Deciphering the Neural Representations of Visual Illusions

Fan Cheng

Supervisor: Prof. Yukiyasu Kamitani

Department of Intelligence Science and Technology
Graduate School of Informatics
Kyoto University

This thesis is submitted for the degree of
Doctor of Philosophy

June 2024

I would like to dedicate this thesis to my loving parents ...

Declaration

I, Fan Cheng, declare that this thesis, entitled "Deciphering the neural representations of visual illusions" is original and my own work. I confirm that:

- This work was done solely while a candidate for the research degree at the Graduate School of Informatics, Kyoto University.
- No part of this work has previously been submitted for a degree at this or any other university.
- References to the work of others have been clearly attributed. Quotations from the work of others have been clearly indicated, and attributed to them.
- In cases where others have contributed to part of this work, such contribution has been clearly acknowledged and distinguished from my own work.

Fan Cheng
June 2024

Acknowledgements

I would like to express my deepest gratitude to Prof. Yukiyasu Kamitani for not only welcoming me into his exceptional research group but also for his unwavering support and guidance throughout this transformative journey. His mentorship has been instrumental in shaping my growth as a researcher and as an individual. Working alongside him and his brilliant team has been an incredible privilege, and I look forward to a lifelong and fruitful collaboration.

I am also immensely grateful to Dr. Tomoyasu Horikawa and Dr. Eizaburo Doi for their invaluable insights and guidance, which have been crucial in navigating the complexities of this field.

To my amazing lab members, I cannot thank you enough for the camaraderie, support, and stimulating discussions that have made this journey so enriching and unforgettable. A special thanks to Haibao Wang, Ken Shirakawa, Dr. Jun Kai Ho, Dr. Jong-Yun Park, Dr. Mohamed Abdelhack, Dr. Misato Tanaka, and Dr. Shuntaro Aoki from the Kyoto University lab, as well as Dr. Mitsuaki Tsukamoto and the fantastic team at ATR.

I would also like to extend my sincere appreciation to our incredible research administrators, Yukari Kado and Yukiko Masuda, for their meticulous handling of the intricacies of our research projects and their unwavering support. To the members of the Kokoro Research Center at Kyoto University, thank you for your invaluable assistance during the experimental processes, ensuring the seamless execution of our studies.

A special thanks to the AI tools, particularly ChatGPT and Claude, for their invaluable suggestions on improving the flow, clarity, and consistency of the text, as well as their assistance in addressing grammatical issues.

Finally, to my beloved family and friends, words cannot express how grateful I am for your unwavering belief in me and your constant encouragement. You have been the bedrock of my academic journey, and I share this achievement with you as I proudly graduate from the illustrious Kyoto University.

Abstract

Visual illusions have long captivated the interest of scientists and philosophers alike, as they provide a unique window into the complex neural mechanisms that underlie our subjective perceptual experiences. Despite the growing body of research on the neural correlates of visual illusions, the precise manner in which brain activity gives rise to illusory percepts remains largely unknown. Elucidating this link is crucial for advancing our understanding of how the brain constructs and represents conscious experiences.

Brain activity can be seen as "codes" of our mental contents. Deep neural network (DNN) representations can be regarded as a different form of "codes" or latent features of the same content, which can be translated from brain activity. By leveraging DNN representations as an intermediary, recent reconstruction techniques have been able to reconstruct arbitrary stimulus images from brain activity patterns, providing a powerful means to visualize a coherent representation of visual experiences encoded in brain activity and mapped onto DNN representations.

Building on these advances, we hypothesized that an illusory stimulus would evoke brain activity patterns similar to those induced by a stimulus that matches the subjective appearance of the illusion. Such activity patterns could then be translated into DNN representations and subsequently reconstructed as images that explicitly depict the illusory attributes absent from the original stimulus. This approach offers a novel way to materialize the subjective contents of illusory experiences and shed light on the underlying neural representations.

To test this hypothesis, we conducted a series of experiments using functional magnetic resonance imaging (fMRI) to measure brain activity in human participants viewing natural images and two types of representative visual illusions: illusory contours and neon color spreading. We employed linear regression models trained with natural images to translate the fMRI activity patterns into DNN features, which were then fed into an image generator to visualize the represented content. Quantitative analyses were performed on the reconstructions to assess the presence and strength of illusory features across different regions of the visual cortex.

For illusory contours, the reconstructions revealed a dissociation between lower and higher visual areas. Both the illusory contours and the inducing elements could be accurately

reconstructed from early regions like V1-V3. Higher areas tended to show less localized illusory contours, indicating a transition to more abstract object representations.

Neon color spreading showed distinct patterns for different illusion configurations. For the Ehrenstein configuration, the illusory color was robustly represented across areas, with lower areas precisely encoding the boundaries of the spread color and higher regions representing the filled-in surfaces. The Varin configuration, however, only showed pronounced effects in higher areas, suggesting potential differences in the underlying neural computations. Reconstructed color strength was modulated by the spatial extent and arrangement of inducers, underscoring the importance of contextual integration.

Our findings provide compelling evidence that visual illusions arise from hierarchical neural computations distributed across the visual cortex. The fact that no single brain region could consistently reconstruct the full gamut of illusory features across different scenarios offers nuanced understanding of how conscious experiences are neurally represented and places important limits on theoretical frameworks of consciousness.

Our approach spans the domains of both "inner" psychophysics, which probes the mind through internal representations in the brain, and "outer" psychophysics, which quantifies subjective percepts using physical stimuli. The present findings not only deepen our understanding of how the brain represents perceptual experiences but also lay the groundwork for future explorations of the neural correlates of subjective experience using more naturalistic and diverse scenarios. It also holds promise as a tool for developing novel brain-computer interfaces that can effectively communicate subjective experiences in both basic research and clinical settings.

Table of contents

List of figures

Chapter 1

Introduction

Have you ever wondered why our perception of the world feels so vivid and coherent, despite the inherent ambiguity in the sensory information we receive? Visual perception extends far beyond mere image formation on our retinas.

While our eyes capture light patterns from the environment, this sensory input is often incomplete and ambiguous. Variations in illumination, reflectance, and transmittance can produce identical retinal images from vastly different physical scenarios. Moreover, the same retinal projection could arise from objects with distinct orientations, sizes, and distances. Consequently, retinal signals alone are insufficient to discern the true physical reality before us. To derive coherent perceptions amidst such ambiguity, the brain employs sophisticated neural computations to actively reconstruct and interpret the visual world, rather than passively encoding retinal inputs. Vision, therefore, is an intelligent guessing game, where the brain infers the most likely interpretation of inherently ambiguous sensory data.

Among the captivating phenomena that provide insights into these interpretive processes are visual illusions – instances where our perceptual experiences diverge from the physical reality of the visual stimulus. These illusions are not mere tricks or errors but rather serve as natural experiments that reveal the computational strategies employed by the brain to translate ambiguous sensory signals into coherent perceptual representations.

The central aim of this thesis is to decipher the neural representations that give rise to representative visual illusions such as illusory contours and color. By unraveling how the brain encodes and processes these illusion-inducing stimuli, we can gain a deeper understanding of the mechanisms underlying conscious visual perception and the neural correlates of subjective experience.

Elucidating the neural underpinnings of visual illusions holds profound significance. It can inform and refine theoretical models of visual processing, shed light on the brain's ability to construct stable perceptions from noisy inputs, and potentially inspire novel applications in

fields such as computer vision, user interface design, and visual aids or therapies for clinical conditions.

In this chapter, we will first introduce visual illusion phenomena and give several examples (Section 1.1), providing a foundation for our investigations. We will then explore the existing theories and proposed mechanisms that attempt to account for these illusions (Section 1.2), followed by a discussion on the challenges and approaches to quantifying subjective illusory experiences (Section 1.3). Subsequently, we will outline our specific research methodology combining psychophysics, computational modeling, and neuroimaging techniques (Section 1.4). Finally, we will provide an overview of the remaining chapters in the thesis (Section 1.5).

1.1 Introduction to visual illusions

Visual illusions are subjective perceptions that arise from the brain's interpretation of visual input, leading to experiences that differ from objective measurements or widely accepted norms of that input. These illusions offer unique insights into the complex mechanisms underlying visual perception and serve as valuable tools for probing the neural processes that give rise to conscious experience (Levinson and Baillet, 2022). Visual illusions differ from other subjective visual experiences such as visual imagery and hallucinations in that they are triggered by specific visual stimuli (Gregory, 1997; Ffytche, 2014). In contrast, visual imagery and hallucinations are internally generated and can vary more significantly between individuals (Shine et al., 2011; Pearson, 2019).

Broadly, visual illusions can be categorized based on the primary visual attributes that deviate from the physical stimulus, such as illusory contours, brightness, color, size, and tilt (Figs 1.1. to 1.3). Illusory contours, for instance, are perceived edges that exist in the absence of actual luminance or color gradients (Schumann, 1900; Ehrenstein, 1941; Kanizsa, 1955; Soriano et al., 1996). Brightness and color illusions, on the other hand, involve the misperception of these properties in regions of identical luminance or chromatic values (Munker, 1970; van Tuijl, 1975; White, 1979; Münsterberg, 1894; Bressan et al., 1997; Pinna et al., 2001; Anderson and Winawer, 2005). Size, tilt, and position illusions are characterized by distortions in perceived dimensions, orientations, and alignments, respectively (Zöllner, 1860; Hering, 1861; Müller-Lyer, 1889; Ebbinghaus, 1902; Fraser, 1908; Ponzo, 1911). Motion illusions, on the other hand, create a perception of movement in static images (Schrauf et al., 1997; Kitaoka and Ashida, 2003).

Attempts to sophisticatedly classify illusions have been made based on various factors such as the type of perceptual distortion, the underlying neural mechanisms, or the nature of

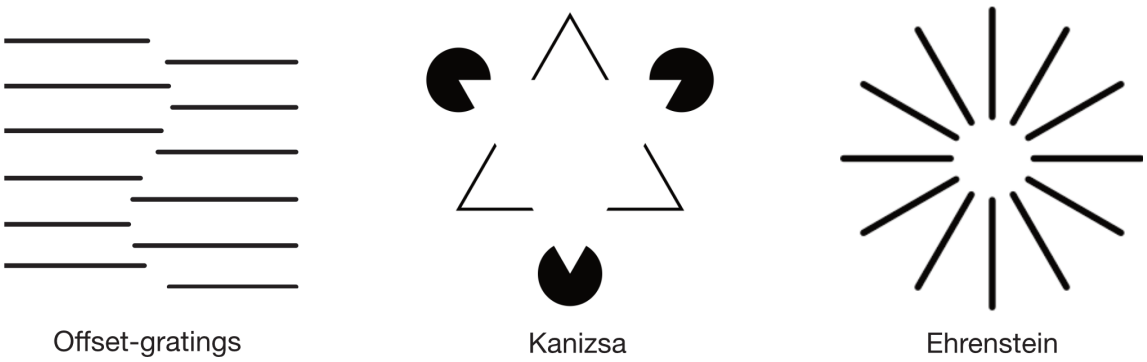

Fig. 1.1 Illusory contours. Examples: offset-gratings (left; Soriano et al., 1996), Kanizsa (middle; Kanizsa, 1955, and Ehrenstein (right; Ehrenstein, 1941).

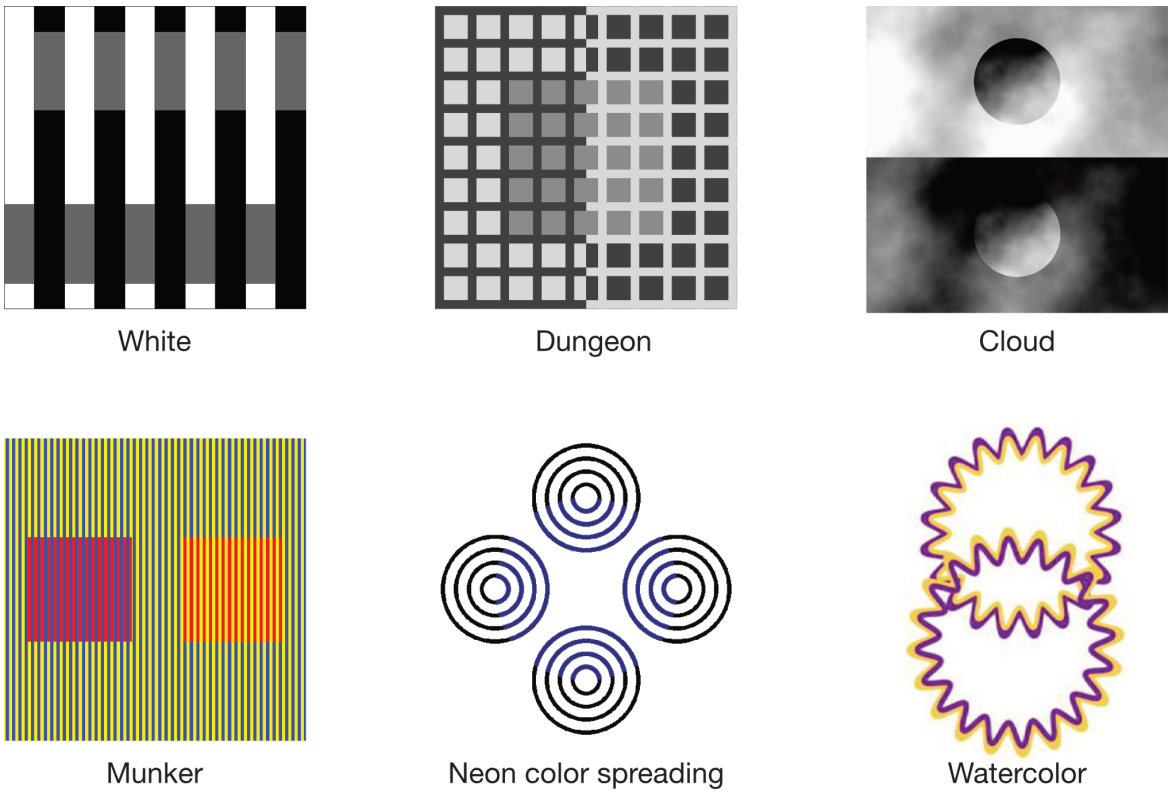

Fig. 1.2 Brightness and color illusions. (Top) Examples of brightness illusions: White's illusion (left; White, 1979), Dungeon illusion (middle; Bressan et al., 1997), and Cloud illusion (right; Anderson and Winawer, 2005). (Bottom) Examples of color illusions: Munker illusion (left; Munker, 1970), neon color spreading (middle; van Tuijl, 1975), and watercolor illusion (right; Pinna et al., 2001).

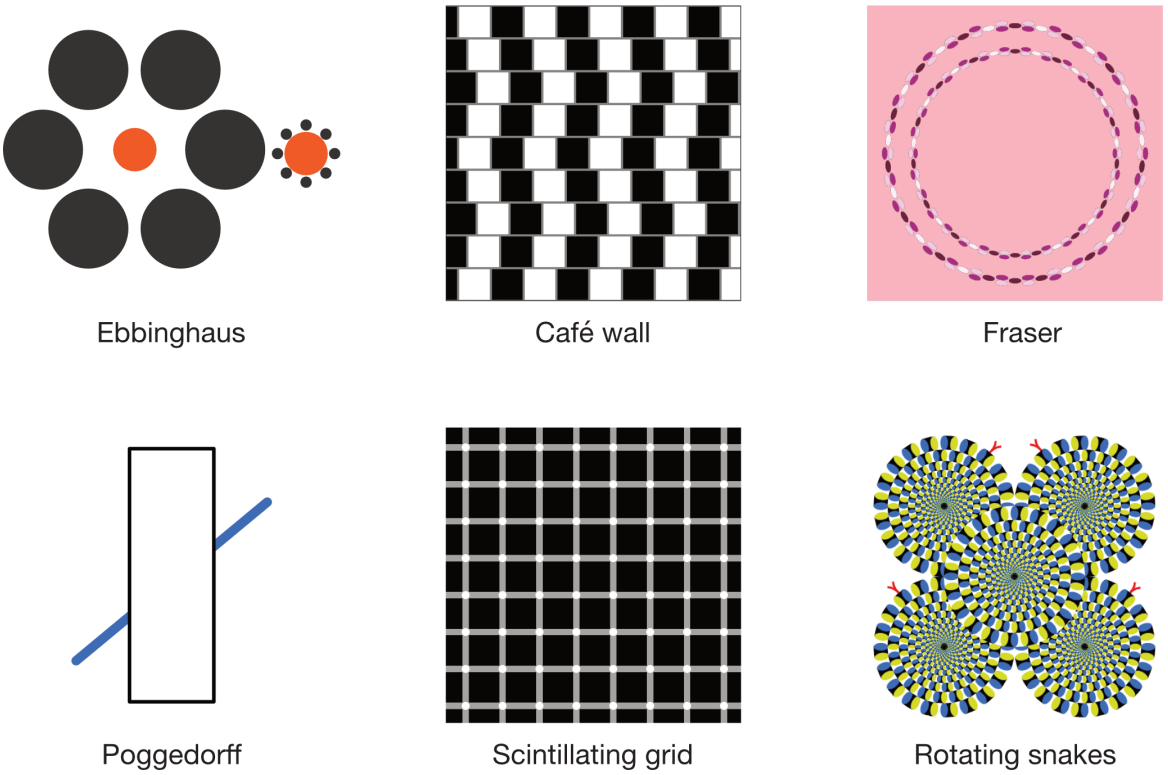

Fig. 1.3 Size, tilt, position, and motion illusions. (Top left) Example of a size illusion: Ebbinghaus illusion (Ebbinghaus, 1902). (Top middle and right) Examples of tilt illusions: Café wall illusion (Münsterberg, 1894) and Fraser spiral illusion (Fraser, 1908). (Bottom left) Example of a position illusion: Scintillating grid illusion (Schrauf et al., 1997). (Bottom middle and right) Examples of motion illusions in static images: Poggendorff illusion (Zöllner, 1860) and Rotating snakes illusion (Kitaoka and Ashida, 2003).

the inducing stimulus (Carbon, 2014; Shapiro and Todorović, 2017). However, a comprehensive and widely accepted classification system remains elusive due to the complexity and diversity of illusory phenomena. Moreover, the definitions of illusions have been ambiguous. Depending on how we define the reality of the physical world, the same visual phenomena can be regarded as illusory or non-illusory. In light of this fluid taxonomy, this thesis concentrates on two specific types of visual illusions: illusory contours and neon color spreading (Schumann, 1900; Soriano et al., 1996; van Tuijl, 1975). These illusions, which involve the perception of lines or colors that are not physically present in the stimulus or the signals reaching the retina, are fascinating despite the fluid taxonomy of visual illusions and are considered as such according to various definitions of illusions. Although extensively studied, debates continue about the precise mechanisms behind these phenomena, underlining the need for further research (M. M. Murray and Herrmann, 2013; Grossberg, 2017).

Psychophysical studies have been instrumental in advancing our understanding of visual illusions. These studies can be broadly categorized into two approaches: "outer" psychophysics and "inner" psychophysics (Fechner, 1860). "Outer" psychophysics relies on behavioral responses, such as verbal reports or button presses, to characterize the perceptual properties of illusions and identify the stimulus features that drive these perceptual experiences (Braddick, 1995; Lu and Dosher, 2013; Kingdom and Prins, 2016). In contrast, "inner" psychophysics directly probes the neural representations underlying illusory percepts by measuring brain activity, offering a more direct window into the neural mechanisms at play (Kamitani and Tong, 2005; Haynes, 2009; Kriegeskorte and Kreiman, 2011).

Recent advancements in neuroscience, particularly the development of reconstruction techniques (Miyawaki et al., 2008; Shen et al., 2019a), have opened up new possibilities for applying "inner" psychophysical methods to the study of visual illusions. These techniques allow us to reconstruct perceptual experiences from brain activity, enabling direct comparisons between the neural representations of illusory percepts and those of physically matched stimuli. This approach holds promise for resolving long-standing debates and providing novel insights into the neural basis of illusory experience.

In the following sections, together with Chapters 2 and 3, we will delve deeper into visual illusions, especially the types of illusions of interest, review the existing literature on their proposed mechanisms and measurement techniques, and outline our approach to investigating their neural underpinnings using a combination of psychophysical, computational, and neuroimaging techniques.

1.2 Mechanisms behind visual illusions

This section will introduce general theories of visual perception and computational models for visual illusions, reviewing how these theories and models have been applied to explain the mechanisms behind visual illusions.

1.2.1 General theories of visual perception

Three theories that have contributed to providing principled explanations for visual perception and illusions include efficient coding, the Bayesian brain hypothesis, and predictive coding. These theories offer complementary perspectives on how the brain optimizes its resources, makes inferences based on prior knowledge, and generates predictions to facilitate efficient processing of sensory information. Examining these theories can provide insights into the mechanisms underlying visual illusions.

Efficient coding

Efficient coding theory posits that the brain optimizes the mutual information between visual inputs and their internal representations, constrained by the brain's information processing capacity. This optimization process involves reducing redundancy within the visual stimuli, thereby ensuring that neural responses are as distinct and informative as possible. The origins of this theory trace back to Attneave (1954) and Barlow (1961), who highlighted the redundancy of visual stimuli and the brain's propensity to recode sensory inputs into more efficient representations. The theory has garnered empirical support, notably through its predictions about the sparse coding observed in the primary visual cortex (Olshausen and Field, 1996, 1997, 2004, 2005). Sparse coding in the primary visual cortex is characterized by neurons responding selectively to specific visual features, such as edges or orientations. This selective response is thought to maximize information transmission while minimizing redundancy, aligning with the principles of efficient coding (Olshausen and Field, 1996). Subsequent research has extended these concepts to the analysis of neural responses to natural stimuli (Simoncelli and Olshausen, 2001) and the development of computational models like independent component analysis (ICA) in machine learning (Bell and Sejnowski, 1997).

Application to visual illusions

Efficient coding theory offers a framework for understanding how visual illusions emerge from the brain's efforts to optimize visual processing under specific contexts. For example, in the tilt illusion, contextual elements and the target are spatially distinct yet temporally

concurrent, leading to perceptual effects that deviate from the actual stimulus (Wenderoth and Johnstone, 1988; Clifford et al., 2000). Neural tuning curve adjustments induced by contextual stimuli, such as suppression, width change, and shift, can collectively result in visual illusions, aligning with efficient coding principles that suggest the brain prioritizes coding for orientations with higher probability, optimizing the use of neural resources and reducing redundancy through decorrelation (Schwartz et al., 2007). Models like the Wilson–Cowan (WC) type (Cowan et al., 2016), which emulate neural dynamics, have successfully reproduced tilt and brightness illusions (Bertalmío et al., 2020). Another model, grounded in the efficient coding principle, has employed spatio-chromatic filters to predict the appearance of approximately 50 brightness and color illusory images (Troscianko and Osorio, 2023).

Efficient coding theory's relevance to visual illusions can be further elucidated through the lens of Bayesian inference (Kersten et al., 2004; Knill and Pouget, 2004; Sohn, 2021). This theoretical framework posits that the brain integrates prior knowledge with current sensory information to generate a perceptual estimate, a process that may lead to visual illusions. Studies demonstrate how the brain's neural architecture could be optimally arranged to reflect prior probabilities, leading to more efficient information processing and sometimes resulting in perceptual distortions or illusions (Ganguli and Simoncelli, 2014; Girshick et al., 2011; Harrison et al., 2023). These studies suggest that the brain's adaptive mechanisms, geared towards optimizing visual processing, can occasionally give rise to the distorted perceptions seen in visual illusions.

Bayesian brain hypothesis

While efficient coding theory provides a framework for understanding how the brain optimizes visual processing, the Bayesian brain hypothesis offers a complementary perspective, emphasizing the role of probabilistic inference in perception. It posits that perception can be regarded as a generative process from internal models using Bayesian probability theory (Kersten et al., 2004; Knill and Pouget, 2004).

The Bayesian brain hypothesis suggests that the internal generative model of the environment implemented by the brain specifies how to generate and predict sensory data from hidden states. The hidden variables can be sampled from a prior distribution, and the sensory data are sampled from a distribution conditioned on the hidden variables. The degree of plausibility is quantified by the likelihood, which is the probability of sensory data given a specific hidden variable. For instance, the spatial frequency is the hidden variable of the visual inputs.

Given sensory data, perception can be regarded as an inverted process of the generative model that maps from hidden states to sensory data to compute the posterior probability of the hidden states. The probability of hidden variables given the sensory data can be estimated by combining the prior and the likelihood, based on Bayes' rule:

$$P(h \mid s) = \frac{P(s \mid h)P(h)}{P(s)} \tag{1.1}$$

where s and h represent the sensory data (e.g. measurements) and hidden variables (e.g. stimulus), respectively. $P(h|s)$ represents the posterior distribution of hidden variables, and $P(s)$ is the marginal likelihood of the sensory data.

Since the Bayesian brain hypothesis is essentially an abstract "computational-level" hypothesis (Marr, 1982), it is crucial to examine how this theory is implemented by the brain (Sohn, 2021; Walker, 2023). Population neurons have the ability to represent the probability distribution depicted in the Bayesian brain hypothesis; the questions would be what form could be adopted by neurons and how the brain performs tasks by utilizing the posterior distribution or estimates in vivo.

The concept of hierarchical Bayesian inference, particularly in the visual cortex, provides a foundational framework for understanding how the brain might implement Bayesian principles (T. S. Lee and Mumford, 2003). This approach suggests that early visual areas such as V1 and V2 participate in dynamic processing, influenced by both feedforward sensory data and feedback contextual information, challenging the traditional linear, feedforward view of visual processing.

