably because of the functional topography that is already in place.

CONCLUDING REMARKS AND FUTURE DIRECTIONS

This review described how researchers have capitalized on sophisticated behavioral, neural, and computational methods to advance understanding of the neural mechanisms of visual feature perception, figure-ground perception, and the processing of visual context. Vision research has also provided critical techniques and powerful computational approaches for characterizing higher cognitive functions of top-down attention and object recognition. The progress made since the new millennium has been truly remarkable.

With this growing knowledge base, new questions have emerged on the horizon. The perception of basic features and features in global contexts is strongly linked to the information processing that takes place in early visual areas. However, it remains puzzling as to how this detailed information is subsequently read out by higher visual areas for perceptual report. As visually precise information is passed from lower to higher areas, what information is maintained and what information is lost or distorted? Studies of perceptual decision making that rely on simple binary decisions have yet to address this thorny issue. Along related lines, what is the role of attentional feedback in this read-out process, and might attention have a critical role in allowing for the flexible transmission of high-fidelity information between early visual areas and higher order areas?

Another question concerns the top-down mechanisms of perceptual inference and how they resemble or differ from the voluntary effects of top-down attention. Powerful automatic effects of feedback have been documented during perceptual filling-in, figure-ground segmentation, and perceptual grouping, indicating that higher areas send feedback signals to early visual areas to signify inferences made based on the broader visual context. Such feedback effects are in accordance with a general predictive coding framework (Friston, 2005; Rao & Ballard, 1999), whereas other models have proposed more specific accounts of contextual processing (Brosch, Neumann, & Roelfsema, 2015; Craft, Schutze, Niebur, & von der Heydt, 2007). Many well-known illusions, such as the tilt-surround illusion, remain to be understood at a neural level (Schwartz, Sejnowski, & Dayan, 2009). At the same

time, the discovery of new and surprising illusions, such as #TheDress, point to the fact that we have only a rudimentary idea of how the visual system makes inferences. It will be of considerable interest to see whether Bayesian accounts of visual perception and population coding, which currently focus on more specific or abstracted problems, can help motivate the development of neural models that can make perceptual inferences in generalized contexts.

Finally, we witnessed how deep convolutional networks, designed with an architecture based on the visual system, have outperformed all prior models of object recognition. Investigations of these networks have revealed that individual units develop response preferences that resemble the visual system. However, current models can also be biased to make gross errors that no human ever would, through simple image modifications such as the addition of adversarial noise. Thus, current models share some but far from all of functional properties of the human visual system. As computer scientists seek to achieve more accurate performance, by training deeper networks with ever-larger data sets, it is not clear that a more accurate characterization of the visual system will emerge. Instead, it will be important to consider design aspects of the network's architecture, the learning algorithm and its implementation, and approach taken to training the network to understand what attributes of deep networks may provide a better characterization of our own visual system (Yamins & DiCarlo, 2016).

Although convolutional neural networks are exceedingly complex and highly nonlinear, strictly speaking they are not black boxes, as their tuning properties can be interrogated. This deep learning approach to visual processing will have an important complementary role in the quest to understand the neural computations of our own visual system, as they provide the best current account of how a network of simple units, with appropriately learned weights, can extract structure and meaningful information from complex natural images.

REFERENCES

Adams, D. L., & Horton, J. C. (2003). Capricious expression of cortical columns in the primate brain. *Nature Neuroscience*, *6*(2), 113–114. doi:10.1038/nn1004

Addams, R. (1834). An account of a peculiar optical phenomenon seen after having looked at a moving body. *London and Edinburgh Philosophical Magazine and Journal of Science*, *5*, 373–374.

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America*, *2*(2), 284–299.

Adelson, E. H., & Movshon, J. A. (1982). Phenomenal coherence of moving visual patterns. *Nature*, *300*(5892), 523–525.

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., & Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature*, *433*(7021), 68–72. doi:10.1038/nature03086

Al-Aidroos, N., Said, C. P., & Turk-Browne, N. B. (2012). Top-down attention switches coupling between low-level and high-level areas of human visual cortex. *Proceedings of the National Academy of Sciences, USA*, *109*(36), 14675–14680. doi:10.1073/pnas.1202095109

Alais, D., & Blake, R. (1999). Neural strength of visual attention gauged by motion adaptation. *Nature Neuroscience*, *2*(11), 1015–1018. doi:10.1038/14814

Alitto, H. J., & Usrey, W. M. (2008). Origin and dynamics of extraclassical suppression in the lateral geniculate nucleus of the macaque monkey. *Neuron*, *57*(1), 135–146.

Allison, T., Ginter, H., McCarthy, G., Nobre, A. C., Puce, A., Luby, M., & Spencer, D. D. (1994). Face recognition in human extrastriate cortex. *Journal of Neurophysiology*, *71*(2), 821–825.

Andolina, I. M., Jones, H. E., Wang, W., & Sillito, A. M. (2007). Corticothalamic feedback enhances stimulus response precision in the visual system. *Proceedings of the National Academy of Sciences, USA*, *104*(5), 1685–1690. doi:10.1073/pnas.0609318104

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, *78*(4), 266–278.

Armstrong, K. M., & Moore, T. (2007). Rapid enhancement of visual cortical response discriminability by microstimulation of the frontal eye field. *Proceedings of the National Academy of Sciences, USA*, *104*(22), 9499–9504. doi: 10.1073/pnas.0701104104

Awh, E., Armstrong, K. M., & Moore, T. (2006). Visual and oculomotor selection: Links, causes and implications for spatial attention. *Trends in Cognitive Sciences*, *10*(3), 124–130. doi:10.1016/j.tics.2006.01.001

Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, *10*(7), e0130140. doi:10.1371/journal.pone.0130140

Baldauf, D., & Desimone, R. (2014). Neural mechanisms of object-based attention. *Science*, *344*(6182), 424–427. doi:10.1126/science.1247003

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(6), 1162–1182.