There are two broad classes of neural implementations for Bayesian inference (Sohn, 2021): the modular and transform perspectives (Figure 1.4). The modular perspective posits that prior and likelihood are encoded as independent entities, supported by empirical evidence from studies on probabilistic population coding (Sanger, 1996; Pouget et al., 2003; Ma et al., 2006; Fetsch et al., 2012; Akrami et al., 2018; Darlington et al., 2018; Hou et al., 2019; Walker, 2020; Singletary et al., 2024) and sampling codes (Fiser et al., 2010; Echeveste et al., 2020). Contrary to the modular perspective, the transform view argues for a direct mapping of uncertain sensory measurements into Bayesian estimates (Simoncelli, 2009; Jazayeri and Shadlen, 2010; Nessler et al., 2013; Ganguli and Simoncelli, 2014; Narain et al., 2018; Gershman, 2019). The transform view can find support in empirical studies of visual behavior data and animal neural recordings, suggesting the latent encoding of priors in cortical circuits or computation through latent dynamics (Girshick et al., 2011; Narain et al., 2018). Together, these studies pave the way for investigating how the Bayesian brain hypothesis is implemented within the neural architecture, with disputes unsettled.

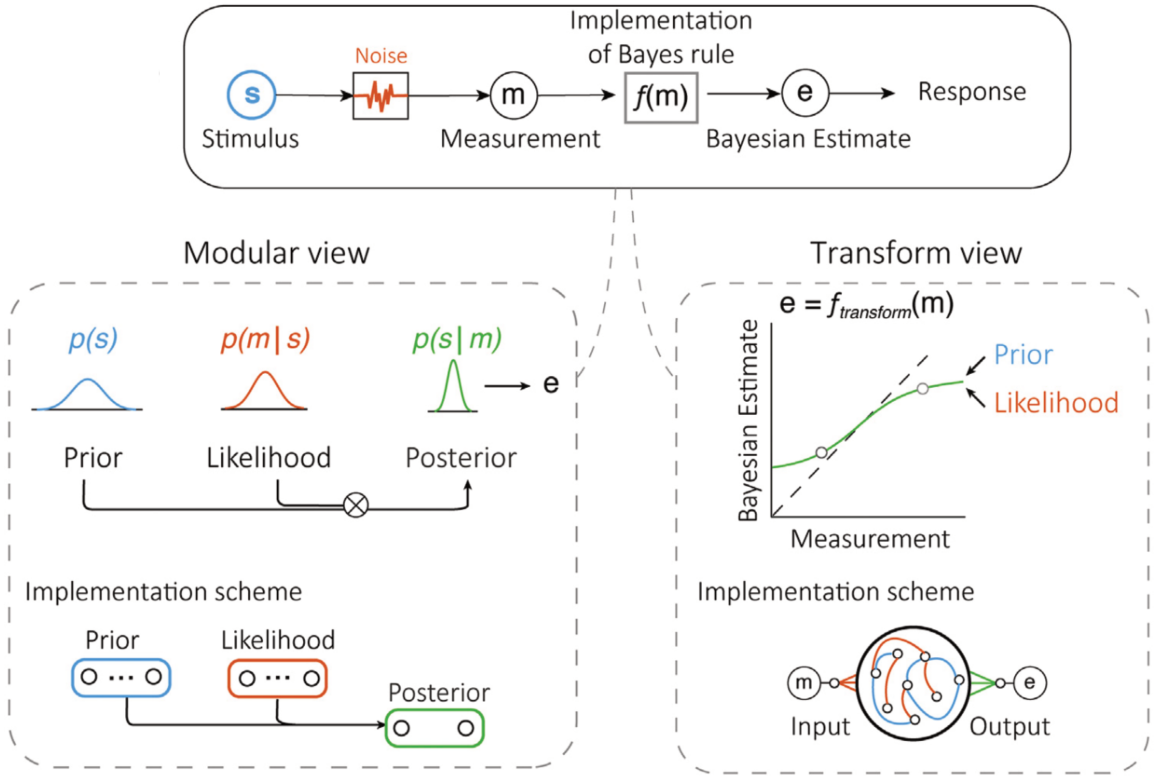

Fig. 1.4 Two views for the Neural implementation of Bayesian inference. Adapted from Sohn (2021).

Application to visual illusions

Both veridical and illusory perceptions are products of the brain's inferential processes, which combine sensory signals with prior expectations to resolve the 'inverse optics' problem (Nour and Nour, 2015). In this view, illusions are natural outcomes of the brain's attempt to infer the most probable state of the world based on incomplete or misleading sensory data.

Bayesian methods have been extensively employed to elucidate visual illusions, which are presumed to be influenced by our priors and internal world models. Examples include color constancy (Brainard, 2006), suppression and filling-in phenomena (Zhaoping and Jingling, 2008), and the brightness illusion (Brown and Friston, 2012). In the context of color constancy, Brainard and colleagues incorporate a Bayesian framework for illuminant estimation, tying closely with psychophysical data to simulate human color perception under varying lighting conditions, demonstrating how prior knowledge in lighting conditions shapes perceptual experience (Brainard, 2006).

However, not all visual illusions align neatly with Bayesian predictions (Gregory, 2006; Anderson et al., 2011). Anderson et al. present evidence of contour synthesis that defy Bayesian explanation, suggesting that while Bayesian models are powerful, they may not capture the full spectrum of perceptual experiences, indicating the need for broader computational frameworks to fully understand the synthesis of visual illusions (Anderson et al., 2011).

These studies collectively underscore the applicability of the Bayesian brain hypothesis to explaining visual illusions, while also highlighting the limitations and the need for nuanced approaches to encompass the diversity of perceptual phenomena encountered in the visual domain.

Predictive coding

Predictive coding has emerged as a prominent theoretical framework for understanding how the brain processes sensory information, particularly in the context of visual perception. The core idea behind predictive coding is that the brain actively generates predictions about incoming sensory data based on prior knowledge and expectations, and it constantly compares these predictions with actual sensory input to minimize prediction errors (Rao and Ballard, 1999; Friston, 2005; Y. Huang and Rao, 2011; Clark, 2013).

The concept of predictive coding suggests that the brain employs a hierarchical generative model, where each level of the hierarchy predicts the activity of the level below it (Dayan et al., 1995; Rao and Ballard, 1999; Jiang and Rao, 2024; Rao et al., 2023; Sprevak, 2024). The difference between the predicted and actual activity, known as the prediction error, is then propagated back up the hierarchy to update the higher-level predictions. This iterative

process of feedforward prediction error and feedback prediction update allows the brain to refine its internal model of the world and adapt to changes in sensory input. Empirical evidence from neurophysiological studies in the visual cortex supports the existence of such hierarchical predictive processing, with findings demonstrating the presence of prediction error and prediction update signals in various cortical areas (Kok and de Lange, 2015; Shipp, 2016; Millidge et al., 2022; Shipp, 2024). However, some researchers have pointed out that the general opinion of predictions excludes the possibility that the predictive information is embedded in the bottom-up signals; it thus requires further examination of the form of the predictions in the brain system (Teufel and Fletcher, 2020; Millidge et al., 2022).

Application to visual illusions

The predictive coding framework has been successfully applied to explain a range of visual illusions, providing an account of how these phenomena arise from the brain's efforts to reconcile sensory input with its internal model of the world. For instance, predictive coding has been employed to explain illusory contours (Lotter et al., 2020; Pang et al., 2021), brightness illusion (Brown and Friston, 2012), motion illusion such as the Rotating Snakes illusion (Watanabe et al., 2018; Kobayashi et al., 2022) and the flash-grab illusion (van Heusden et al., 2019). Through the flash-grab illusion, where the perceived position of an object is influenced by sudden motion reversals, predictive coding has been demonstrated to operate at both monocular and binocular stages of visual processing, contributing to the perception of the illusion. Several studies have shown that the illusory contours and motion could be reproduced by the networks trained on predicting the next movie frame, supporting the idea that predictive coding mechanisms underlie the generation of such illusions (Watanabe et al., 2018; Lotter et al., 2020; Pang et al., 2021).

While predictive coding offers a compelling framework for understanding visual illusions, there are also challenges and limitations to its explanatory power. For example, researchers found inconsistencies in the ability of a predictive coding-based deep neural network to replicate the Rotating Snakes illusion, particularly in grayscale variants, suggesting that predictive coding alone may not fully account for the complexity of motion illusion (Kirubeswaran and Storrs, 2023). Moreover, certain perceptual illusions, such as the Müller-Lyer illusion, pose difficulties for predictive processing accounts, as they seem to persistently defy the error minimization efforts emphasized by these models (Gallagher et al., 2022). Additionally, the activity suppression in early visual cortex during illusory shape perception might be better explained by neural adaptation to perceptually stable input rather than predictive coding (Yan et al., 2021).

In conclusion, predictive coding has proven to be a valuable framework for understanding the neural mechanisms underlying visual illusions. However, the framework also faces challenges in fully capturing the complexity and diversity of illusory experiences, suggesting that further research and potential integration with other theoretical perspectives may be necessary to provide a complete understanding of the neural basis of visual illusions.

Linking theories to explain visual illusions

The three general theories are not mutually exclusive; rather, they have overlapping elements and can be complementary. Efficient coding and predictive coding facilitate the brain's allocation of computational resources to the most informative and relevant aspects of sensory input (Manookin and Rieke, 2023). In fact, the concept of predictive coding was predicted by the principle of efficient coding (Barlow, 1961). Predictive coding minimizes redundancy in both spatial and temporal domains by predicting sensory inputs and only encoding the deviations from these predictions (Y. Huang and Rao, 2011; Manookin and Rieke, 2023). The predictions concerning current or future sensory inputs, which are requisite for predictive coding, can be computed through Bayesian inference, yielding posterior estimates. Furthermore, efficient coding can impose a constraint on the prior probabilities used in Bayesian inference.

There have been efforts to integrate some of or all these theories into a single framework (Friston and Kiebel, 2009; Friston, 2010; Aitchison and Lengyel, 2017; Chalk et al., 2018; Park and Pillow, 2024), which then applied to explain empirical phenomena such as visual illusions. Under the free energy principle, Bayesian inference can be achieved by predictive coding using specific generative models and variational inference algorithms: under the Gaussian assumption, optimizing the empirical prior is equivalent to minimizing the difference between predictions and sensations, which turns out to be predictive coding (Friston and Kiebel, 2009; Friston, 2010). Using this implementation, researchers have successfully replicated brightness illusions including Cornsweet effect and Mach bands (Brown and Friston, 2012). Bayesian efficient coding extends classic efficient coding by posing a constraint on parameters of likelihood and introducing a general loss function characterizing the "goodness" of posterior (Park and Pillow, 2024). Well-constrained observer models that integrate efficient coding with Bayesian inference can explain the perceptual biases in orientation, motion, and color not previously accounted for by traditional Bayesian models, suggesting perceptual biases are determined crucially by the encoding, stimulus range, loss function and characteristics of noise (Wei and Stocker, 2012, 2015; Hahn and Wei, 2024).

1.2.2 Computational models for visual illusions

Inspired by early studies in psychology and neurophysiology, more specific theoretical models for individual illusion phenomena have been proposed. In this review, we will discuss some of the key computational models that have contributed to our understanding of visual illusions, providing a link between the general theories and the neural mechanisms involved.

Illusory contour and surface

Illusory contour and surface are perceptual phenomena where the visual system perceives edges and contours, or surfaces that are not physically present in the visual stimulus. Their mechanisms are believed to be closely related to those for surface perception (Mattingley et al., 1997; Shimojo et al., 2001; Marlow et al., 2017), figure-ground segmentation (Kovács and Julesz, 1993; Grossberg, 1994; Kimchi and Peterson, 2008; Liang et al., 2015; Tadin et al., 2019; L. Huang et al., 2020), and border ownership problems (Zhou et al., 2000; Zhaoping, 2005; Fang et al., 2009; Layton et al., 2012) in visual perception. Several computational models have been proposed to emphasize the role of feature extraction, grouping processes, neural interactions, and probabilistic computations in generating illusory contours (Grossberg and Mingolla, 1985; Manjunath and Chellappa, 1993; Heitger et al., 1994; Anderson and Julesz, 1995; Williams and Jacobs, 1997; Grossberg and Raizada, 2000; Lehar, 2003; Kalar et al., 2010).

Inspired by the properties of receptive fields in the visual cortex and their interactions (Hubel et al., 1977; Zeki, 1983a, 1983b), Grossberg and Mingolla proposed the Boundary Contour System (BCS), which is designed to detect and synthesize the boundaries of visual objects, both real and illusory (Grossberg and Mingolla, 1985). At the stage of boundary detection, BCS starts with the activation of oriented masks or elongated receptive fields, which are sensitive to specific orientations and contrasts but not to the direction of contrast (i.e., they respond similarly to light-dark and dark-light edges). These masks help in detecting the edges and contours of objects in the visual field. When building up boundary contours, there is first a short-range competition among cells representing like-oriented masks at nearby locations, followed by long-range cooperation among aligned masks which takes into account the global configuration of elements in the scene. These mechanisms ensure that the strongest signals, corresponding to the most prominent edges, are emphasized locally; and allow the system to 'complete' boundaries, creating a perceptual whole from fragmented visual inputs. Grossberg and Raizada advanced the foundational concepts of the BCS by integrating them into a more comprehensive and detailed cortical framework, adding the dimensions of attentional modulation and dynamic feedback, and by grounding these processes in a neurophysiological context that can be directly tested and simulated (Grossberg and Raizada,

2000). In particular, lateral connections and “bipole property” in layer 2/3 of V2 explain how perceptual grouping can occur over parts of the visual field that receive no direct visual stimuli but are inferred from the surrounding cues (Figure 1.5).

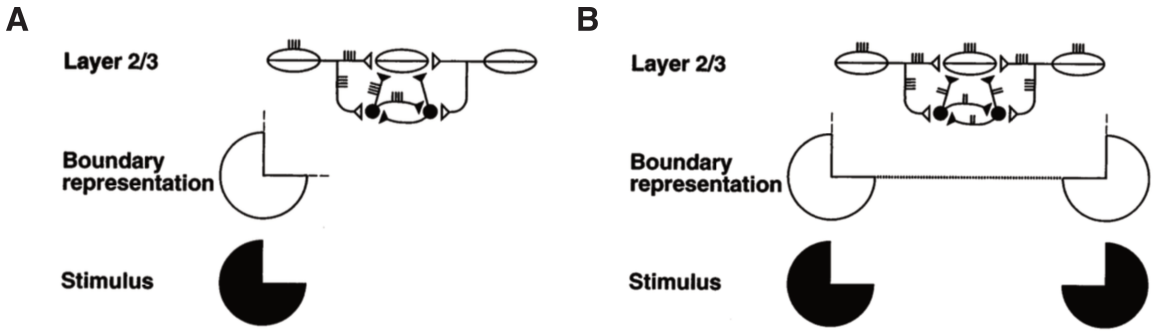

Fig. 1.5 The boundary grouping circuit in layer 2/3 can explain the phenomenon of Kanizsa illusion. Figure from Grossberg and Raizada (2000).

Consistent with neurophysiological observations of simple, complex, and hypercomplex cells in the visual cortex of mammals, models have emphasized multi-scale processing and nonlinear operations (Manjunath and Chellappa, 1993; Heitger et al., 1994). The Multistage System Model extracts and groups salient features in the image at different spatial scales or frequencies (Manjunath and Chellappa, 1993). It uses a Gabor wavelet decomposition to detect significant features such as step and line edges at various scales and orientations. The model incorporates local competitive interactions and interscale interactions to reduce noise, localize line ends and corners, and facilitate boundary completion. The Neural Contour Processing Model employs convolutions and nonlinear operations to simulate neural contour processing and figure-ground segregation (Heitger et al., 1994).

In contour and boundary detection, junctions, where different edges or contours meet or intersect, can introduce ambiguity in the visual scene (e.g., the "T-junctions" often indicate occlusion). Models also specified how to solve the ambiguity introduced by junctions and connect fragments (Heitger et al., 1994; Anderson and Julesz, 1995; Williams and Jacobs, 1997; Kalar et al., 2010). It defines contours by the local maxima of responses from a contour operator that sums contrast borders and a "grouping signal." The grouping mechanism involves convolving a representation of key points (T-junctions, corners, line ends) with orientation-selective kernels and a nonlinear pairing operation. The model successfully reproduces illusory contours and generates representations of occluding contours in natural scenes. The Stereoscopic Matching model focuses on the role of stereoscopic junctions in the formation of illusory contours in untextured stereograms (Anderson and Julesz, 1995). The model assumes that the visual system primarily uses horizontal disparities to match features between the eyes and form the basis for depth perception in illusory contours.

It extends a feed-forward model to detect local features indicative of occlusion (e.g., T-junctions) and interpolates contours through occluded regions (Kalar et al., 2010). The model incorporates contour and end-stop detection, generation of interpolation paths based on geometric constraints and relatability, and a grouping mechanism to link contour fragments.

The Stochastic Completion Fields Model represents the shape of illusory contours using random walks in a lattice space corresponding to positions and orientations in the image plane (Williams and Jacobs, 1997). It calculates the probability of a particle following a random walk through vector field convolutions, representing the likelihood of a path connecting boundary fragments in an image. It employs a parallel distributed representation to mimic the structure of the visual cortex, where each unit represents the probability of a particle passing through a specific position and orientation during its random walk.

These models highlight the importance of multi-scale processing, orientation selectivity, and the integration of local and global information in the perception of illusory contours and surfaces. They suggest that the perception of illusory contours and surfaces arises from the complex interplay of feedforward, feedback, and lateral interactions within the visual cortex. Future research can further refine these models and explore their implications for other aspects of visual perception and cognition.

Brightness and color illusion

Brightness and color illusions involve the misperception of these properties in regions of identical luminance or chromatic values. They are closely related to how our brain interprets the light received by the retina. Bright and color illusions can be explained by different computational models (Grossberg and Todorovic, 1988; Adelson, 1993; Zarandy et al., 1998; Gilchrist et al., 1999; Blakeslee and McCourt, 2004; Blakeslee et al., 2005; Grossberg and Hong, 2006; Robinson et al., 2007; Corney and Lotto, 2007; Otazu et al., 2008; Peters et al., 2010).

Grossberg and Todorovic proposed hierarchical neural network model with six levels representing different neural processing units, including Stimulus Distribution, LGN Cells (On/Off), Simple Cells, Complex Cells, Boundary Contour Units, and Diffusive Filling-In (Grossberg and Todorovic, 1988). Boundary Contour System (BCS) detects and synthesizes object boundaries, while Feature Contour System (FCS) fills in brightness within boundaries. The brightness illusions can be explained through the interaction of BCS and FCS, with the filling-in process creating perceived brightness differences. Later Grossberg and Hong refined the model that includes retinal adaptation, contrast enhancement, and filling-in mechanisms (Grossberg and Hong, 2006). They proposed a filling-in process in cortical areas like V1 and V2, regulated by boundary signals and introduced anchored lightness scaling (anchoring) to

maintain consistent lightness perception across varying conditions. The brightness illusions are explained through the interplay of contrast enhancement and efficient filling-in processes.

Models offer different views on how brightness illusions are generated due to contextual information. Adelson emphasizes the role of higher-level perceptual organization in brightness perception, beyond low-level processes like lateral inhibition (Adelson, 1993). He has demonstrated that changes in perceptual organization can significantly alter brightness illusions, even with the same local luminance contrasts. His view is consistent with the Bayesian framework that highlights the role of inferred properties like reflectance and transparency in brightness perception and the importance of the configurations of gray-level junctions in determining perceptual organization and brightness percepts. Zarándy and colleagues used Cellular Neural Networks to simulate processing and employed logarithmic transformation and subtraction to estimate overall illumination and separate it from reflectance (Zarandy et al., 1998). They explain brightness illusions through the transformation of visual inputs based on local context. The Anchoring Theory introduces the anchoring rule, which determines specific points on the gray scale as references for assigning lightness values. The model differentiates between local and global anchoring based on the luminance context of the visual field and addresses scaling and normalization of luminance ranges within frameworks (Gilchrist et al., 1999).

The models based on oriented multiscale spatial filtering using difference-of-Gaussians (DOG) filters and their variants have successfully predicted various brightness illusions, such as the White effect and simultaneous brightness contrast (Blakeslee and McCourt, 2004; Blakeslee et al., 2005; Robinson et al., 2007). The models incorporated contrast normalization across orientation channels, reflecting cortical processing properties and explained brightness contrast and assimilation as outcomes of the interaction between multiscale, oriented filters and the spatial layout of visual stimuli. FLODOG model has extended the ODOG model with localized energy normalization, providing more accurate predictions of brightness perception across different visual contexts (Robinson et al., 2007).

While many models are inspired by the architecture and processing in the visual cortex (Grossberg and Todorovic, 1988; Grossberg and Hong, 2006; Peters et al., 2010), the model proposed by Corney and Lotto employed artificial neural networks (ANNs) trained on 3D scenes composed of 'dead-leaves' models (Corney and Lotto, 2007). ANNs learn to predict surface reflectance based on luminance in the given environment. The systematic errors made by the trained ANNs are analogous to human brightness illusions, demonstrating that brightness illusions arise as a result of the visual system's attempt to infer physical properties from ambiguous visual input.

Some models emphasize higher-level perceptual organization and anchoring mechanisms (Adelson, 1993; Gilchrist et al., 1999). Other models, such as ODOG and FLODOG, focus on early-stage visual processing and low-level features (Blakeslee and McCourt, 2004; Blakeslee et al., 2005; Robinson et al., 2007). These models highlight the role of multiscale and oriented processing in the emergence of brightness illusions. ODOG, FLODOG, and the multiresolution wavelet framework, have been tested against empirical data and have successfully predicted various brightness illusions (Blakeslee and McCourt, 2004; Blakeslee et al., 2005; Robinson et al., 2007; Otazu et al., 2008). Another model directly compares its predictions with fMRI data, providing neural validation for the proposed mechanisms (Peters et al., 2010).

In summary, these computational models offer diverse perspectives and mechanisms for understanding brightness illusions, ranging from low-level spatial filtering and normalization to higher-level perceptual organization and anchoring. While each model has its strengths and focuses, they collectively contribute to a more comprehensive understanding of the complex processes underlying brightness perception and its illusions. Future research can further integrate these models and explore their complementary aspects to develop a unified framework for explaining brightness illusions.

1.2.3 Summary

In this section, we have explored general theories of visual perception, including efficient coding, the Bayesian brain hypothesis, and predictive coding, and their applications in explaining visual illusions. These theories offer complementary perspectives on how the brain optimizes its resources, makes inferences based on prior knowledge, and generates predictions to facilitate efficient processing of sensory information. They provide valuable insights into the mechanisms underlying visual illusions, such as the brain's efforts to optimize visual processing, the role of probabilistic inference in perception, and the constant comparison between predictions and sensory input.

However, while these general theories contribute to our understanding of visual illusions, they often fail to fully account for the intricacies and diversity of specific illusion types. The complex nature of visual illusions requires more targeted explanations that consider the unique characteristics of each phenomenon.

Inspired by early studies in psychology and neurophysiology, researchers have proposed more specific theoretical models for individual illusion phenomena, such as illusory contours and surfaces, brightness illusions, and color illusions. These models emphasize the role of feature extraction, grouping processes, neural interactions, and probabilistic computations in the generation of visual illusions. They provide detailed explanations for the mechanisms

underlying specific illusions, bridging the gap between the general theories and the observed perceptual experiences.

Notably, the same illusion phenomenon can often be explained by different mechanisms proposed by various models. This highlights the complexity of visual illusions and the need for further research to disambiguate and integrate the diverse explanations. The coexistence of multiple models for the same illusion also suggests that the perception of illusions may arise from the interplay of different levels of neural processing, rather than a single mechanism.

In conclusion, while general theories of visual perception provide valuable insights into the mechanisms of visual illusions, more specific theoretical models are necessary to account for the intricacies of individual illusion types. Future research should focus on empirical validation, the development of comprehensive computational models, and the exploration of the interplay between different levels of neural processing to advance our understanding of the complex nature of visual illusions.

1.3 Measurement techniques for subjective visual experience

A successful computational model should accurately predict phenomena both qualitatively and quantitatively. While it is straightforward to obtain binary answers from participants on whether they experience illusory perceptions from visual input, quantifying these illusory experiences poses a challenge due to their subjective nature.

To measure illusory experiences in our minds, it's essential to establish relationships between them and the physical properties that are measurable. Fechner proposed "outer" and "inner" psychophysics to link perceived mental content to visual input and neural representations in the physical world, respectively (Fechner, 1860) (Figure 1.6).

1.3.1 "Outer" psychophysics

"Outer" psychophysics, commonly referred to as psychophysics today, has been a primary focus of researchers, especially during periods when neurotechnology was less advanced than it currently is. Psychophysics is the scientific study of the relationship between physical stimuli and the sensations and perceptions they produce. It aims to quantify and measure subjective experiences, such as the perception of brightness, color, or the strength of visual illusions, by systematically varying the properties of the physical stimuli and observing the corresponding changes in perception (Gescheider, 1997; Wolfe et al., 2006).

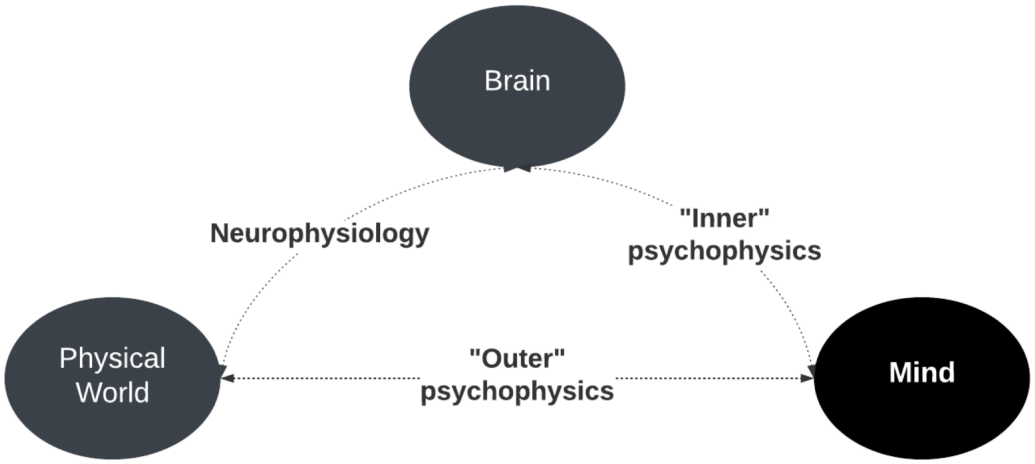

Fig. 1.6 “Outer” and “Inner” psychophysics, and Neurophysiology. While neurophysiology studies how the brain responds to the physical world, “outer” and “inner” psychophysics focus on the relationships between the mental content in mind with physical world and brain activity, respectively.

The origins of psychophysics can be traced back to the work of Gustav Fechner, who formalized the relationship between physical stimuli and sensory experience. Fechner’s law states that the perceived magnitude of a sensation is proportional to the logarithm of the stimulus intensity (Fechner, 1860). This law was based on the earlier work of Ernst Weber, who discovered that the just-noticeable difference (JND) between two stimuli is proportional to the magnitude of the stimuli, a principle known as Weber’s law (Weber, 1834). Stevens later proposed an alternative to Fechner’s law, known as Stevens’ power law, which states that the perceived magnitude of a sensation is related to the physical intensity of the stimulus by a power function (S. S. Stevens, 1957). Signal detection theory, developed by Green and Swets (1966), provided a framework for analyzing the decision-making process in perception tasks, taking into account the observer’s sensitivity and response bias.

Introduction to psychophysical methods

Psychophysical methods can be broadly classified into three categories: detection, discrimination, and scaling methods (Gescheider, 1997). Detection methods, such as the method of constant stimuli and the method of limits, aim to determine the absolute threshold of a stimulus, which is the minimum intensity required for an observer to detect the presence of a stimulus. Discrimination methods, like the two-alternative forced-choice (2AFC) paradigm, measure the observer’s ability to distinguish between two stimuli that differ in a specific property, such as brightness or orientation. Scaling methods, including magnitude estima-

tion and cross-modality matching, allow observers to assign numerical values or match the perceived intensity of a stimulus to another modality, such as sound or touch.

Detection methods

Detection methods aim to determine the absolute threshold of a stimulus, which is the minimum intensity required for an observer to detect the presence of a stimulus.

The method of constant stimuli involves presenting a set of stimuli with fixed intensities in a random order, and the observer reports whether they detect the stimulus or not (Gescheider, 1997). This method allows for the estimation of the absolute threshold and the construction of a psychometric function, which plots the percentage of correct detections against stimulus intensity (Kingdom and Prins, 2016).

In the method of limits, the intensity of a stimulus is gradually increased or decreased until the observer detects or loses the perception of the stimulus (Wolfe et al., 2006). This method provides an estimate of the absolute threshold, but it is susceptible to observer bias and expectation effects.

The method of adjustment allows the observer to adjust the intensity of a stimulus until it is just detectable or just undetectable (Gescheider, 1997). This method is quick and easy to implement but may be influenced by the observer's response bias (Kingdom and Prins, 2016). To efficiently estimate the threshold, adaptive staircase procedures, such as QUEST (Watson and Pelli, 1983) and PEST (Taylor and Creelman, 1967), adjust the stimulus intensity based on the observer's previous responses. They are more time-efficient than the method of constant stimuli and less prone to bias than the method of limits.

Discrimination methods

This group of methods measure the observer's ability to distinguish between two stimuli that differ in a specific property, such as brightness or orientation.

Two-alternative forced-choice (2AFC) method presents two stimuli to the observer, who must choose which stimulus has a higher intensity or a specific attribute (Fechner, 1889; Posner, 1980; Britten et al., 1993). This method is less prone to bias than the method of limits and has been used extensively in studying visual illusions. Three-alternative forced-choice (3AFC) and four-alternative forced-choice (4AFC) methods are similar to 2AFC but present three or four stimuli, respectively, with the observer choosing the stimulus that differs from the others (Schlauch and Rose, 1990; Jäkel and Wichmann, 2006; Victor and Conte, 2012). They are less susceptible to bias than 2AFC but may be more time-consuming (Schlauch and Rose, 1990; Jäkel and Wichmann, 2006).

The second type of task is the same-different task. In this task, observers are presented with two stimuli and must decide whether they are the same or different (Miller and Bauer, 1981). This method has been used to study various aspects of visual perception, such as perceptual grouping (Prinzmetal, 1981) and parallel visual processing (Donderi and Case, 1970).

The third type of task is the oddity task. Observers are presented with multiple stimuli, one of which differs from the others, and must identify the odd stimulus (Iversen and Humphrey, 1971; A. C. Lee et al., 2012). This method has been employed to investigate configural feature ambiguity (Bartko et al., 2007) and visual search (Song and Nakayama, 2008).

Scaling methods

Scaling methods allow observers to assign numerical values or match the perceived intensity of a stimulus to another modality, such as sound or touch (Wolfe et al., 2006).

Magnitude estimation involves assigning numerical values to the perceived intensity of a stimulus relative to a reference stimulus (S. S. Stevens, 1975). Reversely, observers can adjust the intensity of a stimulus to match a given numerical value, which is called Magnitude Production (S. S. Stevens, 1975). This method has been used to investigate the relationship between physical intensity and perceived magnitude in various sensory modalities (S. S. Stevens and Guirao, 1963; J. C. Stevens and Marks, 1965).

The scales can also be accessed through paired comparison: observers compare two stimuli and indicate which one has a higher perceived intensity or quality (Thurstone, 1927). This method has been used to investigate the perception of brightness (Torgerson, 1958; Worthington, 1969), and taste (Amerine et al., 1965; LARSON-POWERS and Pangborn, 1978).

For those experiences with categorical namings, the category scaling method asks observers to assign stimuli to predefined categories based on their perceived intensity or quality (Dunn-Rankin, 2012). This method has been employed to study the perception of color (Boynton and Gordon, 1965; Lindsey and Brown, 2014), pain (Gracely et al., 1978), and emotion (Russell, 1980).

Finally, cross-modality matching requires observers to match the perceived intensity of a stimulus in one sensory modality to the intensity of a stimulus in another modality (S. S. Stevens, 1975). For instance, observers may adjust the loudness of a sound to match the perceived brightness of a visual stimulus.

Application to measuring subjective experiences

Psychophysical methods have been extensively applied to measure various aspects of visual perception, including visual illusions. The brightness illusion has been studied using the method of constant stimuli (Spillmann et al., 1984), magnitude estimation (Dresp et al., 1990), and the 2AFC method (Bindman and Chubb, 2004). These studies have consistently shown that the target appears different from its background of the same luminance, or the perceived brightness of the target can be enhanced due to surface perception.

The Müller-Lyer illusion, where the apparent length of a line is influenced by the orientation of the arrowheads at its ends, has been investigated using the method of limits (Köhler and Fishback, 1950; Pollack and Jaeger, 1991) and magnitude estimation (Dormal et al., 2018). Results from these studies have demonstrated that inward-pointing arrowheads lead to an underestimation of the line length, while outward-pointing arrowheads result in an overestimation, further revealing how the illusion size is affected by factors such as repetition or lightness.

The Ebbinghaus illusion, in which the perceived size of a target circle is affected by the size of its surrounding circles, has been studied using the 2AFC method (Doherty et al., 2010) and the method of adjustment (Manning et al., 2017). These studies have shown that there may be discrepancy in perceiving Ebbinghaus illusion due to the age or psychiatric disorders.

Other visual illusions, such as the Ponzo illusion (D. He et al., 2015), the Zöllner illusion (Morgan and Casco, 1990; Kitaoka and Ishihara, 2000), and the Poggendorff illusion (Weintraub and Krantz, 1971; Weintraub et al., 1980; Westheimer, 2008), have also been investigated using various psychophysical methods. These studies have provided valuable insights into the factors influencing the strength and direction of the illusory effects, such as the spatial configuration of the stimuli, the observer's attentional state, and individual differences in susceptibility to illusions.

While psychophysical methods have been pivotal in advancing our understanding of visual illusions, they primarily rely on behavioral responses and are subject to several limitations. Observer bias, expectation effects, and decision-making strategies can influence the results obtained from these methods (Kingdom and Prins, 2016). Furthermore, they do not provide direct access to the neural representations underlying the subjective experiences of visual illusions, which has motivated the development of "inner" psychophysical methods.

1.3.2 "Inner" psychophysics

"Inner" psychophysical methods measure perceived mental content based on brain activity, offering a more direct approach to studying the neural correlates of subjective experiences (Haynes, 2009). These methods have gained popularity in recent years due to advances in

neurotechnology, such as single-unit recordings, functional magnetic resonance imaging (fMRI), and electroencephalography (EEG) (Haynes, 2009; Mukamel and Fried, 2012; Contini et al., 2017).

Single-unit recordings have been used to demonstrate that the activity of individual neurons can reliably predict an animal's perceptual judgments. In a seminal study, Newsome et al. (1989) showed that the firing rates of neurons in the middle temporal (MT) area of the macaque monkey brain were strongly correlated with the animal's decisions about the direction of motion in a random-dot stimulus. This study introduced the concept of the neurometric function, which relates neural activity to behavioral performance, analogous to the psychometric function in "outer" psychophysics. Similar findings have been reported for other visual attributes, such as orientation (Vogels and Orban, 1990), depth (Uka and DeAngelis, 2003), and color (Kusunoki et al., 2006). However, it is still a hard problem to locate the minimal set of neuronal events leading to the holistic perceptual experiences (Crick and Koch, 1995; Cohen and Dennett, 2011).

The concept of "shared representation" between veridical and subjective content is central to "inner" psychophysics. This idea suggests that the neural representations underlying the perception of a physical stimulus are similar to those underlying the corresponding subjective experience, such as a visual illusion. Single-unit recordings have provided evidence for shared representations in the context of visual illusions. For example, neurons in the primary visual cortex (V1) of monkeys have been shown to respond to illusory contours in a manner similar to their responses to real contours (von der Heydt et al., 1984; Peterhans and von der Heydt, 1989). Similarly, neurons in the inferior temporal cortex (IT) have been found to respond to illusory shapes in the Kanizsa triangle illusion (Sáry et al., 2008).

fMRI has been widely used to investigate the neural correlates of visual illusions in humans. Studies have shown that the primary visual cortex (V1) and higher-order visual areas, such as the lateral occipital complex (LOC), are involved in the processing of various illusions, including the Müller-Lyer illusion (Weidner and Fink, 2007), the Ponzo illusion (S. O. Murray et al., 2006), the Kanizsa triangle illusion (Seghier and Vuilleumier, 2006; Maertens et al., 2008), and the perceived surface lightness in the Cornsweet illusion (Boyaci et al., 2007). These studies have revealed that the neural activity in these areas reflects the perceived illusory content rather than the physical stimulus properties.

EEG and magnetoencephalography (MEG) have been employed to study the temporal dynamics of illusory perception. Studies have shown that the perception of illusory contours and motion involves early visual processing stages, as reflected by modulations in the P1 and N1 components of the event-related potential (ERP) (M. M. Murray et al., 2002). MEG studies have further revealed that the perception of illusory contours is associated

with increased gamma-band oscillations in the visual cortex (Tallon-Baudry et al., 1996; Herrmann and Bosch, 2001).

Multi-voxel pattern analysis (MVPA) has emerged as a powerful tool for decoding subjective experiences from fMRI data (Haxby et al., 2001; Kamitani and Tong, 2005; Norman et al., 2006). MVPA has been used to decode the visual content during nonREM sleep (Horikawa et al., 2013) and perceived color for form-contingent color filling-in (Hong and Tong, 2017). These studies demonstrate that subjective experiences can be reliably predicted from patterns of neural activity, providing strong evidence for the shared representation hypothesis.

Passive viewing experiments, where observers simply fixate on the stimuli without making any overt responses, have been combined with "inner" psychophysical methods to eliminate the potential confounds of decision-making and motor processes (Haynes, 2009). For example, Kok and colleagues used fMRI to measure brain activity while participants passively viewed illusory contours in the Kanizsa triangle illusion. They found that the neural activity in the deep layers of V1 reflected the top-down feedback that influenced perceived illusory contours, even in the absence of a perceptual decision or motor response (Kok and de Lange, 2014; Kok et al., 2016).

In conclusion, "inner" psychophysical methods have greatly expanded our understanding of the neural basis of visual illusions and subjective experiences. By measuring brain activity during the presentation of both veridical and illusory stimuli, these methods have provided compelling evidence for shared neural representations underlying perceptual experiences. The combination of various techniques, such as single-unit recordings, fMRI, EEG, MEG, and MVPA, has enabled researchers to investigate the spatial and temporal dynamics of illusory perception at different levels of the visual system. Passive viewing experiments have further strengthened the link between neural activity and subjective experience by minimizing the influence of decision-making and motor processes. As neurotechnology continues to advance, "inner" psychophysical methods are expected to play an increasingly important role in unraveling the neural mechanisms of visual illusions and other aspects of conscious experience.

1.3.3 Summary

In summary, "outer" and "inner" psychophysics offer complementary approaches to measuring and quantifying visual perceptual experiences, including visual illusions. While "outer" psychophysical methods have been widely used to study the properties and mechanisms of illusions, they rely on behavioral responses and can be influenced by various biases and confounds. "Inner" psychophysical methods, on the other hand, provide a more direct

window into the neural representations underlying subjective experiences by measuring brain activity during the perception of both veridical and illusory stimuli.

Single-neuron recordings, fMRI, MEG, and EEG have been employed to identify the neural substrates of various visual illusions, revealing the involvement of both early visual areas, such as V1, and higher-order visual regions, such as LOC, in the processing of illusory percepts. The concept of "shared representation" suggests that the neural correlates of veridical and illusory perceptions are similar.

Passive viewing experiments in combination with "inner" psychophysical methods offer a promising approach to studying the neural basis of subjective experiences, as they eliminate the potential confounds of decision-making and motor processes. Future research should continue to explore the neural mechanisms underlying visual illusions using a combination of "outer" and "inner" psychophysical methods, with a focus on integrating findings across different levels of analysis, from single neurons to large-scale brain networks.

By bridging the gap between the subjective experience of illusions and their neural correlates, "inner" psychophysics can provide valuable insights into the nature of perception and the mechanisms by which the brain constructs our conscious experiences. This knowledge can inform the development of more comprehensive computational models of visual perception and contribute to our understanding of how the brain processes and interprets the complex and ambiguous sensory information that gives rise to visual illusions.

1.4 Proposed approach

To investigate the neural representations underlying visual illusions, we propose a multi-faceted approach that combines psychophysical experiments, computational modeling, and neuroimaging techniques. This integrated methodology allows us to study the perceptual characteristics of illusions, develop and test hypotheses about their underlying mechanisms, and directly probe the neural correlates of illusory experiences.

Our approach begins with the design of visual stimuli that implemented existing psychophysical literature of the illusions of interest, namely illusory contours and neon color spreading. We will employ "outer" psychophysical methods, such as paired stimuli, to decide physical parameters that induce strongest illusory experiences for human participants.

To directly probe the neural representations of these illusions, we will employ advanced neuroimaging techniques, particularly functional magnetic resonance imaging (fMRI) and reconstruction methods. Participants will undergo fMRI scans while viewing both illusory and non-illusory stimuli, enabling us to measure brain activity patterns associated with the perception of these illusions. We will apply state-of-the-art reconstruction techniques,

such as those based on deep neural networks and image generators, to decode and visualize the illusory percepts from the fMRI data. By reconstructing illusory percepts, we can assess the extent to which the neural representations match the subjective experience of the illusions. This approach allows us to bridge the gap between the perceptual phenomena, the computational theories, and the underlying neural mechanisms.

Furthermore, we will investigate the role of different visual areas in the processing of these illusions by reconstructing the illusory percepts from fMRI data at various stages of the visual hierarchy, from early visual areas (e.g., V1, V2) to higher-order areas (e.g., V4, LOC). This will provide empirical evidence for computational models and help elucidate the neural circuits involved in the perception of illusory contours and neon color spreading.

Our proposed approach, which integrates psychophysics, computational modeling, and neuroimaging, offers a powerful framework for unraveling the neural representations of visual illusions. By combining the strengths of these complementary methods, we aim to provide a deeper understanding of how the brain constructs subjective perceptual experiences from limited sensory inputs and advance our knowledge of the neural basis of conscious perception.

1.5 Thesis organization

This thesis is organized into nine chapters, each focusing on a specific aspect of our investigation into the neural representations of visual illusions. The current chapter, Chapter 1, provides an introduction to visual illusions, their categorization, and the challenges in measuring and understanding these phenomena. It also outlines our proposed approach, which combines psychophysical experiments, computational modeling, and neuroimaging techniques to study the neural mechanisms underlying illusory contours and neon color spreading.

Chapter 2 delves into the fluid taxonomy of visual illusions and reviews the existing literature on "outer" and "inner" psychophysical studies of illusory contours and neon color spreading. This chapter highlights the key findings, debates, and open questions in the field, setting the stage for our own investigations.

Chapter 3 describes our methodology setup, including an overview of the study design, the introduction of brain decoding and reconstruction methods, and the rationale behind our choices of methods. This chapter also discusses the advantages and limitations of our approach and its potential for advancing our understanding of the neural basis of visual illusions.

Chapter 4 presents the experimental design of our psychophysical and neuroimaging studies. It details the visual stimuli used, the fMRI experimental procedures, and the data preprocessing steps. This chapter also describes the regions of interest (ROIs) selected for the reconstruction analyses based on their known involvement in visual processing.

Chapters 5 and 6 focus on the reconstruction of illusory percepts from fMRI data. Chapter 5 presents the results of our decoding and reconstruction analyses, demonstrating the feasibility of reconstructing illusory contours and neon color spreading from brain activity patterns. Chapter 6 examines the role of different components in the reconstruction process, including the choice of deep neural network (DNN) architecture and the image generator module. This chapter evaluates the robustness of our findings to variations in these components and provides insights into the factors that influence the quality of the reconstructed illusory percepts.

Chapters 7 and 8 investigate the reconstruction of illusory contours and neon color spreading along the visual hierarchy, respectively. These chapters present the results of reconstructing the illusory percepts from fMRI data at different stages of visual processing, from early visual areas to higher-order regions. By comparing the reconstructions across the visual hierarchy, these chapters shed light on the contributions of different visual areas to the representation of illusory experiences.

Finally, Chapter 9 provides an integrative discussion of our findings, highlighting the strengths and limitations of our approach, and proposing future directions for research. By synthesizing our findings and situating them within the broader context of psychophysics, computational neuroscience and the study of consciousness, we aim to contribute to the ongoing endeavor of unraveling the mysteries of the mind and brain.

Chapter 2

Visual illusions

In the previous chapter, we introduced the concept of visual illusions, their categorization based on the primary visual attributes that deviate from the physical stimulus, and the challenges in measuring and understanding these phenomena. We also outlined our proposed approach, which combines psychophysical experiments, computational modeling, and neuroimaging techniques to study the neural mechanisms underlying illusory contours and neon color spreading.

In this chapter, we delve deeper into the complex nature of visual illusions, exploring the fluid taxonomy that makes their definition and categorization a challenging task. We will discuss how the interpretation of illusions depends on the definition of reality and the level of deviation from the physical stimulus. We will then focus on two specific types of illusions, illusory contours and neon color spreading, which are of particular interest in this thesis. We will review the existing literature on the psychophysical studies and neural correlates of these illusions, highlighting the key findings, debates, and open questions in the field. Finally, we will discuss the prospects of using reconstruction techniques to expand the application of "inner" psychophysical methods in addressing the challenges posed by visual illusions.

2.1 The fluid taxonomy of visual illusions

The definitions of illusions have been ambiguous, and the classification of visual illusions remains a complex and debated topic in the field of vision science. The ambiguity in defining illusions stems from two main factors: the definition of reality and the level of deviation from the physical stimulus.

First, depending on how we define the reality of the physical world, the same visual phenomenon can be regarded as illusory or non-illusory (Figure 2.1). A prime example of this ambiguity is the dress illusion, where people perceive different colors of a dress under

different lighting conditions. If we consider the ground truth to be the pixel values or the spectral light that reaches the retina, the perceived colors of the dress can be regarded as an illusion. This is because the brain's interpretation of the colors deviates from the objective measurements of the physical stimulus. However, if we define the ground truth as the color of the real dress in the world, then our perception of the dress's color can be considered a correct inference based on the available sensory information. In this case, the perceived colors would not be classified as an illusion.

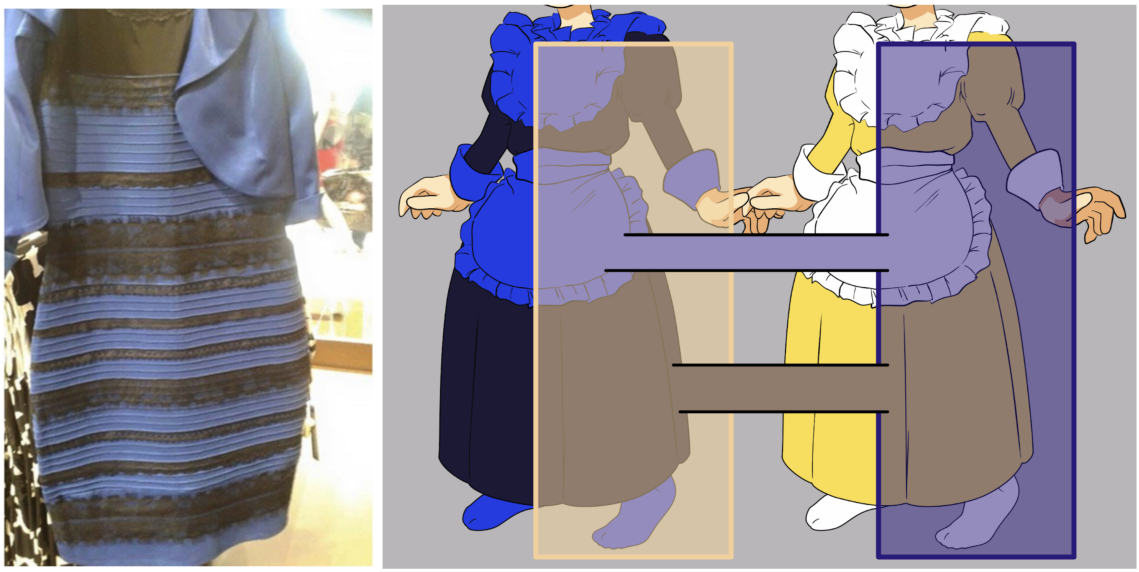

Fig. 2.1 Example of the fluid taxonomy of visual illusions: The dress. If we consider the ground truth to be the pixel values or the spectral light that reaches the retina, the perceived colors of the dress can be regarded as an illusion. However, if we define the ground truth as the color of the real dress in the world, then our perception of the dress's color can be considered a correct inference and would not be classified as an illusion.

Second, the level of deviation from the physical stimulus also plays a role in determining whether a visual phenomenon is considered an illusion. All visual perceptions are not entirely accurate representations of the physical world, as they are the result of the brain's interpretation of sensory information. However, illusions are typically characterized by a clear and significant deviation from the physical stimulus. For example, in the Müller-Lyer illusion, the perceived lengths of two lines are strikingly different, even though the lines are physically identical in length. The magnitude of the perceptual deviation in this case is much larger than the minor inaccuracies that are inherent in all visual perceptions.

The classification of visual illusions has been a subject of ongoing research and debate (Gregory, 1997; Carbon, 2014; Shapiro and Todorović, 2017). Richard L. Gregory proposed that illusions can be categorized into three categories: physical illusions, physiological

illusions, and cognitive illusions (Gregory, 1997). Physical illusions, such as a pencil appearing bent when partially submerged in water, are caused by the physical properties of light and the environment. These phenomena are not considered true visual illusions, as they can be explained by the laws of optics. Physiological illusions, on the other hand, are driven by the inherent wiring and functioning of the visual system. The Café Wall illusion, where parallel horizontal lines appear sloped, is an example of a physiological illusion. This illusion is believed to be caused by the way our visual system processes edges and contrasts. Cognitive illusions, perhaps the most widely known category, are influenced by higher-level cognitive factors such as expectation, context, and prior knowledge. The Hollow-Mask illusion, where a concave mask of a face appears as a normal convex face, is an example of a cognitive illusion (Króliczak et al., 2006). This illusion arises from our brain’s strong bias for perceiving faces as convex, which overrides other depth cues indicating that the object is hollow. Another example of a cognitive illusion is the filling-in illusion, where humans perceive visual features that do not physically exist in the corresponding visual field but are present in the surrounding areas (Komatsu, 2006). Instead of regarding illusions as mistakes, many scientists view them as opportunities to understand the underlying computations implemented by the neural system. For instance, the phenomenon of filling-in can be considered a form of surface interpolation or modal completion, reflecting the brain’s ability to construct coherent perceptual experiences from incomplete sensory information.

In this thesis, we focus on two specific types of visual illusions: illusory contours and neon color spreading (Schumann, 1900; Soriano et al., 1996; van Tuijl, 1975). These illusions are particularly intriguing because they involve the perception of lines or colors that are not physically present in the stimulus or the signals that reach the retina. Despite the fluid taxonomy of visual illusions, illusory contours and neon color spreading are widely recognized as illusions according to all existing definitions, making them compelling targets for investigating the neural mechanisms underlying subjective perceptual experiences.