Blake, R., Overton, R., & Lema-Stern, S. (1981). Interocular transfer of visual aftereffects. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(2), 367–381.

Blaser, E., Pylyshyn, Z. W., & Holcombe, A. O. (2000). Tracking an object through feature space. *Nature*, *408*(6809), 196–199.

Bouvier, S. E., & Engel, S. A. (2006). Behavioral deficits and cortical damage loci in cerebral achromatopsia. *Cerebral Cortex*, *16*(2), 183–191. doi:10.1093/cercor/bhi096

Boynton, G. M., Demb, J. B., Glover, G. H., & Heeger, D. J. (1999). Neuronal basis of contrast discrimination. *Vision Research*, *39*(2), 257–269.

Briggs, F., Kiley, C. W., Callaway, E. M., & Usrey, W. M. (2016). Morphological substrates for parallel streams of corticogeniculate feedback originating in both V1 and V2 of the macaque monkey. *Neuron*, *90*(2), 388–399. doi:10.1016/j.neuron.2016.02.038

Briggs, F., Mangun, G. R., & Usrey, W. M. (2013). Attention enhances synaptic efficacy and the signal-to-noise ratio in neural circuits. *Nature*, *499*(7459), 476–480. doi:10.1038/nature12276

Brosch, T., Neumann, H., & Roelfsema, P. R. (2015). Reinforcement learning of linking and tracing contours in recurrent neural networks. *PLoS Computational Biology*, *11*(10), e1004489. doi:10.1371/journal.pcbi.1004489

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience*, *29*(44), 13992–14003. doi:29/44/13992 [pii] 10.1523/JNEUROSCI.3577-09.2009

Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, *5*, 247–260.

Campbell, F. W., & Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *Journal of Physiology*, *197*, 551–566.

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

Carrasco, M. (2011). Visual attention: the past 25 years. *Vision Research*, *51*(13), 1484–1525. doi:10.1016/j.visres.2011.04.012

Carrasco, M., Ling, S., & Read, S. (2004). Attention alters appearance. *Nature Neuroscience*, *7*(3), 308–313. doi:10.1038/nn1194

Casagrande, V. A., & Xu, X. (2004). Parallel visual pathways: A comparative perspective. In L. C. a. J. S. Werner (Ed.), *The visual neurosciences* (pp. 494–506). Cambridge, MA: MIT Press.

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002a). Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *Journal of Neurophysiology*, *88*(5), 2530–2546. doi:10.1152/jn.00692.2001

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002b). Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *Journal of Neurophysiology*, *88*(5), 2547–2556. doi:10.1152/jn.00693.2001

Chaudhuri, A. (1990). Modulation of the motion aftereffect by selective attention. *Nature*, *344*(6261), 60–62. doi:10.1038/344060a0

Cohen, E. H., & Tong, F. (2015). Neural mechanisms of object-based attention. *Cerebral Cortex*, *25*(4), 1080–1092. doi:10.1093/cercor/bht303

Cooper, E. E., Biederman, I., & Hummel, J. E. (1992). Metric invariance in object recognition: A review and further evidence. *Canadian Journal of Psychology*, *46*(2), 191–214.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. doi:10.1038/nrn755

Craft, E., Schutze, H., Niebur, E., & von der Heydt, R. (2007). A neural model of figure-ground organization. *Journal of Neurophysiology*, *97*(6), 4310–4326. doi:10.1152/jn.00203.2007

Culham, J. C., Cavanagh, P., & Kanwisher, N. G. (2001). Attention response functions: Characterizing brain areas using fMRI activation during parametric variations of attentional load. *Neuron*, *32*(4), 737–745.

Datta, R., & DeYoe, E. A. (2009). I know where you are secretly attending! The topography of human visual attention revealed with fMRI. *Vision Research*, *49*(10), 1037–1044.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222. doi:10.1146/annurev.ne.18.030195.001205

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, *115*(2), 107–117.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. doi:10.1016/j.neuron.2012.01.010

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470–2473. doi:10.1126/science.1063414

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. *Cognitive Neuropsychology*, *24*(4), 419–430. doi:10.1080/02643290701380491

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psycholology: General*, *113*(4), 501–517.

Enroth-Cugell, C., & Robson, J. G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. *Journal of Physiology*, *187*(3), 517–552.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433. doi:10.1038/415429a

Ester, E. F., Sutterer, D. W., Serences, J. T., & Awh, E. (2016). Feature-selective attentional modulations in human frontoparietal cortex. *Journal of Neuroscience*, *36*(31), 8188–8199. doi:10.1523/JNEUROSCI.3935-15.2016

Fang, F., Boyaci, H., Kersten, D., & Murray, S. O. (2008). Attention-dependent representation of a size illusion in human V1. *Current Biology*, *18*(21), 1707–1712. doi:10.1016/j.cub.2008.09.025

Farah, M. J. (2004). *Visual agnosia* (2nd ed.). Cambridge, MA: MIT Press.

Farah, M. J., Wilson, K. D., Drain, H. M., & Tanaka, J. R. (1995). The inverted face inversion effect in prosopagnosia: Evidence for mandatory, face-specific perceptual mechanisms. *Vision Research*, *35*(14), 2089–2093.