2.2 Related “outer” and “inner” psychophysical studies

In this section, we review the existing literature on the psychophysical studies and neural correlates of these illusions, encompassing both "outer" and "inner" psychophysical approaches. Then, we discuss how the evidence from “outer” and “inner” psychophysical studies provides support for several computational models discussed in Section 1.2.

2.2.1 Illusory contour

Illusory contours, also known as subjective contours, are perceived edges or boundaries that exist in the absence of physical gradients in luminance or color. These illusions involve the brain's interpretation of edges and shapes based on the arrangement of inducing elements in the visual scene. Notable examples of illusory contours include the Kanizsa triangle (Kanizsa, 1955), the Ehrenstein illusion (Ehrenstein, 1941), and the abutting line grating illusion (Soriano et al., 1996).

"Outer" psychophysical studies

Psychophysical studies have extensively investigated the perceptual properties and factors influencing the strength of illusory contours. For example, the perceived strength and clarity of illusory contours have been shown to depend on the spatial arrangement, alignment, and contrast of the inducing elements (Banton and Levi, 1992; Shipley and Kellman, 1992; Soriano et al., 1996; Grossberg and Mingolla, 1985). Soriano et al. (1996) systematically manipulated various parameters of abutting line gratings to investigate their effects on the perceived strength of illusory contours. They found that the perception of illusory contours is most robust under specific conditions, such as optimal line density, and certain phase angles. These findings suggest a finely tuned sensitivity in the visual system to specific geometric configurations that facilitate illusory contour perception, likely mediated by the activation and interaction of end-stopped receptive fields in the visual cortex.

The role of perceptual grouping and figure-ground segregation in the formation of illusory contours has also been demonstrated through psychophysical experiments (Kogo et al., 2010). Recent studies have further explored the role of spatial and temporal factors in the perception of illusory contours. For example, Masuda et al. (2015) showed that the temporal dynamics of illusory contour formation are influenced by the synchrony of the inducing elements, with synchronous presentation leading to stronger illusory contours compared to asynchronous presentation.

"Inner" psychophysical studies

Neurophysiological studies, including single-unit recordings in animals and neuroimaging experiments in humans, have provided valuable insights into the neural correlates of illusory contour perception.

Single-unit recordings in monkeys have revealed that neurons in early visual areas, such as V1 and V2, respond to illusory contours in a manner similar to their responses to real contours (von der Heydt et al., 1984; Peterhans and von der Heydt, 1989). By comparing the responses of neurons in different visual areas (V1, V2, and V4), a later study revealed the

hierarchical nature of illusory contour processing, with higher-order areas like V4 playing a crucial role in integrating local information into global percepts (Pan et al., 2012a). This finding challenges the traditional view that early visual areas are primarily responsible for contour perception and suggests a more distributed network involving both early and higher-order visual regions.

Several studies utilized fMRI to investigate the neural mechanisms of illusory contour perception, providing high spatial resolution insights into the brain regions involved in processing these stimuli (Larsson and Amunts, 1999; Mendola, Dale, Fischl, Liu, and Tootell, 1999; D. A. Stanley and Rubin, 2003; Seghier and Vuilleumier, 2006; Maertens et al., 2008). These studies have demonstrated that the blood-oxygen-level-dependent (BOLD) response in early visual areas and lateral occipital complex (LOC) is modulated by the presence and strength of illusory contours, indicating their involvement in the processing of illusory contours. Kok and colleagues used high-resolution fMRI to investigate the laminar profile of illusory contour responses in the human visual cortex. They show that shape perception modulates neural activity in V1 based on predictions and the presence or absence of corresponding sensory input supporting generative models of perception (Kok and de Lange, 2014). Additionally, their follow-up work suggests a specialized role for deep layers in integrating and processing top-down predictions, which are essential for constructing the perceptual experience of illusory contours (Kok et al., 2016) (Figure 2.2). de Haas and Schwarzkopf (2018) found that illusory contours, occluded objects, and subtle luminance contrasts all elicited similar retinotopic responses in the early visual cortex, with no significant differences in signal-to-noise ratios or population receptive field sizes between the conditions. This suggests that responses to these different stimuli are driven by spatial attention rather than unique processing pathways for illusions.

The high spatial resolution provided by fMRI allows for a detailed examination of the brain regions involved in illusory contour perception, revealing the roles of both early visual areas (e.g., V1) and higher-order regions (e.g., LOC) in processing these stimuli. These studies highlight the importance of top-down feedback and predictive processes in shaping neural responses to illusory contours, challenging the traditional view of a purely feedforward processing hierarchy. Additionally, the findings suggest that illusory contours may not be processed through entirely distinct neural circuits compared to real contours, with spatial attention playing a significant role in driving responses to both types of stimuli.

Studies using high temporal resolution techniques such as EEG and MEG have provided insights into the temporal dynamics of illusory contour processing. Halgren et al. (2003) used MEG to measure cortical activation in subjects viewing Kanizsa-type figures. They found that the perception of illusory shapes evoked a complex sequence of cortical activations, beginning

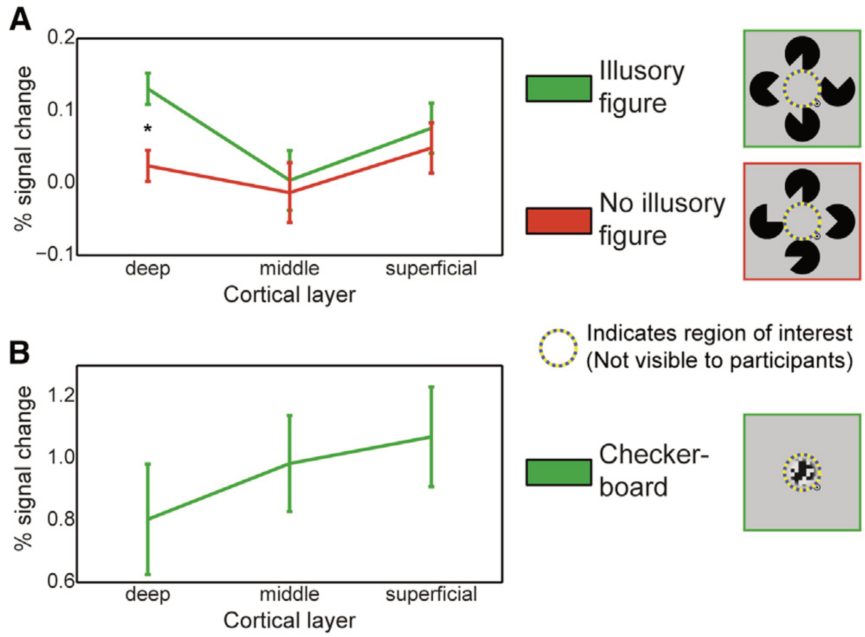

Fig. 2.2 Activations of deep laminar layers to illusory contour. Figure adapted from Kok et al. (2016).

with early activity at the occipital pole, followed by a prominent peak in the lateral occipital cortex, and eventually spreading to ventral occipital, temporal, and orbitofrontal cortices. This study revealed the involvement of both early and mid-level visual areas in illusory contour perception, with significant differential activity in the right hemisphere. Knebel and Murray (2012a) utilized electrical neuroimaging analyses of visual evoked potentials (VEPs) to determine the timing and spatial localization of neural processes involved in illusory contour perception. They found that illusory contour sensitivity primarily occurs first within the lateral occipital cortices (LOC) at about 85 milliseconds post-stimulus onset, followed by subsequent feedback effects in the V1/V2 areas. This finding was independent of the orientation of the inducing grating, indicating a generalized mechanism for illusory contour perception in the LOC before involving early visual cortices.

The high temporal resolution of EEG and MEG allows for a precise characterization of the timing of neural responses to illusory contours. These studies reveal that illusory contour processing involves a dynamic interplay between early visual areas and higher-order regions, with initial sensitivity occurring in the lateral occipital cortex followed by feedback to early visual areas. This temporal sequence supports the idea of top-down influences playing a crucial role in the perception of illusory contours, with higher-order areas guiding the processing in early visual regions.

In conclusion, inner psychophysical studies using single-neuron recordings, fMRI, EEG, and MEG have significantly advanced our understanding of the neural mechanisms underlying illusory contour perception. These studies reveal that illusory contour processing involves a hierarchical and distributed network of visual areas, with higher-order regions like V4 and the lateral occipital complex playing crucial roles in integrating local information into global percepts. The findings also highlight the importance of top-down feedback and predictive processes in shaping neural responses to illusory contours, challenging the traditional view of a purely feedforward processing hierarchy. Furthermore, the temporal dynamics of illusory contour processing, as revealed by EEG and MEG studies, demonstrate the rapid and dynamic interplay between early visual areas and higher-order regions. Overall, these studies contribute to a more comprehensive understanding of how the brain constructs coherent perceptual experiences from incomplete or ambiguous visual information, shedding light on the complex mechanisms underlying visual perception.

2.2.2 Neon color spreading

Neon color spreading is a visual illusion characterized by the perception of a colored glow or spread that extends beyond the boundaries of physically colored elements in an image. This illusion, first documented by van Tuijl (1975), is believed to arise from the brain’s interpretation of surface properties and the interaction between color and luminance information.

“Outer” psychophysical studies

“Outer” psychophysical studies have provided valuable insights into the mechanisms underlying neon color spreading. The perception of this illusion is influenced by a complex interplay of color, transparency, contour, and surface information.

Psychophysical studies have investigated the factors that influence the strength and extent of neon color spreading. The arrangement of inducing elements, such as the continuity and alignment of lines or edges, has been shown to play a crucial role in determining the vividness of the illusory color spread (Redies et al., 1984; Bressan et al., 1997). The contrast between the inducing elements and the background, as well as the chromatic properties of the inducers, also contribute to the perception of neon color spreading. Redies and Spillmann (1981) found that the neon color effect is strongly influenced by the similarity between the color of the background and the color of the inducing cross in Ehrenstein figures. Their experiments varied the color and orientation of crosses, demonstrating that the strength of neon color spreading is affected by these color relationships. Metelli (1985) also explored color relationships by creating stimuli mimicking transparency. The study found that the

same local stimulation can result in the perception of either color fusion or transparency, depending on the background conditions, highlighting the importance of color context in perceptual outcomes.

Several studies show the shared underlying perceptual processes between transparency and neon color spreading. Nakayama et al. (1990) demonstrated that transparency actively influences the perception of color, depth, and contour, and enhancing the perception of transparency intensifies neon color spreading (Figure 2.3). Bressan (1993) further explored the relationship between transparency and neon color spreading, finding that neon color spreading can occur even in the absence of traditional figural prerequisites if the visual configuration supports the interpretation of a transparent overlay. This study challenges the notion that neon spreading relies on specific figural conditions and suggests a more integrated understanding of color perception mechanisms.

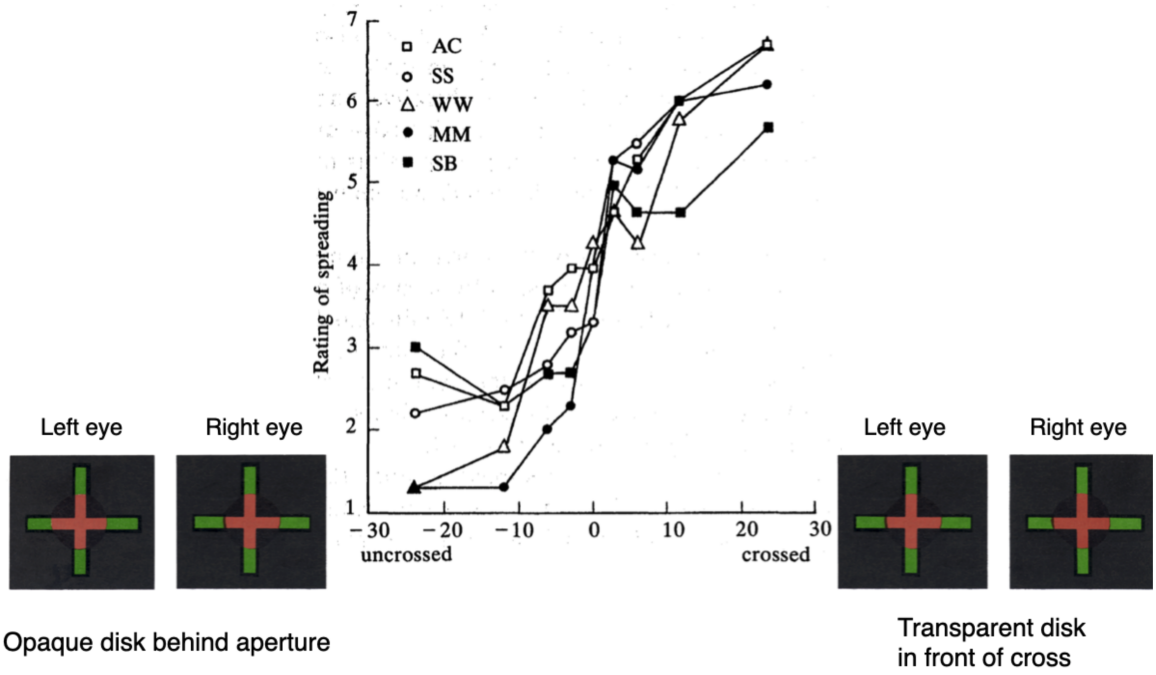

Fig. 2.3 The perceived degree of color spreading is affected by depth relationships. Figure adapted from Nakayama et al. (1990).

Neon color spreading has been associated with the mechanisms of figure-ground segregation and global percept of integrated local parts. De Weerd et al. (1998) explored perceptual filling-in, a process related to neon color spreading, and found that the time required for filling-in varies depending on the size, shape, and eccentricity of the figure. Their findings suggest that the delay before a figure is perceptually filled in reflects the time it takes for figure-ground segregation mechanisms to fail, providing insights into the mechanisms of

neon color spreading. Kamitani and Shimojo (2003) discuss how neon color spreading can be understood as part of the visual system’s broader capability to integrate incomplete local signals into a globally coherent visual experience, occurring early in the visual processing stages. Chen et al. (2018) found that contour interpolation and surface filling-in operate independently of each other. Although their study focused on Kanizsa-type illusory figures, the mechanisms of surface filling-in and contour interpolation could offer insights into related phenomena like neon color spreading.

“Inner” psychophysical studies

While outer psychophysical studies have provided valuable insights into the perceptual aspects of neon color spreading, inner psychophysical studies have sought to understand the neural mechanisms underlying this illusion. However, compared to the extensive research on the neural basis of illusory contours, the inner psychophysical study of neon color spreading has been less explored. This is largely due to the complexity of the phenomenon, which involves the intricate interplay of color, brightness, and surface perception, making it challenging to dissociate these factors in neuroimaging studies.

Despite these challenges, several studies have made significant contributions to our understanding of the neural mechanisms underlying neon color spreading. Sasaki and Watanabe (2004) employed functional MRI (fMRI) to investigate the involvement of the primary visual cortex (V1) in the process of color filling-in, a phenomenon closely related to neon color spreading. Their study found that V1 is actively involved in the color filling-in process, independently of attention modulation. This finding suggests that V1 plays a direct and autonomous role in surface representation and color filling-in, which may extend to the perceptual extension of color in neon color spreading. The study provides further support for the involvement of early visual areas in the processing of this illusion, consistent with models that emphasize low-level mechanisms, such as the Boundary Contour System (BCS) and Feature Contour System (FCS) (Grossberg and Mingolla, 1985), FACADE (Form-And-Color-And-DEpth) (Pinna and Grossberg, 2005), and 3D LAMINART Model (Grossberg and Yazdanbakhsh, 2005). These models propose that neon color spreading arises from the interaction between boundary completion and surface filling-in processes in the early visual cortex.

However, the role of V1 in color filling-in and neon color spreading remains a subject of debate. Cornelissen et al. (2006) also utilized fMRI to investigate neural responses to brightness and color filling-in. They found no evidence of local neural activity changes in V1 or nearby retinotopic maps corresponding to changes in perceived brightness or color induced by modulating a surrounding field. This suggests that perceived brightness and

color changes, which are crucial aspects of neon color spreading, may not be topographically represented in early visual areas. Introducing a novel mouse paradigm for the neon color spreading illusion, Saeedi et al. (2022) provides the evidence from single-unit recordings that a substantial proportion of neurons in V1 respond to illusory brightness with tuning properties matching their responses to real luminance gratings, suggesting that higher-level V1 neurons play a pivotal role in processing illusory qualities under top-down modulation (Figure 2.4).

Further insights into the neural mechanisms of neon color spreading come from studies investigating higher visual areas. Hong and Tong (2017) employed fMRI combined with multivariate pattern analysis to investigate cortical responses to a color filling-in illusion. They found that while early visual areas (V1-V4) could reliably distinguish between two color-induced filled-in conditions based on activity patterns, only higher extrastriate visual areas (V3 and V4) showed activity patterns corresponding with the perceived colors. This indicates that the perception of filled-in surface color, a key aspect of neon color spreading, likely requires more extensive processing by these extrastriate areas.

Gerardin et al. (2018) used fMRI combined with Multi-Voxel Pattern Analysis (MVPA) to investigate the neural circuits involved in long-range color filling-in. They found that long-range color filling-in predominantly engages dorsal cortical areas, specifically V3A and V3B/KO, which are significantly correlated with the perceptual strength of filling-in. In contrast, uniform fields of chromaticity engage ventral areas, such as hV4 and LO, highlighting distinct neural pathways for processing different types of color perception. These findings suggest that similar mechanisms might be at play in neon color spreading, where the perception of color extends beyond its physical boundaries based on contextual cues.

The contribution of higher-level processes, such as perceptual inference and top-down influences, cannot be excluded based on the available evidence. Models that incorporate both bottom-up and top-down processing, such as the predictive coding framework (Rao and Ballard, 1999) and the Bayesian inference model (Kersten et al., 2004), may provide a more comprehensive account of neon color spreading. These models suggest that the perception of illusory color spread involves the integration of sensory evidence with prior knowledge and expectations about the properties of surfaces and illumination.

In conclusion, inner psychophysical studies have provided valuable insights into the neural mechanisms underlying neon color spreading, despite being less explored compared to the study of illusory contours. The complexity of the phenomenon, involving the interplay of color, brightness, and surface perception, has made it challenging to dissociate these factors in neuroimaging studies. While some studies suggest the involvement of early visual areas

like V1 in color filling-in and neon color spreading, others emphasize the role of higher extrastriate areas in processing filled-in surface color and long-range color filling-in. The engagement of distinct neural pathways for processing different types of color perception highlights the intricacy of the mechanisms underlying neon color spreading. Further research is needed to disentangle the contributions of early visual areas and higher extrastriate regions in the perception of neon color spreading, as well as to elucidate the neural interactions between color, brightness, and surface representation in this captivating visual illusion.

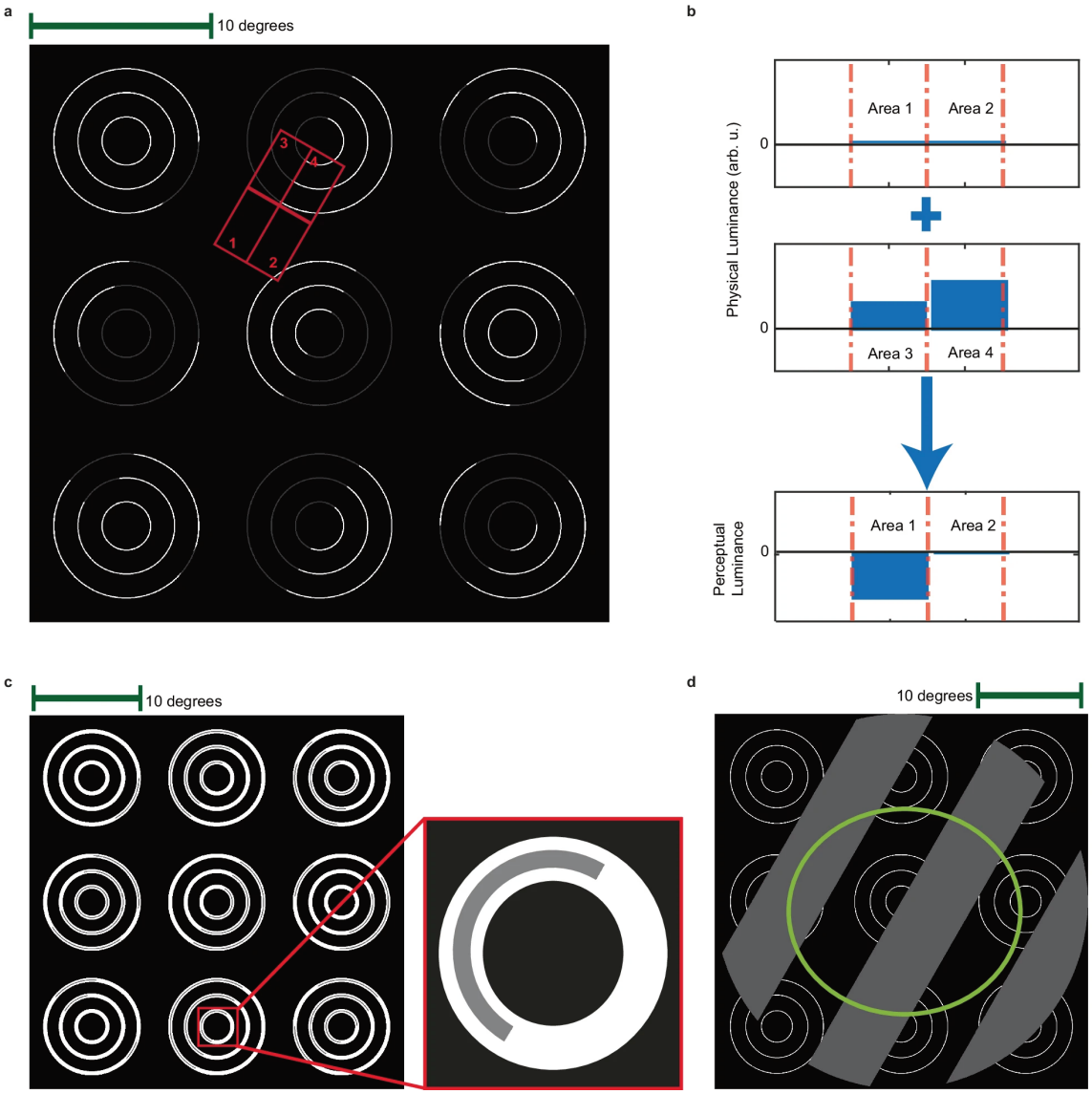

Fig. 2.4 The neurons of primary visual cortex respond to the illusory “darker than black” in neon color spreading, which is affected by feedback signals. Figure adapted from Saeedi et al. (2022).

2.2.3 Discussion

The evidence reviewed in this chapter highlights the ongoing debates regarding the mechanisms underlying illusory contours and neon color spreading, despite extensive research on these phenomena. While psychophysical and neuroimaging studies have provided valuable insights into the perceptual properties and neural correlates of these illusions, there remain open questions about the precise nature of the neural computations and the relative contributions of low-level and high-level processes.

One promising avenue for advancing our understanding of these illusions is the application of "inner" psychophysical methods, particularly reconstruction techniques, to probe the neural representations of illusory contours and neon color spreading. Reconstruction techniques, such as those based on fMRI data (Miyawaki et al., 2008; Shen et al., 2019a), allow researchers to decode and visualize the content of subjective perceptual experiences from patterns of brain activity. By applying these techniques to the study of visual illusions, we can directly compare the neural representations of illusory percepts with those of physically matched stimuli, providing a powerful tool for investigating the neural representations underlying these phenomena.

The use of reconstruction techniques offers several advantages over traditional "outer" psychophysical methods. First, they provide a more direct window into the neural representations of subjective experiences, bypassing the limitations of behavioral reports and the potential confounds of decision-making and motor processes. Second, they enable the study of illusory percepts in the absence of overt responses, allowing researchers to investigate the neural correlates of illusions under passive viewing conditions. Third, they offer the opportunity to compare the neural representations of illusory percepts across different stages of visual processing, from early visual areas to higher-order regions. Fourth, they provide access to holistic visual experiences that are not limited to specific aspects.

Furthermore, the application of reconstruction techniques to the study of visual illusions can inform the development and refinement of computational models of visual processing. By comparing the reconstructed illusory percepts with the predictions of different models, we can assess the validity and explanatory power of these models, guiding future theoretical and experimental work. This iterative process of model development, experimental testing, and model refinement can lead to a more complete and biologically plausible understanding of the neural mechanisms underlying visual perception and its illusions.

In conclusion, the fluid taxonomy of visual illusions and the ongoing debates regarding the mechanisms of illusory contours and neon color spreading highlight the need for novel approaches to investigate these phenomena. The application of reconstruction techniques,

in combination with psychophysical experiments and computational modeling, holds great promise for advancing our understanding of the neural basis of visual illusions.

Chapter 3

Methodology setup

In the previous chapter, we explored the fluid taxonomy of visual illusions and reviewed the existing literature on the psychophysical studies and neural correlates of illusory contours and neon color spreading. We highlighted the ongoing debates regarding the mechanisms underlying these illusions and discussed the potential of reconstruction techniques to advance our understanding of the neural basis of visual illusions.

In this chapter, we present our methodology setup for investigating the neural representations of illusory contours and neon color spreading using a combination of psychophysical experiments, computational modeling, and neuroimaging techniques. We begin by providing an overview of our study design, which leverages the concept of "shared representation" between veridical and illusory percepts to decode and reconstruct subjective experiences from brain activity. We then delve into the details of our decoding and reconstruction methods, discussing the deep neural network (DNN) architectures employed, their representations of visual features, and their correspondence to brain representations. Finally, we discuss the rationale behind our methodological choices and how they contribute to our goal of unraveling the neural mechanisms of visual illusions.