Fechner, G. T. (1860). *Elemente der psychophysik*. Leipzig, Germany: Breitkopf und Härtel.

Ferster, D., Chung, S., & Wheat, H. (1996). Orientation selectivity of thalamic input to simple cells of cat visual cortex. *Nature*, *380*(6571), 249–252. doi:10.1038/380249a0

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America*, *4*(12), 2379–2394.

Field, G. D., & Chichilnisky, E. J. (2007). Information processing in the primate retina: Circuitry and coding. *Annual Review of Neuroscience*, *30*, 1–30. doi:10.1146/annurev.neuro.30.051606.094252

Field, G. D., Gauthier, J. L., Sher, A., Greschner, M., Machado, T. A., Jepson, L. H., … Chichilnisky, E. J. (2010). Functional connectivity in the retina at the resolution of photoreceptors. *Nature*, *467*(7316), 673–677. doi:10.1038/nature09424

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851. doi:10.1126/science.1194908

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, *12*(9), 1187–1196. doi:10.1038/nn.2363

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. doi:10.1098/rstb.2005.1622

Fukushima, K., & Miyake, S. (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition*, *15*(6), 455–469.

Gandhi, S. P., Heeger, D. J., & Boynton, G. M. (1999). Spatial attention affects brain activity in human primary visual cortex. *Proceedings of the National Academy of Sciences, USA*, *96*(6), 3314–3319.

Garcia, J. O., Srinivasan, R., & Serences, J. T. (2013). Near-real-time feature-selective modulations in human cortex. *Current Biology*, *23*(6), 515–522. doi:10.1016/j.cub.2013.02.013

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*(2), 191–197. doi:10.1038/72140

Gauthier, I., & Tarr, M. J. (1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, *37*(12), 1673–1682.

Germine, L. T., Duchaine, B., & Nakayama, K. (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. *Cognition*, *118*(2), 201–210. doi:10.1016/j.cognition.2010.11.002

Golarai, G., Ghahremani, D. G., Whitfield-Gabrieli, S., Reiss, A., Eberhardt, J. L., Gabrieli, J. D., & Grill-Spector, K. (2007). Differential development of high-level visual cortex correlates with category-specific recognition memory. *Nature Neuroscience*, *10*(4), 512–522. doi:10.1038/nn1865

Gomez, J., Barnett, M. A., Natu, V., Mezer, A., Palomero-Gallagher, N., Weiner, K. S., … Grill-Spector, K. (2017). Microstructural proliferation in human cortex is coupled with the development of face processing. *Science*, *355*(6320), 68–71. doi:10.1126/science.aag0311

Gomez, J., Pestilli, F., Witthoft, N., Golarai, G., Liberman, A., Poltoratski, S., … Grill-Spector, K. (2015). Functionally defined white matter reveals segregated pathways in human ventral temporal cortex associated with category-specific processing. *Neuron*, *85*(1), 216–227. doi:10.1016/j.neuron.2014.12.027

Goodale, M. A., & Westwood, D. A. (2004). An evolving view of duplex vision: Separate but interacting cortical pathways for perception and action. *Current Opinion in Neurobiology*, *14*(2), 203–211. doi:10.1016/j.conb.2004.03.002

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261–2271.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10–11), 1409–1422.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron*, *24*(1), 187–203.

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536–548. doi:10.1038/nrn3747

Guclu, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014. doi:10.1523/JNEUROSCI.5023-14.2015

Hadjikhani, N., Liu, A. K., Dale, A. M., Cavanagh, P., & Tootell, R. B. (1998). Retinotopy and color sensitivity in human visual cortical area V8. *Nature Neuroscience*, *1*(3), 235–241.

Harley, E. M., Pope, W. B., Villablanca, J. P., Mumford, J., Suh, R., Mazziotta, J. C., . . . Engel, S. A. (2009). Engagement of fusiform cortex and disengagement of lateral occipital cortex in the acquisition of radiological expertise. *Cerebral Cortex*, *19*(11), 2746–2754. doi:10.1093/cercor/bhp051

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233.

He, K. M., Zhang, X. Y., Ren, S. Q., & Sun, J. (2016). *Deep residual learning for image recognition*. Paper presented at the Computer Vision and Pattern Recognition, Las Vegas, NV.

Hecht, S., Shlaer, S., & Pirenne, M. H. (1941). Energy at the threshold of vision. *Science*, *93*, 585–558.

Hegde, J., & Van Essen, D. C. (2000). Selectivity for complex shapes in primate visual area V2. *Journal of Neuroscience*, *20*(5), RC61.

Heinze, H. J., Mangun, G. R., Burchert, W., Hinrichs, H., Scholz, M., Münte, T. F., . . . Hillyard, S. A. (1994). Combined spatial and temporal imaging of brain activity during visual selective attention in humans. *Nature*, *372*(6506), 543–546. doi:10.1038/372543a0

Herrmann, K., Heeger, D. J., & Carrasco, M. (2012). Feature-based attention enhances performance by increasing response gain. *Vision Research*, *74*, 10–20. doi:10.1016/j.visres.2012.04.016

Herrmann, K., Montaser-Kouhsari, L., Carrasco, M., & Heeger, D. J. (2010). When size matters: Attention affects performance by contrast or response gain. *Nature Neuroscience*, *13*(12), 1554–1559. doi:10.1038/nn.2669

Hilgetag, C. C., Theoret, H., & Pascual-Leone, A. (2001). Enhanced visual spatial attention ipsilateral to rTMS-induced "virtual lesions" of human parietal cortex. *Nature Neuroscience*, *4*(9), 953–957. doi:10.1038/nn0901-953

Hill, H., Schyns, P. G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition*, *62*(2), 201–222.