3.1 Overview of study design

The concept of "shared representation" is central to our approach to investigating the neural basis of visual illusions. As introduced in Chapter 1, "shared representation" is characterized by similar neural representations underlying the perception of a physical stimulus and those underlying the corresponding subjective experience, such as a visual illusion. By leveraging this principle, we can decode and reconstruct illusory percepts from brain activity patterns, providing a window into the neural mechanisms that give rise to these subjective experiences.

Our study design consists of two main stages: training and testing (Figure 3.1). In the training stage, we use a large dataset of natural images to train a decoding model that maps brain activity patterns to DNN features. This decoding model learns the relationship between the neural representations of visual stimuli and their corresponding DNN feature representations. We employ a linear decoding approach, which assumes that the visual features are linearly decodable from brain activity patterns in specific visual areas (Kamitani and Tong, 2005; Horikawa and Kamitani, 2017).

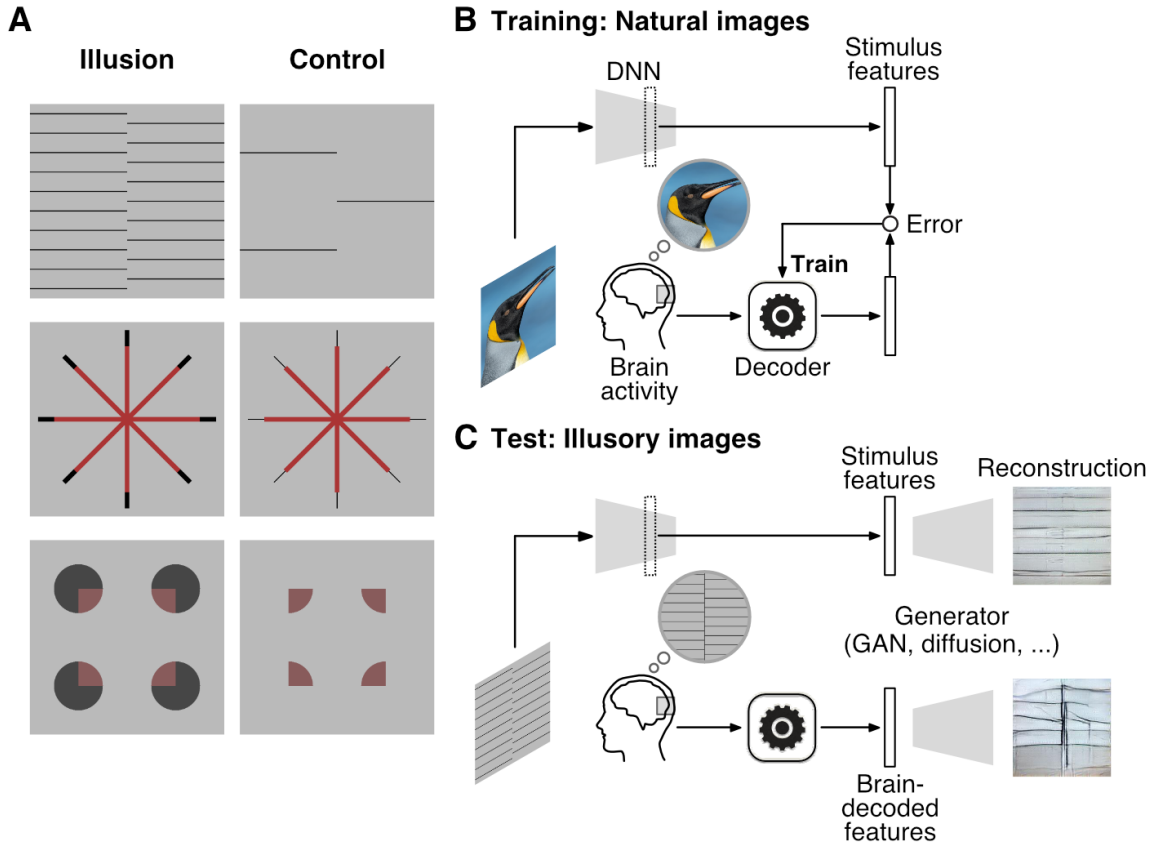

Fig. 3.1 Overview of study design. Figure from Cheng et al. (2023).

In the testing stage, we apply the trained decoding model to brain activity patterns evoked by illusory stimuli, such as illusory contour induced by offset-gratings and neon color spreading. By feeding the decoded DNN features into a pre-trained image generator, we reconstruct the subjective experiences of illusory contours and neon color spreading. This approach allows us to compare the reconstructed illusory percepts with the physical stimuli and the predictions of computational models, providing insights into the neural representations underlying these illusions.

To ensure that the reconstructed illusory features are not contaminated by inherent representations of illusions in the DNNs, we carefully select DNN architectures that do not explicitly encode illusory features. We also include control conditions, such as reconstructions from features directly extracted from illusory images and reconstructions from decoded features of non-illusory control stimuli, to validate the effectiveness of our approach in capturing the subjective experience of illusions.

3.2 Decoding and reconstruction methods

Recent advancements in deep learning have revolutionized the field of brain decoding and reconstruction, enabling researchers to access mental content using DNN representations (Shen et al., 2019a, 2019b; Horikawa and Kamitani, 2017). In this section, we discuss the DNN architectures employed in our study, their representations of visual features, and their correspondence to brain representations. We then delve into the details of our brain decoding and reconstruction methods, which leverage these DNN representations to reconstruct illusory percepts from brain activity patterns.

3.2.1 DNN architectures and representations of visual features

Deep neural networks (DNNs) have emerged as powerful tools for learning hierarchical representations of visual features from large datasets of natural images. These architectures consist of multiple layers of interconnected nodes, which progressively extract more complex and abstract features as the input propagates through the network. In this subsection, we provide an overview of the key DNN architectures used in our study and discuss their representations of visual features.

The Neocognitron is one of the earliest artificial neural network models inspired by the hierarchical processing in the visual cortex (Fukushima, 1980). It consists of alternating layers of simple and complex cells, which extract local features and provide invariance to spatial translations, respectively. This seminal model set the stage for future explorations into the parallels between artificial networks and biological brain structures. The theoretical framework for understanding the primate cortex's distributed and hierarchical processing, emphasizing how interconnected areas build complex representations, was provided (Felleman and Van Essen, 1991). Further refinement came from the formalization of hierarchical object recognition in the cortex, detailing the combination of simple to complex features for object representation (Riesenhuber and Poggio, 1999). While the Neocognitron laid the

groundwork for future DNN architectures, its feature representations are relatively simple compared to more modern networks.

LeNet is a pioneering convolutional neural network (CNN) architecture designed for handwritten digit recognition (LeCun et al., 1989). It consists of convolutional layers for feature extraction, followed by fully connected layers for classification. LeNet's success demonstrated the potential of CNNs for learning hierarchical feature representations, although its architecture is relatively shallow compared to more recent networks. AlexNet is a landmark CNN architecture that achieved state-of-the-art performance on the ImageNet object recognition challenge in 2012 (Krizhevsky et al., 2012). It consists of five convolutional layers, followed by three fully connected layers, and uses rectified linear units (ReLUs) as activation functions. AlexNet's success popularized the use of deep CNNs for computer vision tasks and sparked a renewed interest in deep learning. VGGNet is a deeper CNN architecture that builds upon the success of AlexNet (Simonyan and Zisserman, 2015). It consists of 16 or 19 layers, with smaller convolutional filters (3x3) and more layers than AlexNet. VGGNet achieved state-of-the-art performance on the ImageNet challenge in 2014 and has been widely used as a feature extractor for various computer vision tasks.

GoogLeNet, also known as Inception, introduced the concept of inception modules, which consist of parallel convolutional layers with different filter sizes (Szegedy et al., 2015). This architecture allows for the extraction of features at multiple scales and reduces the number of parameters compared to previous networks. GoogLeNet achieved state-of-the-art performance on the ImageNet challenge in 2014 and has been further refined in subsequent versions (Inception-v2, v3, and v4) Szegedy et al., 2016.

ResNet introduced the concept of residual connections, which allow for the training of much deeper networks (up to 152 layers) without suffering from the vanishing gradient problem (K. He et al., 2016). ResNet achieved state-of-the-art performance on the ImageNet challenge in 2015 and has become a popular backbone architecture for various computer vision tasks. DenseNet builds upon the idea of residual connections by introducing dense connections, where each layer receives inputs from all preceding layers (G. Huang et al., 2018). This architecture encourages feature reuse and reduces the number of parameters, while achieving state-of-the-art performance on various image classification benchmarks.

SENet introduced the concept of squeeze-and-excitation modules, which adaptively recalibrate channel-wise feature responses based on global context (Hu et al., 2018). These modules can be easily integrated into existing CNN architectures, such as ResNet and Inception, and have been shown to improve their performance on various image classification tasks.

EfficientNet is a family of CNN architectures that are optimized for both accuracy and efficiency (Tan and Le, 2020). These networks are designed using a compound scaling method, which uniformly scales the depth, width, and resolution of the network. EfficientNets have achieved state-of-the-art performance on various image classification benchmarks while being much smaller and faster than previous architectures.

Vision Transformers (ViT) are a recent class of DNN architectures that adapt the self-attention mechanism from natural language processing to computer vision tasks (Dosovitskiy et al., 2021). ViT models split an image into patches and process them using a transformer encoder, which captures global dependencies between the patches. ViT models have achieved state-of-the-art performance on various image classification benchmarks and have sparked a new direction in DNN architecture design.

Some DNN architectures like CNN have been shown to learn hierarchical representations of visual features that are similar to those found in the primate visual cortex (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014) and human visual cortex (Güçlü and Gerven, 2015; Horikawa and Kamitani, 2017) (Figure 3.2A). Cichy et al. (2016) compared the temporal dynamics of human object recognition with a deep convolutional neural network (CNN), finding hierarchical correspondence between brain activity patterns and DNN activations during object recognition, thus strengthening the link between artificial and biological intelligence. Studies have demonstrated that the activations of units in these networks correspond to the responses of neurons in various regions of the visual cortex, such as V1, V2, V4, and IT (Güçlü and Gerven, 2015; Cadieu et al., 2014). This correspondence has been further quantified using metrics such as the brain-score (Schrimpf et al., 2018), the brain hierarchy score (Nonaka et al., 2021) and the representational similarity analysis (RSA) framework (Kriegeskorte et al., 2008).

However, the exact nature and extent of the correspondence between DNN and brain representations remain an active area of research. Some studies have challenged the notion of a strong correspondence, suggesting that there may be limited overlap in visual representations between the human brain and CNNs (Xu and Vaziri-Pashkam, 2021). Others have proposed that the correspondence may be metric-dependent (Sexton and Love, 2022) (Figure 3.2B).

In our study, we leverage the hierarchical feature representations learned by DNNs to decode and reconstruct illusory percepts from brain activity patterns. We carefully select DNN architectures that have been shown to exhibit a strong correspondence to brain representations in the visual cortex, such as AlexNet and VGGNet. By using these networks as feature extractors, we aim to capture the neural representations underlying the subjective experience of illusory contours and neon color spreading.

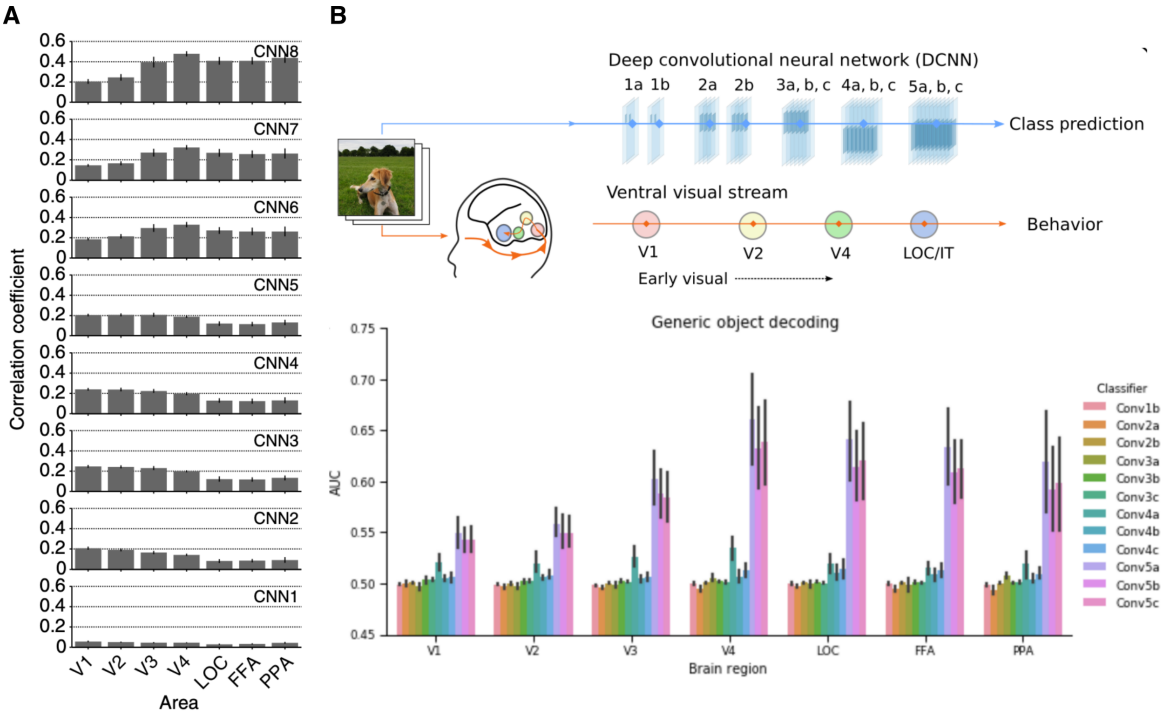

Fig. 3.2 Hierarchical representational correspondence between brain and DNN. While (A) evidence suggests some degree of hierarchical correspondence, (B) the extent and nature of the relationship between brain processing and DNNs remain debated. Figure adapted from Horikawa and Kamitani (2017) and Sexton and Love (2022).

3.2.2 Brain decoding and reconstruction

Brain decoding and reconstruction techniques have emerged as powerful tools for investigating the neural representations underlying visual perception and cognition. In this subsection, we review the key methods used for decoding and reconstructing the visual mental image from brain activity patterns.

Brain decoding of visual information

Brain decoding is a method of inferring cognitive states or stimulus properties from brain activity. Haynes (2015) provided a comprehensive review of brain decoding methods, discussing the challenges and opportunities in decoding mental states from brain activity. The review highlighted the importance of multivariate pattern analysis and the need for robust statistical methods to ensure the reliability of decoding results.

Multi-Voxel Pattern Analysis (MVPA), Linear Discriminant Analysis (LDA), and Support Vector Machines (SVMs) are some of the key methods used in brain decoding. MVPA analyzes patterns of activity across multiple voxels, LDA separates classes of brain activity patterns, and SVMs classify these patterns using hyperplanes. The accuracy of brain decoding varies greatly depending on the method, brain region, and cognitive function being studied. Additionally, the choice of neuroimaging modality (fMRI, EEG, MEG, etc.) is crucial for capturing specific aspects of brain activity.

Kamitani and Tong (2005) made a notable contribution by using fMRI to decode the orientation of gratings displayed to subjects (Figure 3.3). They trained an MVPA classifier on brain activity patterns in the primary visual cortex, achieving impressive accuracy in mapping subjective visual content. Brouwer and Heeger (2009) took a different approach, employing fMRI to reconstruct perceived colors from brain activity in the early visual cortex and developing a model that mapped these patterns to color representations. Their study suggested that brain regions encode color information beyond simple RGB values. Another pivotal study by Haxby et al. (2001) investigated the distributed and overlapping representations of faces and objects in the ventral temporal cortex, revealing distinct yet intermingled activation patterns for different object categories and laying the foundation for understanding object recognition in the brain.

Inspired by the brain's architecture, DNNs have emerged as powerful tools for brain decoding. They can learn complex feature representations from brain activity, leading to more accurate decoding of visual experiences (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and Gerven, 2015; Cichy et al., 2016; Horikawa and Kamitani, 2017). Horikawa and Kamitani (2017) introduced the concept of "generic object decoding." They trained a CNN on a large dataset of natural images and used the learned features

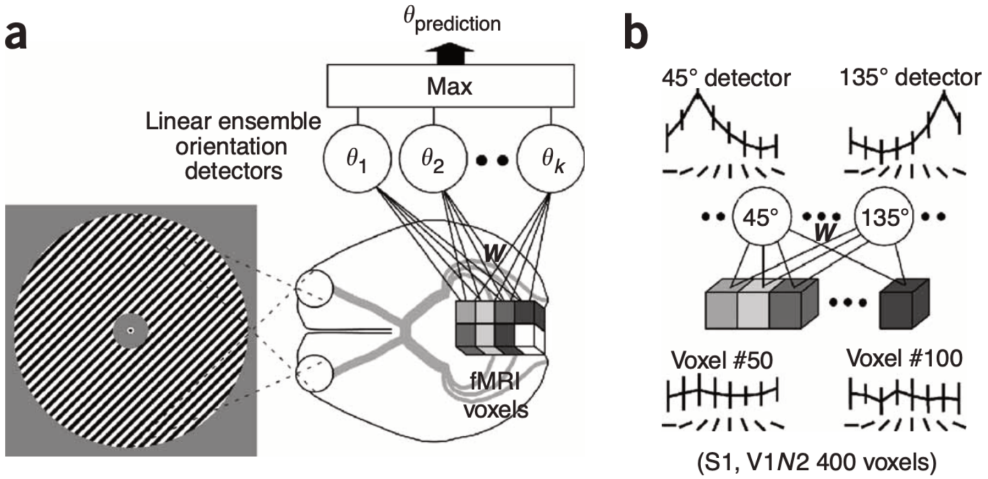

Fig. 3.3 Decoding perceived and attended orientation. Figure adapted from Kamitani and Tong (2005).

to decode object categories from fMRI data, even for objects that were not included in the training set (Figure 3.4). This approach highlighted the generalization capabilities of deep learning models and their potential for uncovering the neural representations of novel stimuli. Wen et al. (2018) decoded dynamic visual experiences from fMRI data. They used a combination of CNNs and recurrent neural networks (RNNs) to map brain activity patterns to sequences of video frames, enabling the reconstruction of dynamic visual experiences. This study showcased the potential of deep learning models for capturing the temporal dynamics of visual perception.

Reconstruction of visual mental image

The field of visual content reconstruction from brain activity is an area of research combining neuroscience and machine learning, aiming to translate neural signals into visually-interpreted materials such as images and movies. This interdisciplinary approach has seen remarkable progress, though it continues to face challenges due to the complexity of brain activity and the substantial data requirements for training deep learning models. Starting with the work of G. B. Stanley et al. (1999), a method was proposed for reconstructing natural scenes from the ensemble responses of neurons in the lateral geniculate nucleus (LGN). They employed a linear regression model to predict the pixel intensities of a natural scene based on the firing rates of LGN neurons. This early approach set the stage for more sophisticated methods in visual reconstruction.

Miyawaki et al. (2008) took a significant step forward by developing a method to reconstruct visual images from human brain activity using fMRI. They used a combination of

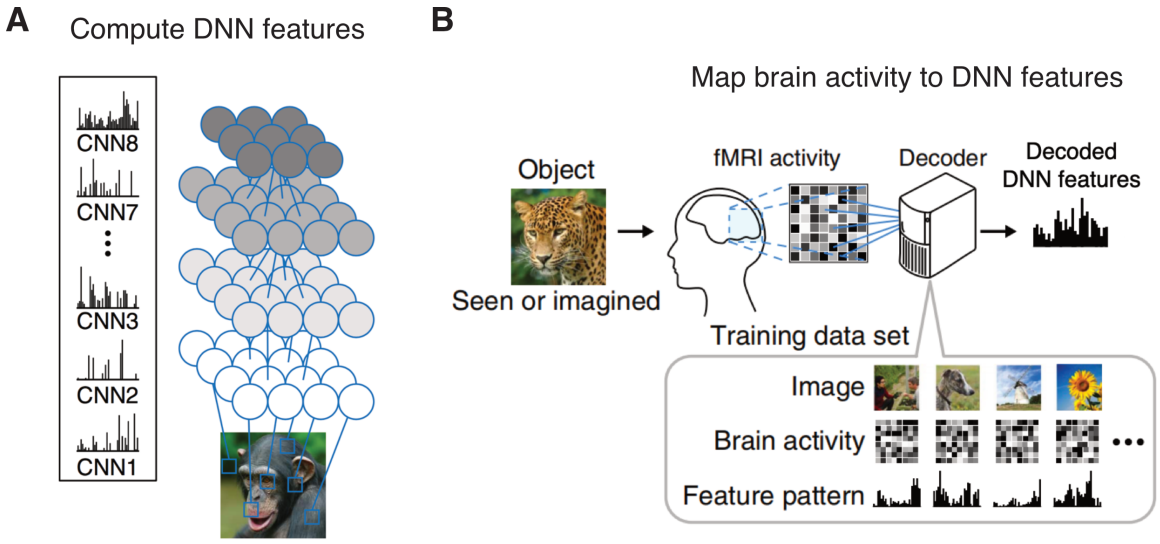

Fig. 3.4 Generic decoding of object category based on DNN representations. Figure adapted from Horikawa and Kamitani (2017).

multiscale local image decoders to extract features from fMRI data, followed by a support vector machine to reconstruct the image. This was complemented by the work of Naselaris et al. (2009), who introduced a Bayesian method for reconstructing natural images from brain activity. Their approach utilized a Bayesian framework to model the brain activity-image relationship, employing an iterative algorithm for image estimation. Nishimoto et al. (2011) further advanced this field by developing a method to reconstruct visual experiences from brain activity evoked by natural movies, using a linear regression model to predict movie frame pixel intensities from fMRI data.

Recent years have seen significant advancements in the field of visual image reconstruction from brain activity, driven by the development of deep learning techniques and the availability of large-scale neuroimaging datasets (Güçlütürk et al., 2017; Seeliger et al., 2018; Beliy et al., 2019; Shen et al., 2019a, 2019b; Gaziv et al., 2022). Güçlütürk et al. (2017) explored the use of deep adversarial neural networks (DNNs) to reconstruct perceived faces from brain activations. This involved training a generator network to create images resembling perceived faces, coupled with a discriminator network to differentiate between real and generated images. Seeliger et al. (2018) also utilized DNNs, specifically training a generative adversarial network (GAN) to generate images consistent with brain activity data. Shen et al. (2019a) developed a deep image reconstruction method using a convolutional neural network (CNN) to extract features from fMRI data, followed by an interactive process for image reconstruction (Figure 3.5). They also proposed an end-to-end deep image reconstruction

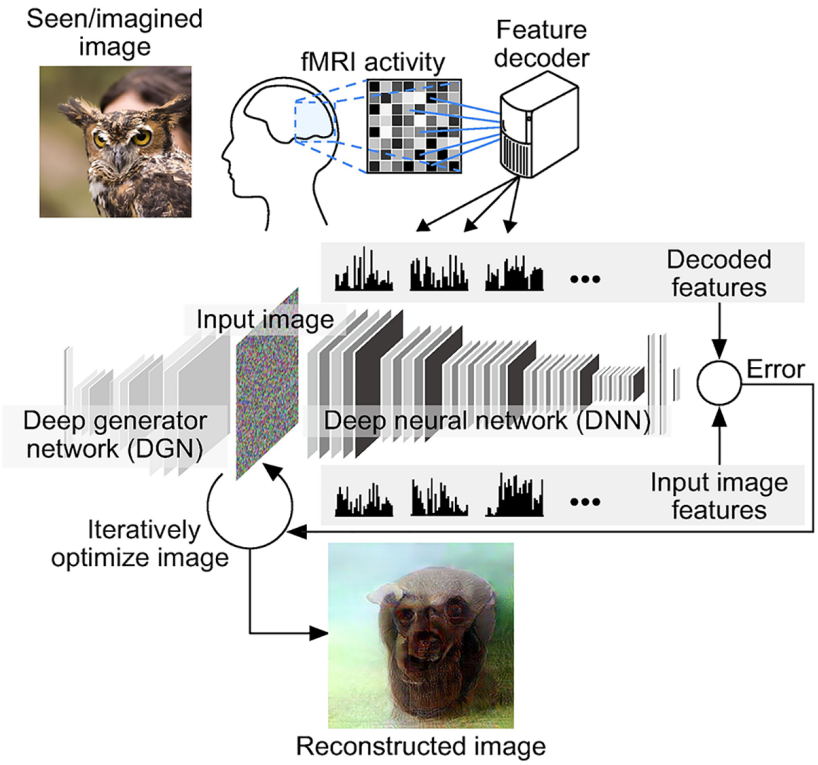

Fig. 3.5 Reconstruction of seen, imagined, or attended images. Figure from Shen et al. (2019a).

method, employing a GAN to directly map brain activity data to reconstructed images (Shen et al., 2019b).

Methods such as Ozcelik and VanRullen (2023) leveraged recently developed diffusion models and tested on a fMRI dataset not designed for reconstruction. However, the results failed to faithfully reflect the perceived image and are likely due to the "hallucinations" of the model (Shirakawa et al., 2023).

Despite these advances, the field continues to grapple with challenges. The complex and not fully understood relationship between brain activity and visual content poses a significant obstacle. Additionally, the large amount of data required to effectively train deep learning models remains a hurdle. Nevertheless, the potential impact of visual content reconstruction from brain activity on our understanding of the brain and its representation of the world is substantial. As this field continues to evolve, it promises to offer profound insights into the neural basis of perception and cognition.