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience*, *3*(1), 80–84. doi:10.1038/71152

Hubel, D. H. (1982). Exploration of the primary visual cortex, 1955–78. *Nature*, *299*(5883), 515–524.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224. doi:10.1016/j.neuron.2012.10.014

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts

of visual attention. *Vision Research*, *40*(10–12), 1489–1506.

James, W. (1890/1981). *The principles of psychology*. Cambridge, MA: Harvard University Press.

Jehee, J. F., Ling, S., Swisher, J. D., van Bergen, R. S., & Tong, F. (2012). Perceptual learning selectively refines orientation representations in early visual cortex. *Journal of Neuroscience*, *32*(47), 16747–16753. doi:10.1523/JNEUROSCI.6112-11.2012

Jones, H. E., Andolina, I. M., Ahmed, B., Shipp, S. D., Clements, J. T., Grieve, K. L., . . . Sillito, A. M. (2012). Differential feedback modulation of center and surround mechanisms in parvocellular cells in the visual thalamus. *Journal of Neuroscience*, *32*(45), 15946–15951.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685.

Kamitani, Y., & Tong, F. (2006). Decoding seen and attended motion directions from activity in the human visual cortex. *Current Biology*, *16*(11), 1096–1102.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–4311.

Kanwisher, N., Tong, F., & Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition*, *68*(1), B1–B11.

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, *22*(4), 751–761.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355. doi:nature06713 [pii] 10.1038/nature06713

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304. doi:10.1146/annurev.psych.55.090902.142005

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915. doi:10.1371/journal.pcbi.1003915

Kietzmann, T. C., Poltoratski, S., Konig, P., Blake, R., Tong, F., & Ling, S. (2015). The cccipital face area is causally involved in facial viewpoint perception. *Journal of Neuroscience*, *35*(50), 16398–16403. doi:10.1523/JNEUROSCI.2493-15.2015

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. doi:10.1016/j.tins.2004.10.007

Kok, P., Bains, L. J., van Mourik, T., Norris, D. G., & de Lange, F. P. (2016). Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Current Biology*, *26*(3), 371–376. doi:10.1016/j.cub.2015.12.038

Kondo, S., & Ohki, K. (2016). Laminar differences in the orientation selectivity of geniculate afferents in mouse primary visual cortex. *Nature Neuroscience*, *19*(2), 316–319. doi:10.1038/nn.4215

Kourtzi, Z., Tolias, A. S., Altmann, C. F., Augath, M., & Logothetis, N. K. (2003). Integration of local features into global shapes: Monkey and human fMRI studies. *Neuron*, *37*(2), 333–346.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., . . . Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141. doi:S0896-6273(08)00943-4[pii]10.1016/j.neuron.2008.10.043

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet classification with deep convolutional neural networks*. Paper presented at the Advances in Neural Information Processing Systems, Lake Tahoe, NV.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, *16*(1), 37–68.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neuroscience*, *23*(11), 571–579.

Lamme, V. A., Zipser, K., & Spekreijse, H. (1998). Figure-ground activity in primary visual cortex is suppressed by anesthesia. *Proceedings of the National Academy of Sciences, USA*, *95*(6), 3263–3268.

Larsson, J., & Heeger, D. J. (2006). Two retinotopic visual areas in human lateral occipital cortex. *Journal of Neuroscience*, *26*(51), 13128–13142. doi:10.1523/JNEUROSCI.1657-06.2006

Le, Q., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G., . . . Ng, A. (2012). *Building high-level features using large scale unsupervised learning*. Paper presented at the International Conference in Machine Learning, Edinburgh, Scotland.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi:10.1038/nature14539

Lee, T. S., & Nguyen, M. (2001). Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences, USA*, *98*(4), 1907–1911. doi:10.1073/pnas.031579998

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, *442*(7102), 572–575. doi:10.1038/nature04951

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*(1), 89–94. doi:10.1038/82947

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, *6*(1), 9–16.

Lien, A. D., & Scanziani, M. (2013). Tuned thalamic excitation is amplified by visual cortical circuits. *Nature Neuroscience*, *16*(9), 1315–1323. doi:10.1038/nn.3488

Lind, O., & Kelber, A. (2011). The spatial tuning of achromatic and chromatic vision in budgerigars. *Journal of Vision*, *11*(7), 2. doi:10.1167/11.7.2

Ling, S., & Carrasco, M. (2006). Sustained and transient covert attention enhance the signal via different contrast response functions. *Vision Research*, *46*(8–9), 1210–1220. doi:10.1016/j.visres.2005.05.008

Ling, S., Pratte, M. S., & Tong, F. (2015). Attention alters orientation processing in the human lateral geniculate nucleus. *Nature Neuroscience*, *18*(4), 496–498. doi:10.1038/nn.3967

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: An MEG study. *Nature Neuroscience*, *5*(9), 910–916. doi:10.1038/nn909

Liu, T., Hospadaruk, L., Zhu, D. C., & Gardner, J. L. (2011). Feature-specific attentional priority signals in human cortex. *Journal of Neuroscience*, *31*(12), 4484–4495. doi:10.1523/JNEUROSCI.5745-10.2011

Liu, T., Slotnick, S. D., Serences, J. T., & Yantis, S. (2003). Cortical mechanisms of feature-based attentional control. *Cerebral Cortex*, *13*(12), 1334–1343.

Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nature Neuroscience*, *8*(10), 1386–1390. doi:10.1038/nn1538

Logothetis, N. K., Pauls, J., Bulthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology*, *4*(5), 401–414.

Lorenc, E. S., Pratte, M. S., Angeloni, C. F., & Tong, F. (2014). Expertise for upright faces improves the precision but not the capacity of visual working memory. *Attention, Perception & Psychophysics*, *76*(7), 1975–1984. doi:10.3758/s13414-014-0653-z

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *Journal of Neurophysiology*, *77*(1), 24–42.

Luck, S. J., Woodman, G. F., & Vogel, E. K. (2000). Event-related potential studies of attention. *Trends in Cognitive Sciences*, *4*(11), 432–440.

Marcum, J. I. (1947). *A statistical theory of target detection by pulsed radar*. Technical Report.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: W. H. Freeman & Co.

Masland, R. H. (2012). The neuronal organization of the retina. *Neuron*, *76*(2), 266–280. doi:10.1016/j.neuron.2012.10.002

Maunsell, J. H., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences*, *29*(6), 317–322. doi:10.1016/j.tins.2006.04.001

McAlonan, K., Cavanaugh, J., & Wurtz, R. H. (2008). Guarding the gateway to cortex with attention in visual thalamus. *Nature*, *456*(7220), 391–394. doi:10.1038/nature07382

McCarthy, G., Puce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience*, *9*(5), 605–610.

McKone, E., & Yovel, G. (2009). Why does picture-plane inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? Toward a new theory of holistic processing. *Psychonomic Bulletin & Review*, *16*(5), 778–797. doi:10.3758/PBR.16.5.778

McDonald, J. S., Seymour, K. J., Schira, M. M., Spehar, B., & Clifford, C. W. (2009). Orientation-specific contextual modulation of the fMRI BOLD response to luminance and chromatic gratings in human visual cortex. *Vision Research*, *49*(11), 1397–1405. doi:10.1016/j.visres.2008.12.014

McGugin, R. W., Gatenby, J. C., Gore, J. C., & Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. *Proceedings of the National Academy of Sciences, USA*, *109*(42), 17063–17068. doi:10.1073/pnas.1116333109

McKeeff, T. J., & Tong, F. (2007). The timing of perceptual decisions for ambiguous face stimuli in the human ventral visual cortex. *Cerebral Cortex*, *17*(3), 669–678.

McMains, S. A., & Somers, D. C. (2004). Multiple spotlights of attentional selection in human visual cortex. *Neuron*, *42*(4), 677–686.

Meadows, J. C. (1974). The anatomical basis of prosopagnosia. *Journal of Neurology, Neurosurgery, and Psychiatry*, *37*(5), 489–501.

Mendola, J. D., Dale, A. M., Fischl, B., Liu, A. K., & Tootell, R. B. (1999). The representation of illusory and real contours in human cortical visual areas revealed by functional magnetic resonance imaging. *Journal of Neuroscience*, *19*(19), 8560–8572.

Meng, M., Remus, D. A., & Tong, F. (2005). Filling-in of visual phantoms in the human brain. *Nature Neuroscience*, *8*(9), 1248–1254. doi:10.1038/nn1518

Merigan, W. H., Nealey, T. A., & Maunsell, J. H. (1993). Visual effects of lesions of cortical area V2 in macaques. *Journal of Neuroscience*, *13*(7), 3180–3191.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, *32*, 89–115.

Moeller, S., Freiwald, W. A., & Tsao, D. Y. (2008). Patches with links: A unified system for processing faces in the macaque temporal lobe. *Science*, *320*(5881), 1355–1359. doi:10.1126/science.1157436

Moore, T., & Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, *421*(6921), 370–373. doi:10.1038/nature01341

Moore, T., & Fallah, M. (2001). Control of eye movements and spatial attention. *Proceedings of the National Academy of Sciences, USA*, *98*(3), 1273–1276. doi:10.1073/pnas.021549498

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience*, *9*(5), 555–604. doi:10.1162/jocn.1997.9.5.555

Muller, N. G., & Kleinschmidt, A. (2003). Dynamic interaction of object- and space-based attention in retinotopic visual areas. *Journal of Neuroscience*, *23*(30), 9812–9816. doi:23/30/9812 [pii]

Murray, S. O., Boyaci, H., & Kersten, D. (2006). The representation of perceived angular size in human primary visual cortex. *Nature Neuroscience*, *9*(3), 429–434. doi:10.1038/nn1641

Nakayama, K., & Mackeben, M. (1989). Sustained and transient components of focal visual attention. *Vision Research*, *29*(11), 1631–1647.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding

in fMRI. *NeuroImage*, *56*(2), 400–410. doi: S1053-8119(10)01065-7 [pii] 10.1016/j.neuroimage.2010.07.073

Ni, A. M., Murray, S. O., & Horwitz, G. D. (2014). Object-centered shifts of receptive field positions in monkey primary visual cortex. *Current Biology*, *24*(14), 1653–1658. doi:10.1016/j.cub.2014.06.003

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience*, *15*, 15.

O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. *Nature*, *401*(6753), 584–587. doi:10.1038/44134

Obermayer, K., & Blasdel, G. G. (1993). Geometry of orientation and ocular dominance columns in monkey striate cortex. *Journal of Neuroscience*, *13*(10), 4114–4129.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609. doi:10.1038/381607a0

Op de Beeck, H. P., Haushofer, J., & Kanwisher, N. G. (2008). Interpreting fMRI data: Maps, modules and dimensions. *Nature Reviews Neuroscience*, *9*(2), 123–135. doi:10.1038/nrn2314

Orban, G. A., Van Essen, D., & Vanduffel, W. (2004). Comparative mapping of higher visual areas in monkeys and humans. *Trends in Cognitive Sciences*, *8*(7), 315–324. doi:10.1016/j.tics.2004.05.009

Paik, S. B., & Ringach, D. L. (2011). Retinal origin of orientation maps in visual cortex. *Nature Neuroscience*, *14*(7), 919–925. doi:10.1038/nn.2824

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, *5*(4), 291–303. doi:10.1038/nrn1364

Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area V4. *Nature Neuroscience*, *5*(12), 1332–1338. doi:10.1038/nn972

Perrett, D. I., Oram, M. W., Harries, M. H., Bevan, R., Hietanen, J. K., Benson, P. J., & Thomas, S. (1991). Viewer-centred and object-centred coding of heads in the macaque temporal cortex. *Experimental Brain Research*, *86*(1), 159–173.

Pinel, P., Lalanne, C., Bourgeron, T., Fauchereau, F., Poupon, C., Artiges, E., . . . Dehaene, S. (2015). Genetic and environmental influences on the visual word form and fusiform face areas. *Cerebral Cortex*, *25*(9), 2478–2493. doi:10.1093/cercor/bhu048

Pitcher, D., Charles, L., Devlin, J. T., Walsh, V., & Duchaine, B. (2009). Triple dissociation of faces, bodies, and objects in extrastriate cortex. *Current Biology*, *19*(4), 319–324. doi:10.1016/j.cub.2009.01.007

Pitcher, D., Walsh, V., & Duchaine, B. (2011). The role of the occipital face area in the cortical face perception network. *Experimental Brain Research*, *209*(4), 481–493. doi:10.1007/s00221-011-2579-1

Pitcher, D., Walsh, V., Yovel, G., & Duchaine, B. (2007). TMS evidence for the involvement of the right occipital face area in early face processing. *Current Biology*, *17*(18), 1568–1573. doi:10.1016/j.cub.2007.07.063

Polonsky, A., Blake, R., Braun, J., & Heeger, D. J. (2000). Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nature Neuroscience*, *3*(11), 1153–1159.

Poort, J., Raudies, F., Wannig, A., Lamme, V. A., Neumann, H., & Roelfsema, P. R. (2012). The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. *Neuron*, *75*(1), 143–156. doi:10.1016/j.neuron.2012.04.032

Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology*, *109*(2), 160–174.

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience*, *26*, 381–410. doi:10.1146/annurev.neuro.26.041002.131112

Priebe, N. J., & Ferster, D. (2008). Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. *Neuron*, *57*(4), 482–497. doi:10.1016/j.neuron.2008.02.005

Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience*, *18*(6), 2188–2199.

Qiu, F. T., & von der Heydt, R. (2005). Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. *Neuron*, *47*(1), 155–166. doi:10.1016/j.neuron.2005.05.028

Rangarajan, V., Hermes, D., Foster, B. L., Weiner, K. S., Jacques, C., Grill-Spector, K., & Parvizi, J. (2014). Electrical stimulation of the left and right human fusiform gyrus causes different effects in conscious face perception. *Journal of Neuroscience*, *34*(38), 12828–12836. doi:10.1523/JNEUROSCI.0527-14.2014

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. doi:10.1038/4580

Ress, D., Backus, B. T., & Heeger, D. J. (2000). Activity in primary visual cortex predicts performance in a visual detection task. *Nature Neuroscience*, *3*(9), 940–945.

Ress, D., & Heeger, D. J. (2003). Neuronal correlates of perception in early visual cortex. *Nature Neuroscience*, *6*(4), 414–420. doi:10.1038/nn1024

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185. doi:10.1016/j.neuron.2009.01.002

Rezlescu, C., Barton, J. J., Pitcher, D., & Duchaine, B. (2014). Normal acquisition of expertise with greebles in two cases of acquired prosopagnosia. *Proceedings of the National Academy of Sciences, USA*, *111*(14), 5123–5128. doi:10.1073/pnas.1317125111

Rhodes, G., Ewing, L., Hayward, W. G., Maurer, D., Mondloch, C. J., & Tanaka, J. W. (2009). Contact and other-race effects in configural and component processing of faces. *British Journal of Psychology*, *100*(Pt 4), 717–728. doi:10.1348/000712608X396503

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025. doi:10.1038/14819

Roelfsema, P. R., Lamme, V. A., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, *395*(6700), 376–381. doi:10.1038/26475

Rotshtein, P., Henson, R. N., Treves, A., Driver, J., & Dolan, R. J. (2005). Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. *Nature Neuroscience*, *8*(1), 107–113. doi:10.1038/nn1370

Sasaki, Y., & Watanabe, T. (2004). The primary visual cortex fills in color. *Proceedings of the National Academy of Sciences, USA*, *101*(52), 18251–18256. doi:10.1073/pnas.0406293102

Saygin, Z. M., Osher, D. E., Koldewyn, K., Reynolds, G., Gabrieli, J. D., & Saxe, R. R. (2011). Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. *Nature Neuroscience*, *15*(2), 321–327. doi:10.1038/nn.3001

Schall, J. D., Perry, V. H., & Leventhal, A. G. (1986). Retinal ganglion cell dendritic fields in old-world monkeys are oriented radially. *Brain Research*, *368*(1), 18–23.

Schneider, K. A., & Kastner, S. (2009). Effects of sustained spatial attention in the human lateral geniculate nucleus and superior colliculus. *Journal of Neuroscience*, *29*(6), 1784–1795. doi:10.1523/JNEUROSCI.4452-08.2009

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, *80*(1–2), 1–46.