3.3 Discussion

In this chapter, we presented our methodology setup for investigating the neural representations of illusory contours and neon color spreading. We discussed the rationale behind our methodological choices and how they contribute to our goal of unraveling the neural representations of visual illusions.

Selection of DNN architecture

One of the key considerations in our study is the selection of DNN architectures for feature extraction and decoding. We deliberately chose DNNs that do not inherently represent illusory features, such as AlexNet. This choice ensures that the reconstructed illusory features are not a result of the DNNs' pre-existing representations but rather reflect the neural representations of the illusions in the brain. By using DNNs that have been shown to exhibit a strong correspondence to brain representations in the visual cortex, we aim to capture the hierarchical processing of illusory percepts from early visual areas to higher-order regions.

The rationale behind this choice is twofold. First, if the DNNs were trained to represent illusory features, such as illusory contours or color, the reconstructed images would contain these features even in the absence of corresponding neural representations. By using DNNs that do not inherently represent illusory features, we can rule out this possibility and attribute the reconstructed illusory features to the neural representations encoded in the brain activity patterns.

Second, we want to leverage the representational correspondence between DNN layers and visual cortical areas to investigate the neural representations of illusory percepts at different levels of the visual hierarchy. Previous studies have shown that the activations of DNN layers exhibit a strong correspondence to the activations of visual cortical areas. Specifically, the outputs of the ReLU6 layer in CaffeNet, a variant of AlexNet, have shown comparable decoding performance across visual cortical areas (Horikawa and Kamitani, 2017). By using DNNs with known correspondences to visual cortical areas, we train a decoding model that can successfully predict visual DNN features from brain activity.

Linear decoding model

Another important aspect of our methodology is the use of a linear decoding model to map brain activity patterns to DNN features. This choice is motivated by the implications of linear decodability of visual features from fMRI data. By using a linear model, we ensure that the decoded features reflect the neural representations that are linearly accessible from brain activity patterns in specific visual areas. This approach allows us to investigate the form of illusory representations at different levels of the visual hierarchy.

The rationale behind using a linear decoding model is to focus on the neural representations that are most directly related to the illusory percepts. Non-linear decoding models, such as deep neural networks, can potentially capture visual features that may not be directly readout by a downstream area. By using a linear model, we constrain the decoded features to those that are linearly accessible from brain activity patterns, which are more likely to reflect the neural representations that give rise to the illusory percepts.

Moreover, using a linear decoding model allows us to directly compare the decoded features across different visual areas and stimulus conditions. If we were to use a non-linear decoding model, the decoded features would be transformed by the non-linear operations of the model, making it difficult to attribute differences in the decoded features to differences in the underlying neural representations (Naselaris et al., 2011). By using a linear model, we ensure that the differences in the decoded features reflect differences in the neural representations, allowing us to investigate the hierarchical processing of illusory percepts in the visual cortex.

Choice of reconstruction method

For the reconstruction of illusory percepts, we employ a feature decoding-to-generator approach, where the decoded DNN features are fed into a pre-trained image generator, such as a GAN. We fix the parameters of the generator during the reconstruction process to ensure that the differences in the reconstructed images arise solely from the decoded features, which in turn reflect the brain activity patterns. This approach enables us to compare the reconstructions from different visual areas.

The use of a pre-trained generator network for image reconstruction is motivated by the need to map the decoded DNN features back to the pixel space. Previous studies have demonstrated the effectiveness of generator networks, such as GANs and VAEs, for reconstructing visual images from brain activity patterns (Shen et al., 2019a; Beliy et al., 2019; Gaziv et al., 2022). However, training a generator network from scratch on a limited dataset of brain activity patterns can be challenging, leading to suboptimal reconstructions. To overcome this challenge, we use a pre-trained generator network that has been trained on a large dataset of natural images.

Criteria for successful reconstruction of illusory features

To evaluate the effectiveness of our reconstruction approach, we conduct several control experiments. First, we reconstruct images from DNN features directly extracted from illusory images, without involving brain activity data. We expect these reconstructions to not contain illusory features, such as illusory contours or color, as the DNN features are not explicitly

trained to represent these features. This control experiment allows us to rule out the possibility that the illusory features in the reconstructions are generated by the pre-trained generator network or the DNN features themselves.

Second, we reconstruct images from decoded DNN features of control images that do not induce illusory percepts. These reconstructions should also be devoid of illusory features, confirming that our decoding and reconstruction pipeline correctly reflected perceptual content. This control experiment allows us to ensure that the illusory features in the reconstructions are not artifacts of the noises from brain signals.

Finally, we reconstruct images from decoded DNN features of illusory images and expect these reconstructions to contain illusory features, reflecting the subjective experience of the illusions. By comparing the reconstructions from illusory and non-illusory images, we can investigate the neural representations of illusory percepts and test the predictions of computational models.

Overall, our methodology setup combines the strengths of linear decoding models, pre-trained generator networks, and carefully designed control experiments to investigate the neural representations of illusory contours and neon color spreading. By comparing the reconstructions generated from different visual areas and stimulus conditions, we aim to shed light on the hierarchical representations of illusory percepts in the visual cortex.

Chapter 4

Experimental design

In the previous chapters, we introduced the concept of visual illusions, their categorization, and the challenges in measuring and understanding these phenomena. We also discussed the general theories of visual perception and computational models that have been proposed to explain the mechanisms behind visual illusions. We outlined our proposed approach, which combines psychophysical experiments, computational modeling, and neuroimaging techniques to investigate the neural representations of illusory contours and neon color spreading.

In this chapter, we present the experimental design of our study, focusing on the types of illusions investigated, the visual stimuli used, and the fMRI data acquisition and preprocessing procedures. We introduce the two main types of illusions studied in this thesis: illusory contours induced by offset gratings and neon color spreading in the Ehrenstein and Varin configurations. We discuss the design of the visual stimuli, including the parameters chosen based on the intensity of perceived illusory features and the color selection for neon color spreading images. We then describe the fMRI experimental procedure, including the training and test sessions, and the data acquisition parameters. Finally, we present the data preprocessing steps, including the exclusion criteria and the delineation of regions of interest (ROIs) in the visual cortex. The content of this chapter is based on "MATERIALS AND METHODS" of Cheng et al. (2023).

4.1 Design of visual stimuli

Training images

In this study, a dataset of 3,200 natural images was selected from various online databases to facilitate robust training of decoding models:

ImageNet is a large-scale image database designed for use in visual object recognition software research (Russakovsky et al., 2015). It contains more than 14 million annotated images, categorized according to the WordNet hierarchy. Each image in ImageNet has been manually annotated with labels and bounding boxes around the object of interest. It's particularly famous for its annual competition, the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which has played a crucial role in advancing deep learning and computer vision research.

Flickr Material Database (FMD) is a collection of images specifically focused on materials (Sharan et al., 2014). It contains images depicting various materials like fabric, foliage, glass, leather, metal, paper, plastic, stone, water, and wood. Each category in FMD consists of 100 high-resolution images, carefully curated to represent the material under various conditions. This database is widely used in research related to material recognition and texture analysis.

COCO (Common Objects in Context) dataset is renowned for object detection, segmentation, and captioning (Lin et al., 2014). It contains over 330,000 images with more than 2.5 million labeled instances in 91 object types. COCO is distinguished by its focus on capturing objects in everyday scenes, thus providing context to the object of interest. This dataset is instrumental in training and evaluating algorithms for scene understanding and has been a benchmark dataset for several computer vision tasks.

The dataset included 1,200 object-specific images from ImageNet. Additionally, 1,000 material-specific images were sourced from the Flickr Material Database FMD, providing a diverse range of textural and material content. Another 1,000 images, encompassing both objects and scenes, were obtained from the COCO database. This amalgamation of images from diverse sources, including a targeted selection of 150 object categories from ImageNet and a strategic inclusion of images from FMD and COCO, was aimed at enhancing the dataset's diversity and quality (Figure 4.1). This approach also sought to mitigate biases typically observed in image datasets, such as a predominant focus on centrally-placed objects. The images were standardized by cropping them to a square format and resizing to 500 × 500 pixels, ensuring consistency across the dataset.

Test images

Images regarding illusory lines induced by offset-gratings and neon color spreading were designed for fMRI experiments. Shifted line gratings can induce the perception of illusory lines. A total of six images were created, each featuring illusory lines induced by offset gratings, as depicted in Figure 4.2A. The experimental setup involved two gratings against

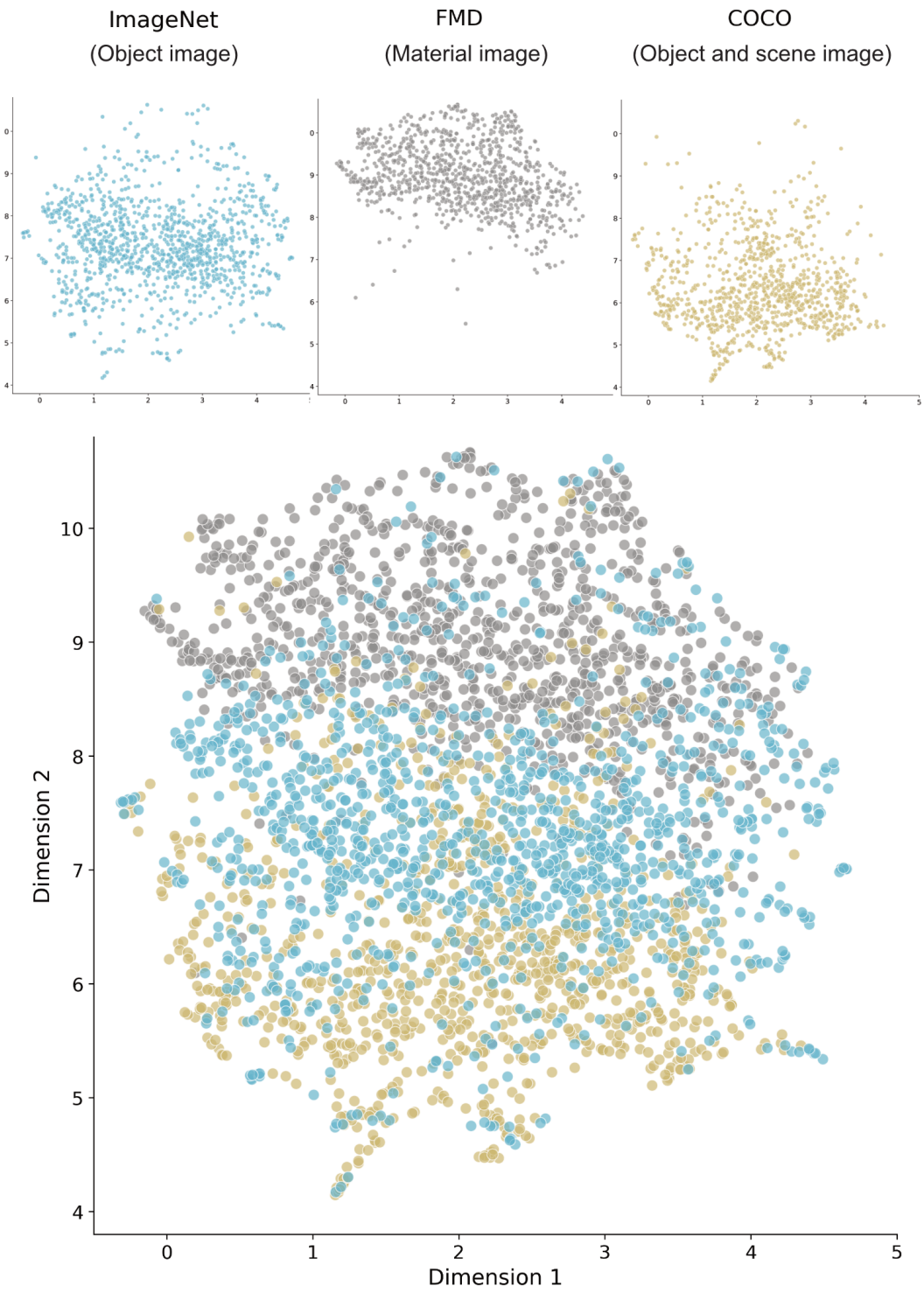

Fig. 4.1 Distribution of training images based on activations of DNN units (fc6 of CaffeNet). Dimension reduction was performed by UMAP using all 3,200 feature vectors derived from the images. Both individual datasets (top) and pooled datasets (bottom) are shown. Each dot represents a distinct image.

a gray background, marked by black lines at regular intervals and a half-cycle phase shift. The experiment was designed around three key parameters: the orientation of the illusory line (illusory orientation), the orientation of the inducer lines (inducer orientation), and the quantity of inducer lines. The inducer lines were fixed at 19 in number, with variances in illusory orientations (0°, 45°, 90°, and 135°) and inducer orientations (0° and 90°). This configuration yielded six distinct images of illusory lines, as the orientations of the illusory and inducer lines were not identical in any single image. For control, the number of inducer lines was reduced to 9 and 3 in cases of vertical or horizontal illusory orientations, respectively, resulting in four control images that elicited a diminished illusory perception. These experimental parameters were informed by previous behavioral studies involving human subjects (Soriano et al., 1996) and tailored for the current research. Additionally, ten positive control images were included, each featuring an actual line drawn where the illusory line was typically perceived. Preceding the fMRI studies, subjects were shown the stimulus images in the MRI scanner without being informed of the nature (illusory or control) of the images. This process confirmed that the illusory images consistently induced a clear illusion, whereas the control images resulted in a weaker or no illusory perception. This protocol was uniformly applied across all types of illusions examined in the study, encompassing both line illusions and neon color spreading, as detailed in the subsequent section.

In our investigation, two configurations of neon color spreading images were examined: the Ehrenstein and Varin configurations. For the Ehrenstein setup (Figure 4.2B), we developed four variations, distinguished by the number of lines (four and two) and the size (small and large) of the colored sections. These images were designed to fulfill transparency perception criteria by setting the luminance of colored lines between that of the surrounding black lines and the gray background. The illusion of color filling-in was induced by connecting colored and black lines of equal width, simulating a transparent colored disk. Control images were created by narrowing the black lines, disrupting the color filling-in perception. Additionally, positive control images with uniform color in the filling-in regions were included. In the Varin configuration (Figure 4.2C), the illusion was created using four disks arranged to suggest a rectangle, each containing a colored 90° sector and a black Pacman, with luminance values calibrated between the black Pacman and the gray background. Control images omitted the black Pacman, retaining only the colored sectors to diminish the color filling-in effect. Additional control images utilized either only black Pacmans or disks. Two positive control images with uniform color in the rectangular region were also prepared.

In total, 12 images for the Ehrenstein and 6 for the Varin configurations were produced. Red was chosen for the colored elements due to its effective reconstruction quality. The

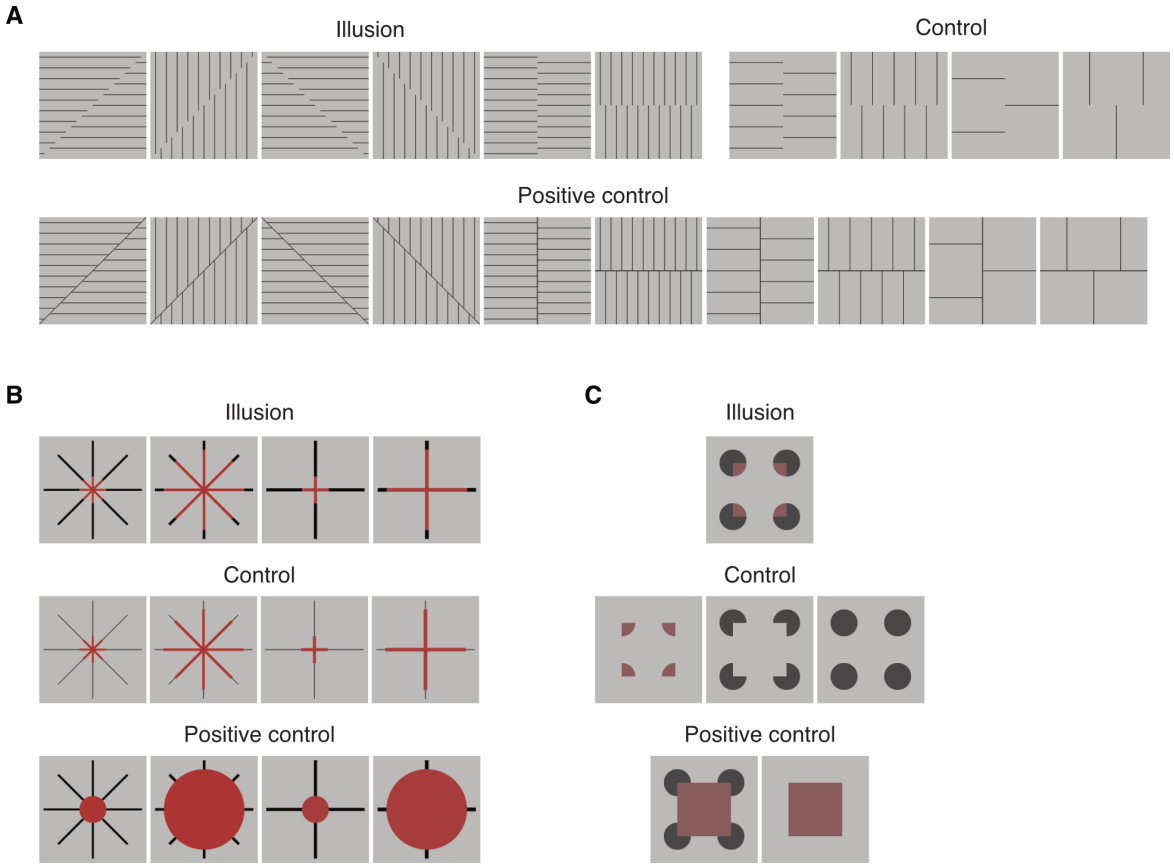

Fig. 4.2 Visual stimuli design. (A) Illusory and corresponding positive control images, featuring a central line in varied orientations. (B) and (C) Ehrenstein and Varin configuration for neon color spreading. Figure from Cheng et al. (2023).

saturation levels of the red color were adjusted to ensure clear perception of color filling-in across all subjects.

4.2 Design of fMRI experiments

fMRI data were acquired from a cohort of seven healthy individuals, comprising four males and three females aged between 25 and 36 years. Prior to participation, each subject gave informed consent in accordance with the study protocol, which had received approval from the Ethics Committee of the Graduate School of Informatics at Kyoto University. All participants possessed either normal vision or vision corrected to normal standards.

Training session

In our study, visual stimuli were presented to participants through rear-projection on a screen located within the bore of an MRI scanner, utilizing a luminance-calibrated liquid-crystal display projector. The stimuli, positioned at the screen's center, were displayed within a 12° by 12° visual angle against a uniform gray background. Participants were instructed to maintain fixation on the center of each image, guided by a small fixation circle measuring 0.3° by 0.3° in visual angle. The participants engaged in a one-back repetition detection task, responding to cues from the fixation spot. To minimize head movement during the fMRI data acquisition, each subject was provided with a custom-molded bite bar and/or a personalized headcase, supplied by CaseForge, Inc. Data collection involved multiple scanning sessions for each participant. This protocol was uniformly applied across all image presentation fMRI experiments, including those described in section 4.2.2.

A single set of training sessions was conducted, encompassing the presentation of each of the 3,200 natural images, distributed across 64 runs. The images sourced from different databases (ImageNet, FMD, and COCO) were allocated to distinct runs. Each run commenced with a 32-second rest period and concluded with a 6-second rest. Comprising 55 trials per run, the setup included 50 trials of unique images and five repetition trials, where the same image was displayed consecutively. The images were exhibited at a frequency of 1 Hz during each 8-second trial. For four subjects (subjects S1–4), the training data from previous studies (24, 26) employing a 1 Hz rate for the ImageNet dataset was repurposed. This rate was consistently used for the remaining subjects and for the FMD and COCO datasets. Between trials, no rest period was scheduled. The onset of each trial was indicated by changing the fixation spot color to red 0.5 seconds beforehand, reverting to white at trial commencement. The participants were tasked with a one-back repetition detection task, pressing a button upon identifying a repeat stimulus. Throughout the run, subjects were instructed to maintain steady

fixation, with their alertness assessed via performance in the one-back task. For subjects S1–4, five sets of training sessions were repeated (yielding 16,000 training samples), while for subjects S5–7, two sets were conducted (resulting in 6,400 training samples). Prior to analyzing the illusion test data, we verified through an independent test dataset that training data with fewer repetitions were sufficient for comparable model performance.

Reuse of fMRI data from previous studies

For four subjects (S1–4), who participated in prior studies (Shen et al., 2019a; Horikawa and Kamitani, 2022), we utilized their previously published data as part of our training dataset. The training session data for these subjects are available online at OpenNeuro (`https://openneuro.org/datasets/ds003430/versions/1.1.1` for subjects S1–3 and `https://openneuro.org/datasets/ds001506/versions/1.3.1` for subject S4). Additional training data were also collected to augment the pre-existing datasets.

Test session

Each of the 38 test images (comprising 10 illusory line images, 10 real line images, 12 Ehrenstein images, and 6 Varin images) was presented to the subjects 20 times. However, for participant S4, data were collected for only 32 images due to their unavailability for the Varin configurations. To ensure comparable fMRI signal baselines with the training sessions, an equal number of random natural images were included. The test session comprised 40 runs, each starting with a 32-second rest period and ending with a 6-second rest. Each run contained 42 trials, including 38 distinct images (19 test and 19 natural images) and four repetition trials, with the presentation order randomized. Between successive illusion image trials, natural images were inserted to mitigate any after-effects. The images were displayed at a frequency of 0.625 Hz during each 8-second trial, a rate chosen based on preliminary experiments to enhance the stability of illusory perceptions. For the neon-color illusion stimuli (Ehrenstein and Varin), only the colored portions were flashed to accentuate the illusion.

MRI acquisition

Functional Magnetic Resonance Imaging (fMRI) is a non-invasive technique used to capture neural activity in the brain (Ogawa et al., 1990; Belliveau et al., 1991). It works based on the principle that cerebral blood flow and neuronal activation are coupled. When a brain region is more active, it consumes more oxygen, leading to an increase in blood flow to that area. fMRI detects these changes through the blood-oxygen-level-dependent (BOLD) contrast. This contrast arises because oxygenated and deoxygenated hemoglobin have different magnetic

properties. fMRI scanners detect these differences, allowing researchers to map and measure brain activity by observing changes in blood flow, thus indirectly inferring neural activity.

fMRI data were acquired using a 3.0-Tesla Siemens MAGNETOM Verio scanner located at the Kyoto University Institute for the Future of Human Society. Functional imaging of the entire brain was performed via interleaved T2*-weighted gradient-echo echo-planar imaging (EPI) scan, characterized by specific parameters: repetition time (TR) of 2000 ms, echo time (TE) of 43 ms, a flip angle of 80°, a field of view (FOV) of 192×192 mm², and a voxel size of $2 \times 2 \times 2$ mm³. The scanning protocol involved 76 slices with no gap, employing a multiband factor of 4. Additionally, high-resolution T1-weighted (T1w) magnetization-prepared rapid acquisition gradient-echo (MP-RAGE) images of the entire head were obtained, featuring a TR of 2250 ms, TE of 3.06 ms, inversion time (TI) of 900 ms, a flip angle of 9°, an FOV of 256×256 mm², and a voxel size of $1.0 \times 1.0 \times 1.0$ mm³.

4.3 Data preprocessing

Preprocessing and exclusion procedure

The MRI data preprocessing in our study was conducted using the fMRIPrep pipeline (version 1.2.1). Each run's functional data began with estimating a BOLD reference image via FMRIPREP's custom methodology. Subsequent steps included motion correction using MCFLIRT (FSL version 5.0.9) and slice time correction with 3dTshift (AFNI version 16.2.07), both based on the BOLD reference. Co-registration of the corresponding T1w image was executed with FreeSurfer's bbregister (version 6.0.1), followed by resampling the BOLD time-series to its original space using antsApplyTransforms from ANTs (version 2.1.0) with Lanczos interpolation. The resampled BOLD time series were adjusted for hemodynamic delays and nuisance parameters, including baseline, linear trends, and motion-related components, were regressed out.

Prior to the commencement of fMRI data collection, criteria for data exclusion were already established, with the exclusion process completed before advancing to the main analyses. Initially, any runs demonstrating low performance (defined as a hit rate of 50% or less) in the one-back repetition detection task were excluded. This criterion aimed to eliminate data from periods where the subject's level of alertness might have been compromised. Consequently, this led to the exclusion of one run for subject S1 and two runs for subject S2. Furthermore, runs characterized by significant head motion, specifically those with a maximum translation of 2 mm or more, were also omitted (Figure 4.3). This resulted in the

exclusion of two runs from subject S5. Post data preprocessing, we successfully acquired between 18 to 20 single-trial samples per image for each participant.