Scholte, H. S., Jolij, J., Fahrenfort, J. J., & Lamme, V. A. (2008). Feedforward and recurrent processing in scene segmentation: Electroencephalography and functional magnetic resonance imaging. *Journal of Cognitive Neuroscience*, *20*(11), 2097–2109. doi:10.1162/jocn.2008.20142

Schoups, A., Vogels, R., Qian, N., & Orban, G. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature*, *412*(6846), 549–553.

Schwartz, O., Sejnowski, T. J., & Dayan, P. (2009). Perceptual organization in the tilt illusion. *Journal of Vision*, *9*(4), 19 11–20. doi:10.1167/9.4.19

Scolari, M., & Serences, J. T. (2009). Adaptive allocation of attentional gain. *Journal of Neuroscience*, *29*(38), 11933–11942. doi:10.1523/JNEUROSCI.5642-08.2009

Scolari, M., & Serences, J. T. (2010). Basing perceptual decisions on the most informative sensory neurons. *Journal of Neurophysiology*, *104*(4), 2266–2273. doi:10.1152/jn.00273.2010

Self, M. W., Peters, J. C., Possel, J. K., Reithler, J., Goebel, R., Ris, P., . . . Roelfsema, P. R. (2016). The effects of context and attention on spiking activity in human early visual cortex. *PLoS Biology*, *14*(3), e1002420. doi:10.1371/journal.pbio.1002420

Serences, J. T., & Boynton, G. M. (2007). Feature-based attentional modulations in the absence of direct visual stimulation. *Neuron*, *55*(2), 301–312. doi:10.1016/j.neuron.2007.06.015

Serences, J. T., Saproo, S., Scolari, M., Ho, T., & Muftuler, L. T. (2009). Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. *NeuroImage*, *44*(1), 223–231. doi:S1053-8119(08)00865-3 [pii] 10.1016/j.neuroimage.2008.07.043

Serences, J. T., Schwarzbach, J., Courtney, S. M., Golay, X., & Yantis, S. (2004). Control of object-based attention in human cortex. *Cerebral Cortex*, *14*(12), 1346–1357. doi:10.1093/cercor/bhh095

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426. doi:10.1109/TPAMI.2007.56

Shibata, K., Watanabe, T., Sasaki, Y., & Kawato, M. (2011). Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science*, *334*(6061), 1413–1415. doi:10.1126/science.1212003

Sillito, A. M., Cudeiro, J., & Jones, H. E. (2006). Always returning: Feedback and sensory processing in visual cortex and thalamus. *Trends in Neurosciences*, *29*(6), 307–316. doi:10.1016/j.tins.2006.05.001

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., . . . Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. doi:10.1038/nature16961

Silver, M. A., & Kastner, S. (2009). Topographic maps in human frontal and parietal cortex. *Trends in Cognitive Sciences*, *13*(11), 488–495. doi:10.1016/j.tics.2009.08.005

Skottun, B. C., Bradley, A., Sclar, G., Ohzawa, I., & Freeman, R. D. (1987). The effects of contrast on visual orientation and spatial frequency discrimination: A comparison of single cells and behavior. *Journal of Neurophysiology*, *57*(3), 773–786.

Smith, V. C., & Pokorny, J. (1975). Spectral sensitivity of the foveal cone photopigments between 400 and 500 nm. *Vision Research*, *15*(2), 161–171.

Somers, D. C., Dale, A. M., Seiffert, A. E., & Tootell, R. B. (1999). Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. *Proceedings of the National Academy of Sciences, USA*, *96*(4), 1663–1668.

Srihasam, K., Mandeville, J. B., Morocz, I. A., Sullivan, K. J., & Livingstone, M. S. (2012). Behavioral and anatomical consequences of early versus late symbol training in macaques. *Neuron*, *73*(3), 608–619. doi:10.1016/j.neuron.2011.12.022

Stockman, A., MacLeod, D. I., & Johnson, N. E. (1993). Spectral sensitivities of the human cones. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *10*(12), 2491–2521.

Stoerig, P. (2006). Blindsight, conscious vision, and the role of primary visual cortex. *Progress in Brain Research*, *155*, 217–234. doi:10.1016/S0079-6123(06)55012-5

Strother, L., Lavell, C., & Vilis, T. (2012). Figure-ground representation and its decay in primary visual cortex. *Journal of Cognitive Neuroscience*, *24*(4), 905–914. doi:10.1162/jocn_a_00190

Suematsu, N., Naito, T., Miyoshi, T., Sawai, H., & Sato, H. (2013). Spatiotemporal receptive field structures in retinogeniculate connections of cat. *Frontiers in Systems Neuroscience*, *7*, 103. doi:10.3389/fnsys.2013.00103

Sun, W., Tan, Z., Mensh, B. D., & Ji, N. (2016). Thalamus provides layer 4 of primary visual cortex with orientation- and direction-tuned inputs. *Nature Neuroscience*, *19*(2), 308–315. doi:10.1038/nn.4196

Swisher, J. D., Halko, M. A., Merabet, L. B., McMains, S. A., & Somers, D. C. (2007). Visual topography of human intraparietal sulcus. *Journal of Neuroscience*, *27*(20), 5326–5337. doi:10.1523/JNEUROSCI.0991-07.2007

Szegedy, C., Liu, W., Jia, Y. Q., Sermanet, P., Reed, S., Anguelov, D., . . . Rabinovich, A. (2015). *Going deeper with convolutions*. Paper presented at the Computer Vision and Pattern Recognition, Boston, MA. https://arxiv.org/abs/1409.4842

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*, 401–409.

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience*, *1*(4), 275–277. doi:10.1038/1089

Tinsley, J. N., Molodtsov, M. I., Prevedel, R., Wartmann, D., Espigule-Pons, J., Lauwers, M., & Vaziri, A. (2016). Direct detection of a single photon by humans. *Nature Communications*, *7*, 12172. doi:10.1038/ncomms12172

Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception*, *9*(4), 483–484. doi:10.1068/p090483

Tong, F., & Engel, S. A. (2001). Interocular rivalry revealed in the human cortical blind-spot representation. *Nature*, *411*(6834), 195–199.

Tong, F., Nakayama, K., Moscovitch, M., Weinrib, O., & Kanwisher, N. (2000). Response properties of the human fusiform face area. *Cognitive Neuropsychology*, *17*, 257–279.

Tong, F., Nakayama, K., Vaughan, J. T., & Kanwisher, N. (1998). Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron*, *21*(4), 753–759.

Tong, F., & Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual Review of Psychology*, *63*, 483–509. doi:10.1146/annurev-psych-120710-100412

Tootell, R. B., Reppas, J. B., Dale, A. M., Look, R. B., Sereno, M. I., Malach, R., . . . Rosen, B. R. (1995). Visual motion aftereffect in human cortical area MT revealed by functional magnetic resonance imaging. *Nature*, *375*(6527), 139–141. doi:10.1038/375139a0

Treue, S., & Martinez-Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736), 575–579. doi:10.1038/21176

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*(5761), 670–674. doi:10.1126/science.1119983

Tsao, D. Y., Moeller, S., & Freiwald, W. A. (2008). Comparing face patch systems in macaques and humans. *Proceedings of the National Academy of Sciences, USA*, *105*(49), 19514–19519. doi:10.1073/pnas.0809662105

Usrey, W. M., & Alitto, H. J. (2015). Visual functions of the thalamus. *Annual Review of Vision Science*, *1*, 351–371.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, *79*(4), 471–491.

van Bergen, R. S., Ma, W. J., Pratte, M. S., & Jehee, J. F. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, *18*(12), 1728–1730. doi:10.1038/nn.4150

Vidyasagar, T. R., & Eysel, U. T. (2015). Origins of feature selectivities and maps in the mammalian primary visual cortex. *Trends in Neurosciences*, *38*(8), 475–485. doi:10.1016/j.tins.2015.06.003

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, *287*(5456), 1273–1276.

Yovel, G., & Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Current Biology*, *15*(24), 2256–2262. doi:10.1016/j.cub.2005.10.072

Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron*, *56*(2), 366–383. doi:10.1016/j.neuron.2007.10.012

Wang, W., Jones, H. E., Andolina, I. M., Salt, T. E., & Sillito, A. M. (2006). Functional alignment of feedback effects from visual cortex to thalamus. *Nature Neuroscience*, *9*(10), 1330–1336. doi:10.1038/nn1768

Wang, Y., Jin, J., Kremkow, J., Lashgari, R., Komban, S. J., & Alonso, J. M. (2015). Columnar organization of spatial phase in visual cortex. *Nature Neuroscience*, *18*(1), 97–103. doi:10.1038/nn.3878

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature*, *428*(6982), 557–561. doi:10.1038/nature02420

Weiner, K. S., & Grill-Spector, K. (2012). The improbable simplicity of the fusiform face area. *Trends in Cognitive Sciences*, *16*(5), 251–254. doi:10.1016/j.tics.2012.03.003

Wenderoth, P., & Johnstone, S. (1988). The different mechanisms of the direct and indirect tilt illusions. *Vision Research*, *28*(2), 301–312.

Westheimer, G. (2003). Meridional anisotropy in visual processing: Implications for the neural site of the oblique effect. *Vision Research*, *43*(22), 2281–2289.

Westheimer, G., & McKee, S. P. (1977). Spatial configurations for visual hyperacuity. *Vision Research*, *17*(8), 941–947.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., . . . Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences, USA*, *107*(11), 5238–5241. doi:10.1073/pnas.0913053107

Winkler, A. D., Spillmann, L., Werner, J. S., & Webster, M. A. (2015). Asymmetries in blue-yellow color perception and in the color of "the dress." *Curr Biol*, *25*(13), R547–548. doi:10.1016/j.cub.2015.05.004

Yacoub, E., Harel, N., & Ugurbil, K. (2008). High-field fMRI unveils orientation columns in humans. *Proceedings of the National Academy of Sciences, USA*, *105*(30), 10607–10612. doi:10.1073/pnas.0804110105

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. doi:10.1038/nn.4244

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, USA*, *111*(23), 8619–8624. doi:10.1073/pnas.1403112111

Yeshurun, Y., & Carrasco, M. (1998). Attention improves or impairs visual performance by enhancing spatial resolution. *Nature*, *396*(6706), 72–75. doi:10.1038/23936

Zeiler, M. D., & Fergus, R. (2014). *Visualizing and understanding convolutional networks*. Paper presented at the European Conference on Computer Vision, Zurich, Switzerland.

Zenger-Landolt, B., & Heeger, D. J. (2003). Response suppression in V1 agrees with psychophysics of surround masking. *Journal of Neuroscience*, *23*(17), 6884–6893.

Zipser, K., Lamme, V. A., & Schiller, P. H. (1996). Contextual modulation in primary visual cortex. *Journal of Neuroscience*, *16*(22), 7376–7389.