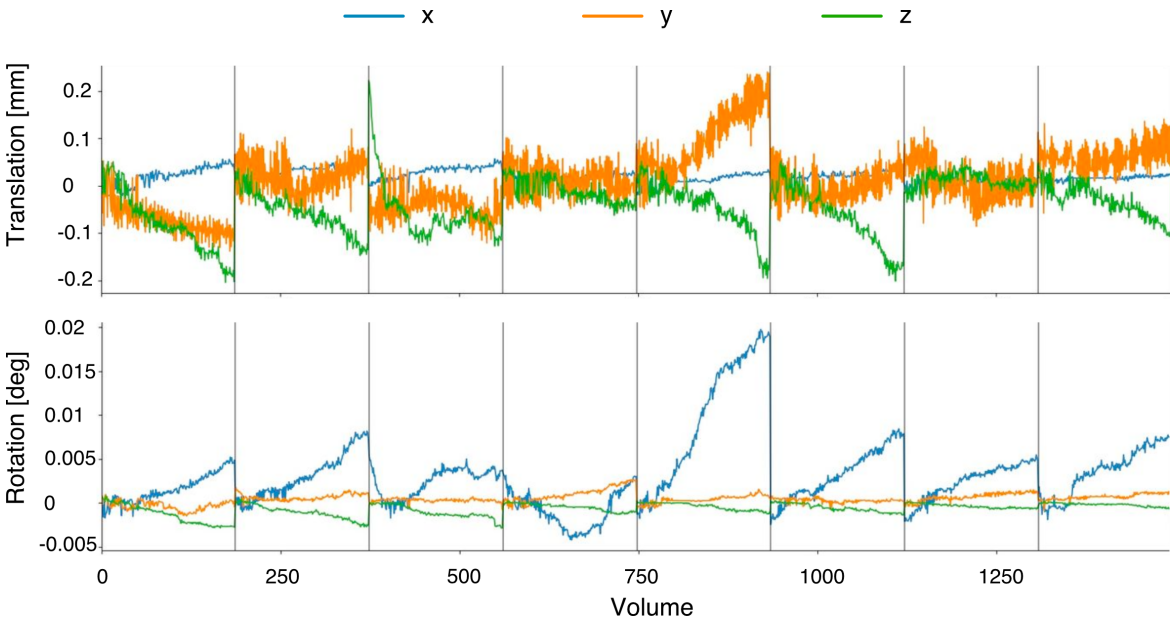

Fig. 4.3 Head motion during one example session. The top and bottom panels show translation (mm) and rotation (degree), respectively.

Region of interest (ROI)

Specific visual areas were delineated using retinotopy and functional localizer experiments, each with its distinct mechanisms.

Retinotopy experiments provide detailed information about the correspondence between the visual field and its cortical representation. fMRI retinotopic mapping is distinct from traditional volume-based analyses in that it often involves a surface-based analysis of data and a phase-encoded paradigm. Standard retinotopy experiments (Engel et al., 1994; Sereno et al., 1995) were used to delineate V1, V2, V3, and V4, respectively (Figure 4.4). Functional localizers are used in neuroimaging studies to pinpoint specific brain areas involved in certain tasks. The tasks and stimuli used to localize particular regions vary, but the method is crucial for identifying functional architecture within the brain. These experiments often involve presenting subjects with specific stimuli (e.g., faces, places, or objects) and observing the brain regions that show increased activity in response. The lateral occipital complex (LOC), fusiform face area (FFA), and parahippocampal place area (PPA) were identified using conventional functional localizers (Kanwisher et al., 1997; Epstein and Kanwisher, 1998; Kourtzi and Kanwisher, 2000 2). A higher visual cortex (HVC) region encompassing

the LOC, FFA, and PPA was manually delineated, and the visual cortex (VC) was defined as a combination of V1–V4 and HVC.

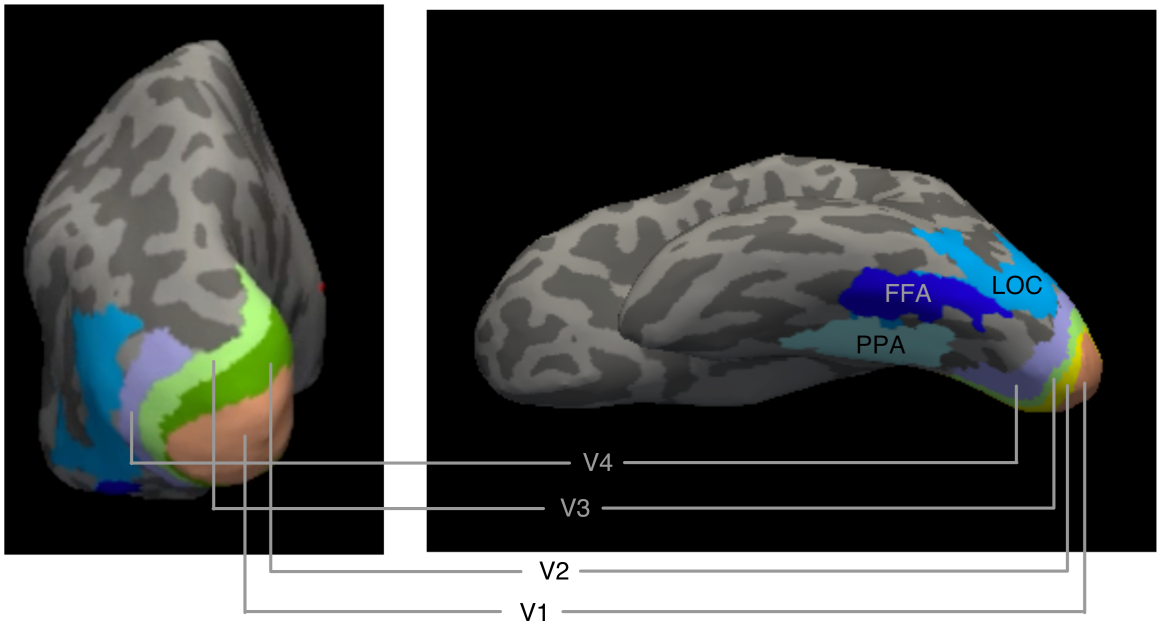

Fig. 4.4 Functional visual areas of one participant. Left and right panels show posterior and bottom views of the brain.

4.4 Discussion

The experimental design of our study was carefully considered to ensure the effective investigation of the neural representations of illusory contours and neon color spreading.

As for the design of visual stimuli, we used natural images with diversified styles and categories for the images in the training sessions. This helps us train better decoding models that map brain activity to feature values of DNN units by ensuring DNN units are activated with enough frequency and increasing the variance of their activations across images. For the test session, we designed the visual stimuli to include illusory, control, and positive control conditions, allowing us to compare the neural responses to illusory percepts with those to non-illusory stimuli and stimuli that mimic illusory percepts. The parameters of the visual stimuli were chosen based on the intensity of perceived illusory features, ensuring that different individual subjects had the strongest illusory perception under the same sets of parameters. For neon color spreading images, we selected the color red based on previous reconstruction studies showing the most reliable performance for this color compared to others (Shen et al., 2019a).

The fMRI data collection procedure was designed to maximize the quality and reliability of the neural data. We collected data from multiple scanning sessions for each participant, using a custom-molded bite bar and/or a personalized headcase to minimize head movement during data acquisition. The exclusion criteria for data were decided prior to data collection, ensuring an unbiased approach to data analysis. The delineation of ROIs in the visual cortex was performed by researchers based on retinotopy and functional localizer experiments.

However, our experimental design also has some limitations and potential areas for improvement. The brain data collection has a long time span, which is time and resource-intensive. In the future, the direction of training models with less data while maintaining performance is worth investigating. This could involve the development of more efficient data acquisition protocols or the use of transfer learning techniques to leverage pre-existing datasets (J. K. Ho et al., 2023). Additionally, the delineation of ROIs is currently performed manually by researchers, which can be time-consuming and subject to variability. The development of more automatic methods for determining ROIs, such as using functional connectivity-based parcellation or machine learning techniques, could improve the efficiency and reproducibility of this process (Jiahui et al., 2023).

Despite these limitations, our experimental design provides a strong foundation for investigating the neural representations of illusory contours and neon color spreading. By carefully designing the visual stimuli, optimizing the fMRI data acquisition procedure, and employing rigorous data preprocessing steps, we aim to uncover the neural mechanisms underlying these fascinating perceptual phenomena. The insights gained from this study could have important implications for our understanding of visual perception, the development of computational models of vision, and the design of visual interfaces and technologies.

Chapter 5

Reconstructing illusory percepts

In the previous chapter, we presented the experimental design of our study, focusing on the types of illusions investigated, the visual stimuli used, and the fMRI data acquisition and preprocessing procedures. We discussed the rationale behind our methodological choices and how they contribute to our goal of unraveling the neural mechanisms of visual illusions.

In this chapter, we describe our decoding and reconstruction analysis, which employs a feature decoding-to-generator approach to reconstruct illusory percepts from decoded DNN features. We present the results of our reconstruction analysis, demonstrating the successful reconstruction of illusory contours and neon color spreading from brain activity patterns. The content of this chapter is based on sections "RESULTS: Illusory stimuli and the reconstruction model" and "RESULTS: Reconstructed images", and "Supplementary Materials" of Cheng et al. (2023).

5.1 Decoding analysis

5.1.1 Method

Multivoxel decoders were developed by employing linear regression models to predict stimulus features from fMRI voxel signals, a technique consistent with previous studies (Horikawa and Kamitani, 2017; Shen et al., 2019a; Horikawa and Kamitani, 2022). For training, fMRI samples from the training session were utilized, encompassing 16,000 trials for subjects S1–4 and 6,400 trials for subjects S5–7. A separate decoder was trained for each combination of DNN units and specific brain areas, which included the whole visual cortex as well as individual visual subareas. Let Y represent the stimulus feature vector from a DNN unit, and X denote the matrix of fMRI voxel signals, where each column corresponds to a

voxel's signal and each row represents a trial. The linear regression model aims to predict Y from X using the equation:

$$Y = XW + \varepsilon \tag{5.1}$$

Here, W is the weight matrix that the model seeks to determine, and represents the error term. The model is trained to find the optimal W that minimizes the difference between the predicted Y and the actual DNN feature values. The optimization of W is achieved through least-square minimization with L2 regularization, which can be expressed as:

$$\min \|Y - XW\|^2 + \lambda \|W\|^2 \tag{5.2}$$

In this formula, $\|Y - XW\|^2$ is the least-square term representing the sum of squared differences between the predicted and actual values, and $\|W\|^2$ is the L2 regularization term, with being the regularization parameter (set to 100 in this study). The regularization term prevents overfitting by penalizing large weights. For each target DNN unit, voxels exhibiting the highest correlation (as determined by the absolute Pearson correlation coefficient) were selected from each brain area based on the training data. These selected voxels, capped at 500, served as inputs to the decoder. It's important to note that while a decoder for a particular DNN unit did not utilize the full extent of the brain region, the collective use of different decoders across various DNN units potentially encompassed a substantial portion of the entire brain region. These trained decoders were then applied to the test dataset's fMRI data to predict individual DNN units' feature values, termed "decoded features." Each visual image's decoded features were derived from a single-trial fMRI sample. To ensure consistency in the distributions between stimulus and decoded features, these decoded features underwent normalization, aligning the variance of normalized decoded features within a layer with the mean variance of DNN feature values. This mean variance was computed from an independent set of 10,000 natural images. Furthermore, the average of the normalized decoded features was maintained at the same level as that of the original, unnormalized features, facilitating subsequent image reconstruction analyses.

5.1.2 Results

Verification on natural test images

To test the performance of decoders individually trained for each subject, the trained decoder was applied to the fMRI activity induced by 50 natural test images. The fMRI data of subjects S1–4 were reused from previous studies (Shen et al., 2019a; Horikawa and Kamitani,

2022). Comparable performance was observed among subjects with all available training data (Figure 5.1).

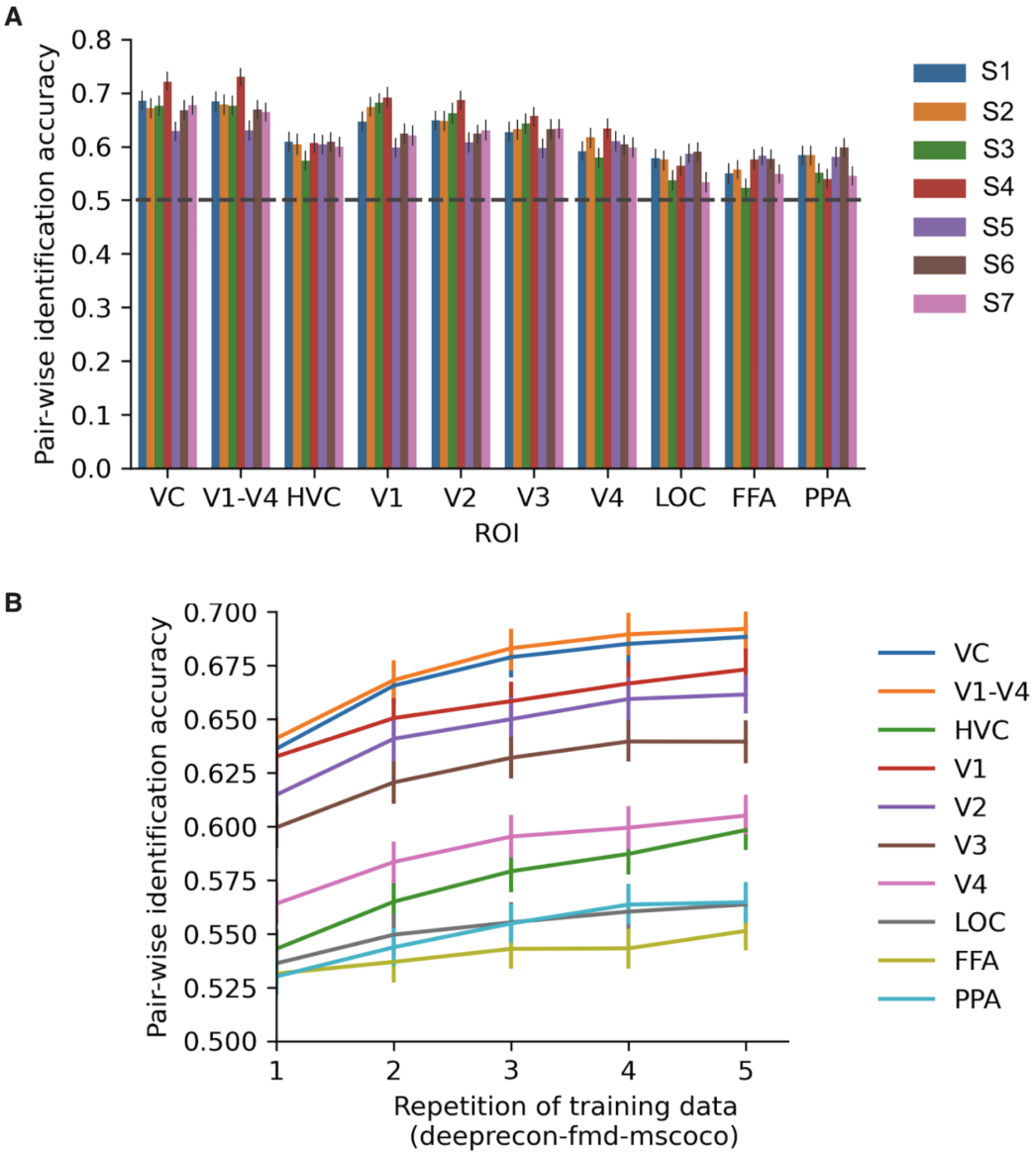

Fig. 5.1 Pair-wise identification accuracy based on decoded fc6 features of natural images from single-trial fMRI samples. (A) Comparable performance among subjects with all available training data. (B) The effect of increasing repetitions of training data. Bars denote 95% confidence interval.

5.2 Reconstruction analysis

5.2.1 Method

In our study, we utilized a generator network (Dosovitskiy and Brox, 2016) initially pre-trained within a Generative Adversarial Network (GAN) framework, accessible from the provided link (`https://lmb.informatik.uni-freiburg.de/resources/binaries`). The purpose of this image generator was to reconstruct the original input images from the rectified outputs of the fc6 layer of the CaffeNet model, utilizing training data derived from the ImageNet dataset. The training framework of GAN consists of three modules: generator, discriminator, and comparator. The generator in a GAN is a deep neural network whose role is to create data that is similar to the real data it's been trained on. The input to the generator is typically a random noise vector, but the pre-trained generator here uses DNN feature as input. Through the training process, this input is transformed into data (in this case, images) by passing it through layers of the network. Each layer learns to transform its input more and more until the output is a realistic image. The features learned and represented in these deep layers are complex and help in generating high-quality images. The discriminator is another deep neural network that receives two kinds of inputs: real images from the training dataset and fake images produced by the generator. Its task is to distinguish between the two. Essentially, it's a binary classifier being trained to output high probability for real images and low probability for generated images. Through training, it becomes better at detecting subtle and complex features that differentiate real images from generated ones. The inputs of the comparator are the same as that of the discriminator. The weights of the comparator are fixed during training. It compares the high-level feature space of the real and fake images and the differences are used as loss to optimize the generator. According to experiments, the conv5 layer of CaffeNet can be used for comparison. During training, generator and discriminator undergo backpropagation, where they receive feedback on their performance and update their weights and biases to perform better in the next iteration. The generator improves to create more realistic images, while the discriminator improves its ability to classify real and fake images. This process is repeated until the generator produces high-quality images that the discriminator can't easily distinguish from real images.

In this study, we incorporated reconstruction models that included DNN feature decoders and an image generator as depicted in Figure 3.1. The DNN feature decoders were akin to those utilized in our prior investigations (refer to section 5.1). We targeted the unit activations of a feedforward convolutional neural network (Krizhevsky et al., 2012) for decoding purposes. We collected fMRI signals from the visual cortex (VC) of seven subjects, encompassing both the early visual areas and the ventral object-responsive areas (refer to

Chapter 4). Our hypothesis was predicated on the assumption that the majority of natural images evoke perceptions that closely align with their physical attributes (veridical perception) and that our trained decoders would be capable of effectively mapping brain activity to these perceptual features, even in the absence of explicit information regarding subjective appearances.

During the testing phase, we employed the fMRI dataset specifically collected for this study, which included exposures of seven subjects to illusory images, along with control and positive control images, interleaved with sequences of natural images (refer to Chapter 4). Each test image was flashing at a frequency of 0.625Hz for a duration of 8 seconds in each trial, with the entire sequence being repeated across twenty trials. Utilizing the trained DNN feature decoders (Figure 3.1B), we obtained the decoded features from the single-trial brain activity of the test dataset (Figure 3.1C). These decoded features were subsequently fed into a generator. This generator, previously trained on a vast dataset of natural images, was tasked with reverting the stimulus DNN features of each image back to its original visual representation.

When using GAN as the generator, we integrated decoded fc6 features from either a single-trial or trial-averaged fMRI sample into the generator network, resulting in a reconstructed image. The selection of fc6 was purposefully aligned with our objective to examine the representations of illusory percepts across various individual brain regions. This particular layer, fc6, has exhibited substantial and uniform linear decodability across all visual areas (Horikawa and Kamitani, 2017), supporting its use in our study. Corroborating our choice, recent research has indicated a strong correspondence between all ventral stream visual areas and DNN layers of comparable depth (Sexton and Love, 2022). Further empirical verification ascertained that the fc6 layer, when mapped back to the original image, yielded more precise reconstructions than its subsequent layers (Dosovitskiy and Brox, 2016).

Consistent with the main methodology, the decoded fc6 features were inputted into a diffusion-based image generator for subsequent reconstruction. When using iCNN, CaffeNet was selected as the target for feature decoding to facilitate comparative analysis with other methodologies. Consistent with the approach delineated in reference (Shen et al., 2019a), we extracted feature values prior to the rectification operation across eight distinct layers, encompassing conv1-5 and all fully connected layers. To optimize the reconstruction, we employed the identical loss function and natural image prior as established in previous works. The optimization problem was addressed using stochastic gradient descent with momentum, iterating through this process for a total of 200 iterations to refine and enhance the accuracy of our reconstructed images. This methodical approach was aimed at ensuring a robust and

reliable comparison while adhering to the proven strategies in feature decoding and image reconstruction.

5.2.2 Results

Verification of main reconstruction method using natural images

We adopted GAN as the main reconstruction method within our study, influenced by preliminary findings that highlighted its propensity to generate reconstructions with heightened contrast, particularly evident in geometric shapes and patterns, in contrast to other evaluated methods. The selected methodology also demonstrated the capability to reconstruct natural images (Fig. 5.2), achieving a quality commensurate with that observed in a prior research (Shen et al., 2019a).

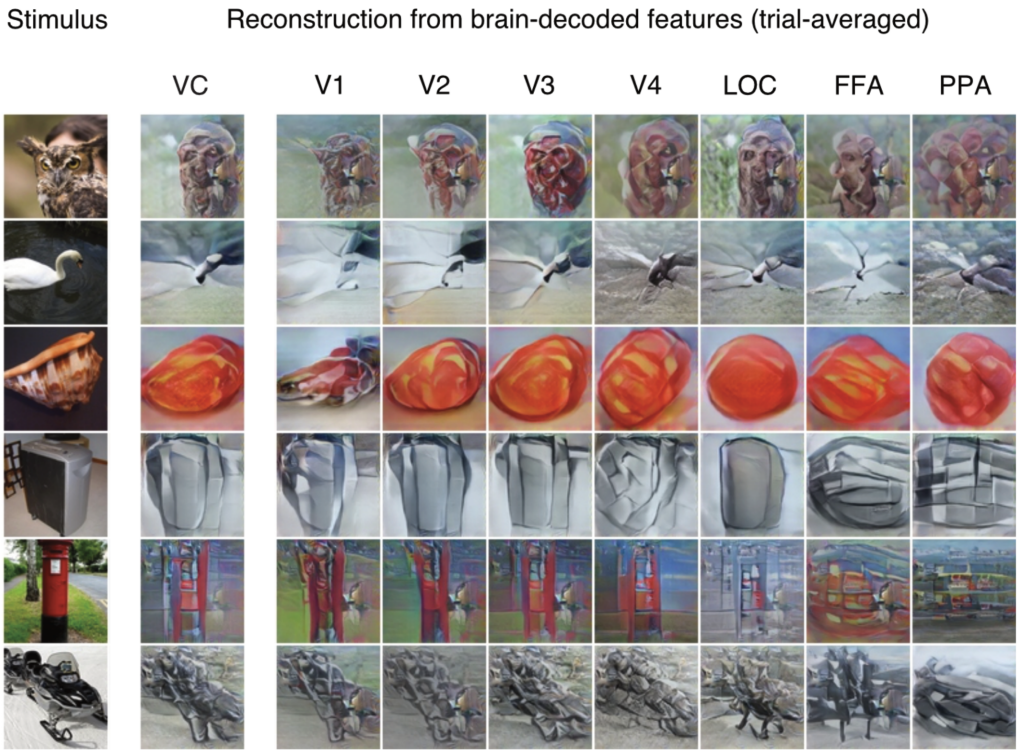

Fig. 5.2 Reconstructions from brain activity for natural images. The results in each column were generated using averaged fMRI signals across 24 trials, obtained from the entire visual cortex (VC) and specific visual areas of subject S2. Figure from Cheng et al. (2023).

Reconstructions from stimulus features

In our study, we analyzed the unit activations of a DNN as stimulus features in response to

visual image inputs, including illusory images, corresponding control images, and natural images. In order to evaluate whether the involvement of brain activity is essential for the reconstruction of illusory features, the stimulus DNN features derived from the test illusory images were input into the generator. We ascertained that our reconstruction framework did not inadvertently generate any artificial lines or colors that aligned with the perceived illusions. Reconstructions were derived from stimulus DNN features for various representative configurations ("Stimulus features" in Figure 21). Across all tested configurations, it was observed that reconstructions based solely on stimulus features did not manifest any components of illusory percepts. While certain DNN models might be conditioned to exhibit illusory representations independently of brain activity (Watanabe et al., 2018; Gomez-Villa et al., 2020; Lotter et al., 2020; Sun and Dekel, 2021), the DNN model employed in our feature decoding is fundamentally a feedforward convolutional neural network, designed primarily for object classification (Krizhevsky et al., 2012). Consequently, it is less probable for it to encode contextual features associated with illusions, such as illusory lines and colors. This assertion was substantiated by our discovery that individual units within the DNN representation did not demonstrate orientation or color tuning that corresponded between actual and illusory features.

To investigate the impact of stimulus-independent noise in brain-decoded features potentially leading to the reconstruction of illusory components, we incorporated an experimental control by introducing noise to the stimulus fc6 features prior to their introduction into the generator. Given the absence of definitive prior knowledge regarding the noise distribution, a non-parametric approach was employed to approximate the noise. This approach was based on the assumption that the noise associated with each individual DNN unit adheres to a consistent yet unknown probability distribution across the non-illusory trials for a given subject, while acknowledging that different units might follow distinct distributions. The analysis was conducted on a subject-specific basis. We commenced by calculating the empirical noise distributions, achieved by aggregating the discrepancies observed between the decoded features and the actual stimulus features across trials devoid of illusory components. Subsequently, for each DNN unit, a noise value was randomly sampled from the corresponding empirical distribution and then superimposed onto the stimulus features of an illusory image. This methodology was meticulously designed to elucidate the influence of inherent noise on the perceived reconstruction of illusory elements within the generated images. Similar to the results from stimulus features, reconstructions after introducing noise to stimulus features did not contain any components of illusory percepts (Figure 5.3). Hence, it appears that our reconstruction model faithfully translates visual information represented in brain activity, adhering to the coding principles of accurate perception.

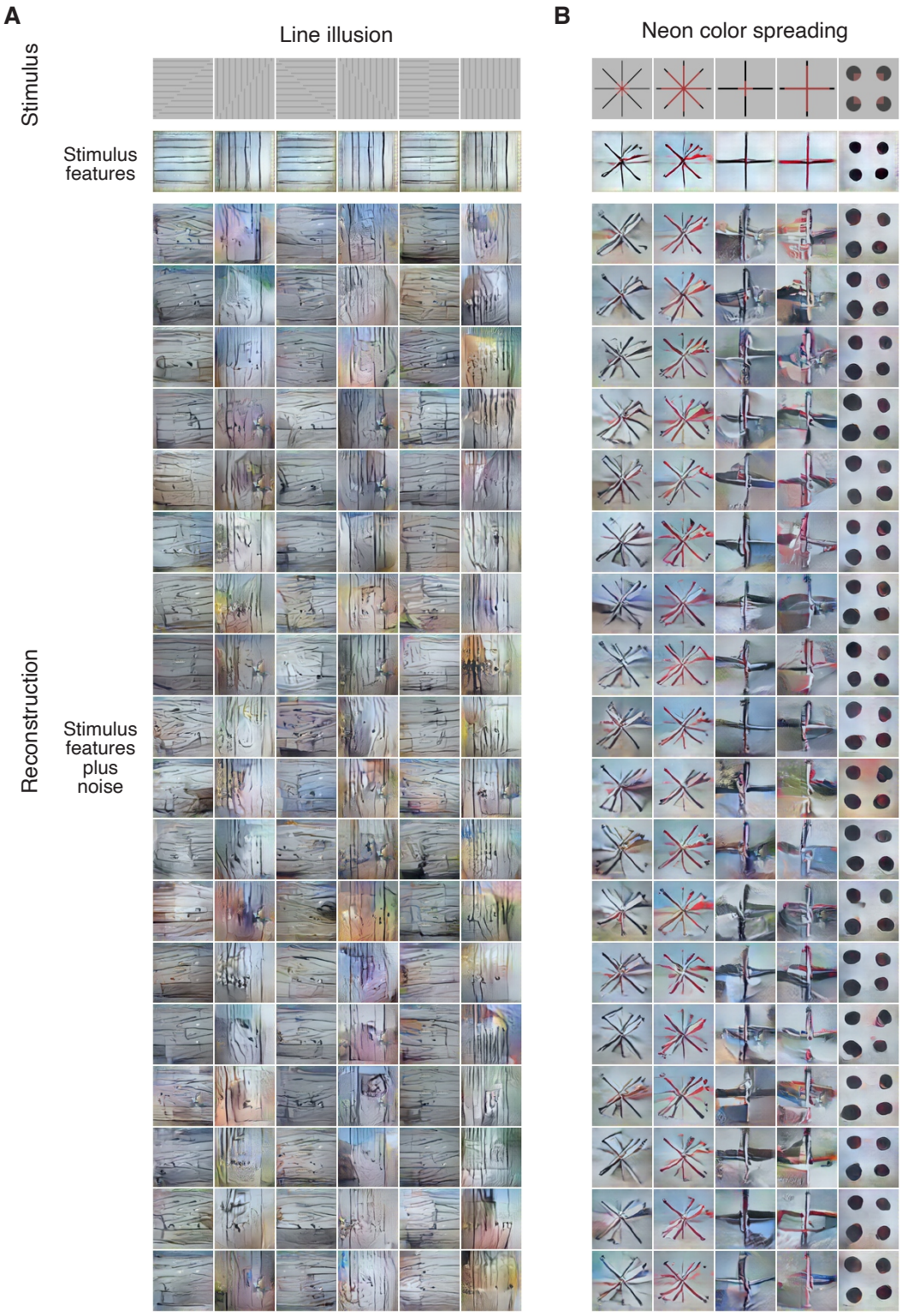

Fig. 5.3 Reconstructions of illusory images from stimulus features combined with noise. The noise factors were derived from the empirical distribution of noise, which was computed from the brain-decoded features associated with non-illusory images. (A) Line illusion. (B) Neon color spreading. Figure from Cheng et al. (2023).

Reconstructions from brain-decoded features (VC)

Based on the observations from the previous subsection, our investigation further delved into the reconstructions utilizing DNN features decoded from single-trial brain activity across the entire visual cortex (VC) corresponding to each stimulus image.

Regarding line illusions, the reconstructed images manifested components of both the illusory and inducer orientations (Figure 5.4). Notably, the illusory orientation frequently emerged more distinctly than the inducer orientation, a phenomenon not confined to the specific image area where the illusory line was perceived. Although reconstructions are similar between illusion and positive control conditions, the line components corresponding to the illusory orientation in the reconstructed positive control images tended to be more pronounced. Conversely, in the control condition, characterized by a reduced number of grating lines to diminish the illusion, the line components pertaining to the illusory orientation were less pronounced than the illusion condition (Figure 5.5). The positive control condition that corresponded to both illusion and control conditions showed reconstructions of clear central lines, indicating our methods can faithfully reconstruct central real lines.

As for the Ehrenstein configuration of neon color spreading, reconstructed images showcased an expanded colored area relative to the control, wherein a line-width gap was introduced to negate the illusory perception (Figure 5.6). As for the Varin configuration, the control image was engineered to inhibit color spreading without affecting the contour or shape aspect (Figure 5.7). Reconstructions in both the illusory and control conditions revealed a contour-like intensity profile; however, color spreading was significantly more evident in the illusory condition.

In the cases of both Ehrenstein and Varin configurations, the outer inducer segments generally suffered from inadequate reconstruction, possibly attributed to the selective emphasis on color regions typically situated more centrally. Additionally, the diminished resolution in peripheral representation might also have contributed to the less accurate reconstructions observed in the peripheral areas. The reconstructed color seemed weaker in the illusion condition compared to the positive control condition, especially for the large cross version of Ehrenstein configuration and the Varin configuration. Interestingly, these configurations often induce weaker perception of illusory color than other configurations.

5.3 Discussion

Before applying decoding models to illusory images, we first evaluated them using single-trial brain activity induced by natural images not included in the decoder training image sets. We also performed reconstruction analyses for these natural images using brain activity from

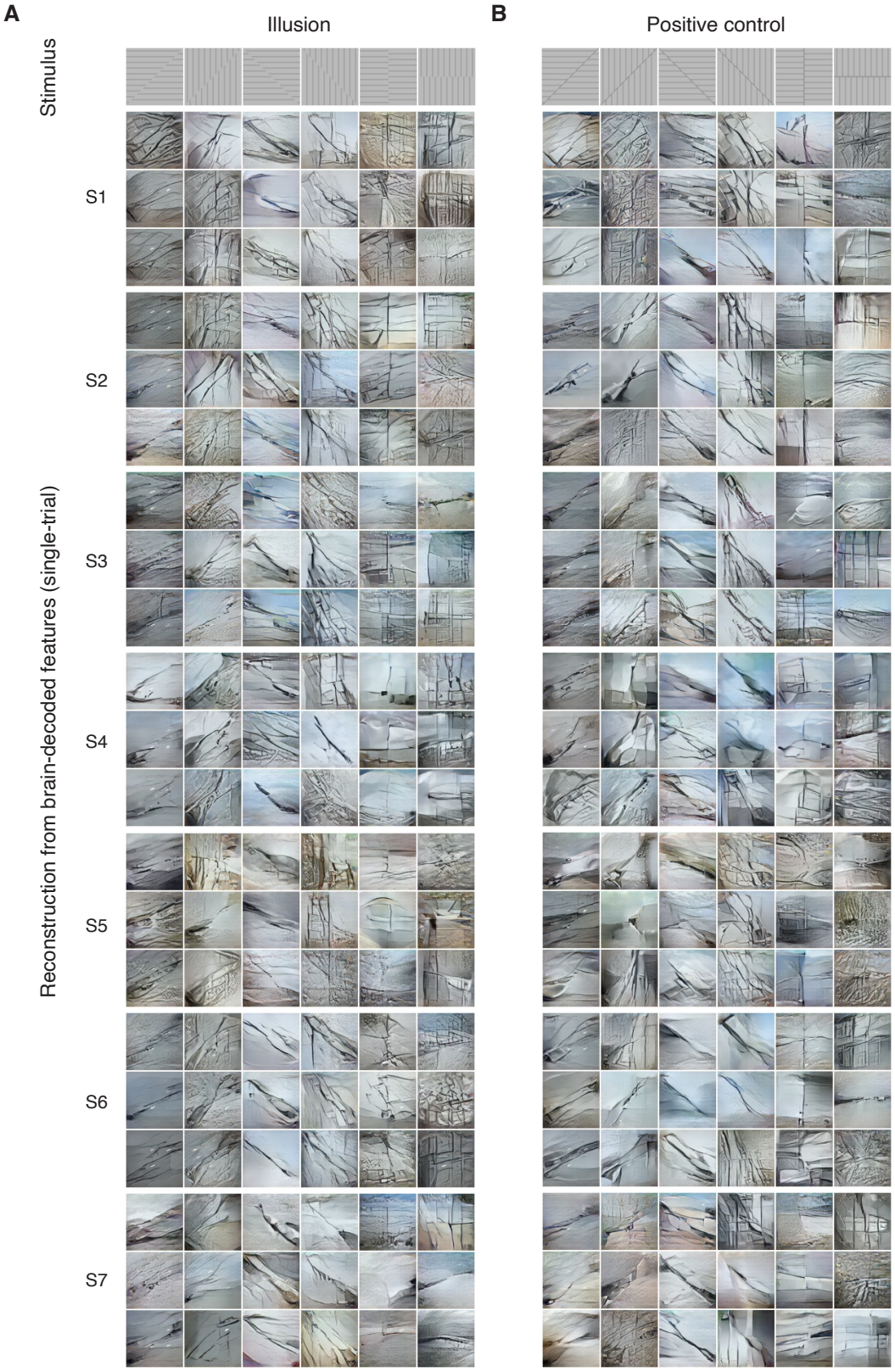

Fig. 5.4 Reconstructions of line illusion for various configurations from brain activity. Single-trial reconstructions were generated using fMRI signals from the whole visual cortex (VC). For each subject, the figure exhibits representative reconstructions from three independent trials, organized in a tripartite panel. (A) Illusion condition. (B) Positive control condition. Figure from Cheng et al. (2023).

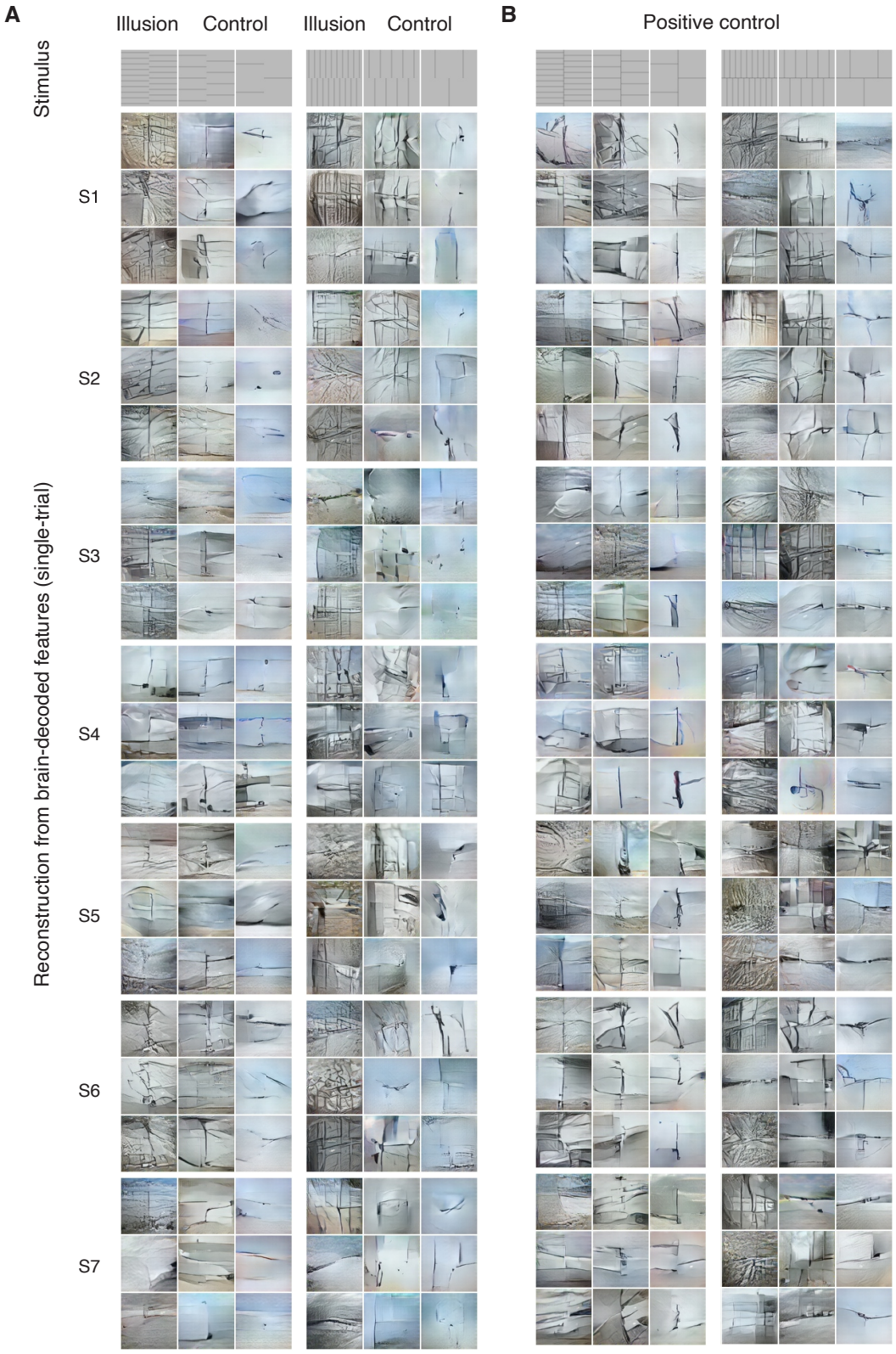

Fig. 5.5 Reconstructions of line illusion and controls from brain activity. Single-trial reconstructions were generated using fMRI signals from the whole visual cortex (VC). For each subject, the figure displays representative reconstructions from three independent trials, organized in a tripartite panel. Figure from Cheng et al. (2023).

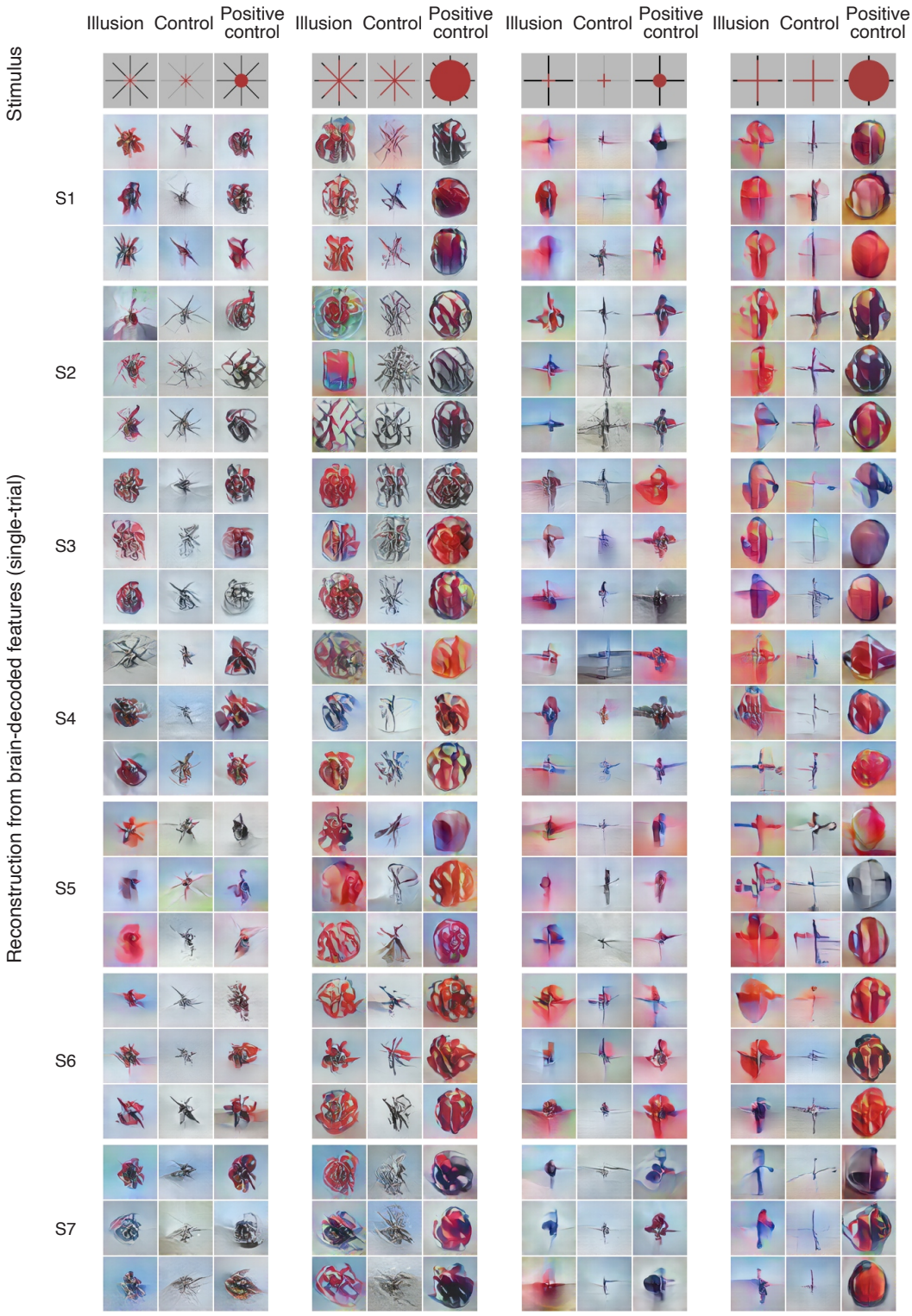

Fig. 5.6 Reconstructions of neon color spreading (Ehrenstein) from brain activity. Single-trial reconstructions were generated using fMRI signals from the entire visual cortex (VC). For each subject, the figure presents representative reconstructions from three separate trials, arranged in a tripartite panel. Each triad of columns depicts, from left to right, the illusion condition, the control condition, and the positive control condition, all corresponding to the same configuration. Figure from Cheng et al. (2023).

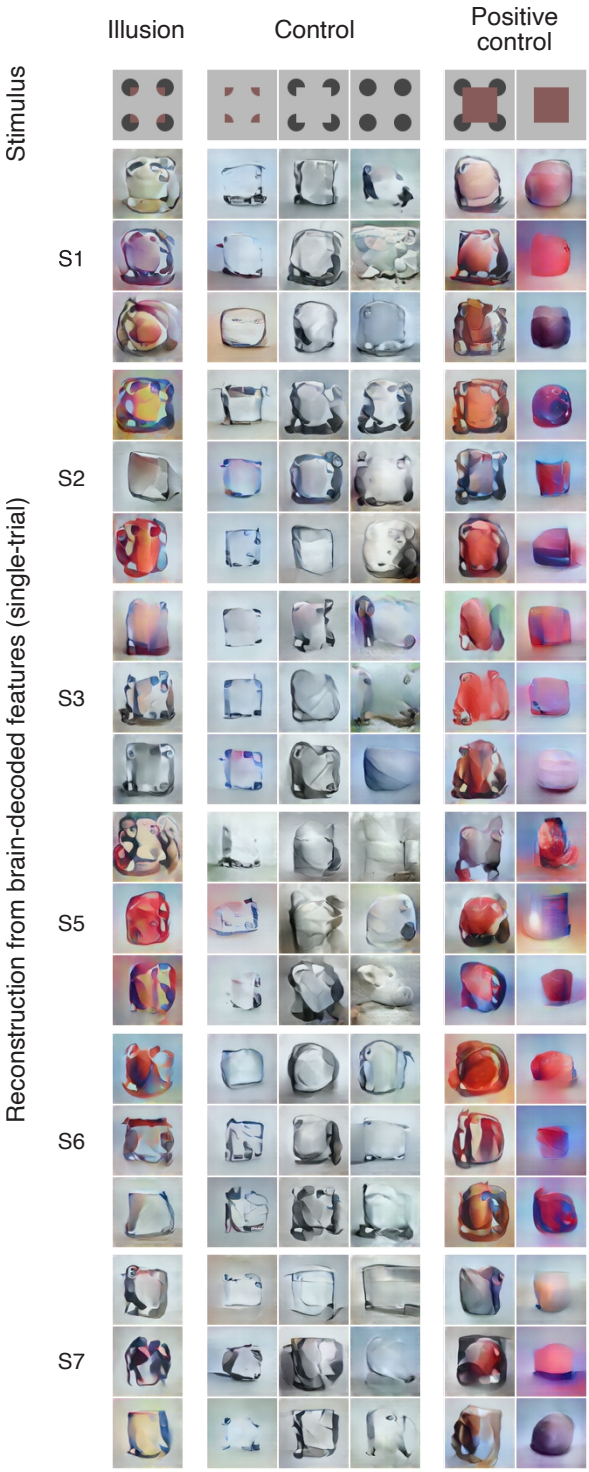

Fig. 5.7 Reconstructions of neon color spreading (Varin) from brain activity. Single-trial reconstructions were generated using fMRI signals from the entire visual cortex (VC). For each subject, the figure presents representative reconstructions from three separate trials, arranged in a tripartite panel. Each triad of columns depicts, from left to right, the illusion condition, the control condition, and the positive control condition, all corresponding to the same configuration. Figure from Cheng et al. (2023).

individual visual areas. These results provide a baseline for how the model works for each subject and brain area. The reconstructions from higher-level visual areas are surprisingly similar to the presented natural images with respect to some visual features and might reflect the unique coding schemes of different visual areas.

Our decoding and reconstruction analyses provide compelling evidence for the neural representations of illusory contours and neon color spreading in the visual cortex. While reconstructions revealed features consistent with our illusory experiences, reconstructions from DNN features directly extracted from illusory images, as well as decoded DNN features of control images did not contain illusory features, such as illusory contours or color.

The success of our reconstruction analysis demonstrates the effectiveness of the feature decoding-to-generator approach in reconstructing illusory percepts from brain activity patterns. By using multiple linear decoding models to map brain activity to DNN features and a pre-trained generator network to map the decoded features back to the pixel space, we were able to reconstruct images that closely matched the subjective experience of illusory contours and neon color spreading. The successful reconstruction of illusory percepts from single-trial brain activity highlights the robustness of our approach and the reliability of the neural representations of these illusions.

However, our decoding and reconstruction analyses also have some limitations and potential areas for improvement. The spatial resolution of reconstructions is not satisfying. Moreover, many inducer parts of the presented images were not accurately reflected in our reconstructions. Despite these limitations, our decoding and reconstruction analyses provide a powerful framework for investigating the neural representations of visual illusions.

Chapter 6

Examination of DNN and Generator Modules

In the previous chapter, we demonstrated the successful reconstruction of illusory contours and neon color spreading from brain activity patterns using a combination of linear regression models and pre-trained generator networks. We showed that the subjective experience of these illusions can be reliably reconstructed from single-trial fMRI data.

In this chapter, we examine the role of the deep neural network (DNN) and generator modules in our reconstruction pipeline. We begin by analyzing the responses of individual DNN units to illusory stimuli, investigating whether these units exhibit tuning properties that are consistent with the perception of illusory contours and neon color spreading. We describe the design of our analysis, including the types of DNNs examined, the unit selection methods, and the criteria for evaluating the tuning properties of the units. We present the results of our analysis, demonstrating that the DNN units do not inherently represent illusory features and that the reconstructed illusory percepts are not a result of the DNNs' pre-existing representations.

Next, we examine the robustness of our reconstruction results to the choice of generator module. We compare the reconstructions obtained using three different types of generators: generative adversarial networks (GANs), diffusion models, and pixel optimization. We discuss the advantages and limitations of each type of generator and evaluate the qualitative similarity of the reconstructions obtained using each method. We present the results of our analysis, demonstrating that the reconstructed illusory features are consistent across different generator modules, supporting the robustness of our approach. The content of this chapter is based on sections "RESULTS: Illusory stimuli and the reconstruction model" and "RESULTS: Reconstructed images", and "Supplementary Materials" of Cheng et al. (2023).

6.1 Analysis of DNN units' responses

6.1.1 Design of analysis

We investigated if DNN units responded to an illusory line similarly to a real line. We identified the units that were selective to a specific orientation and compared the tuning curves of these units with respect to the orientation between illusory and positive control images. An illusory image consisted of two patterns of concentric curves that induced the perception of a straight illusory line in the center (Figure 6.1). Since the spacing between adjacent concentric curves was fixed, each of the two patterns was determined by the phase of the innermost curve. The corresponding positive control image was made by adding a real control line at the same position as the illusory line.

For the color spreading, we identified units that were sensitive to real color in the image regions of interest (where illusory color was expected to be seen) and compared the unit activations between illusory, control, and positive control images (Figure 6.2). If the selected units also responded to illusory color, we would see higher activations in response to both illusory and positive control images than control images.

Type of DNN examined

The analyses of this thesis mainly utilized a variant of the AlexNet architecture, the BAIR/BVLC CaffeNet model, pre-trained on the ImageNet dataset to classify 1,000 object categories (`https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet`). Introduced in 2012 by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, AlexNet marked a major breakthrough in the ImageNet Large Scale Visual Recognition Challenge. This deep convolutional neural network significantly outperformed previous models in image classification tasks. AlexNet is a feed-forward neural network consisting of five convolutional layers followed by three fully connected layers and employs techniques such as ReLU activations, dropout, and data augmentation. Its success ignited a renaissance in deep learning, particularly in computer vision. Images were resized to 227×227 pixels before input into the CaffeNet model. The outputs of the first seven layers (conv1–5, fc6, and fc7; post-rectification) were reshaped into vectors for each visual stimulus. The unit counts for the CaffeNet layers are: conv1, 209400; conv2, 186624; conv3 and conv4, 64896; conv5, 43264; fc6 and fc7, 4096.

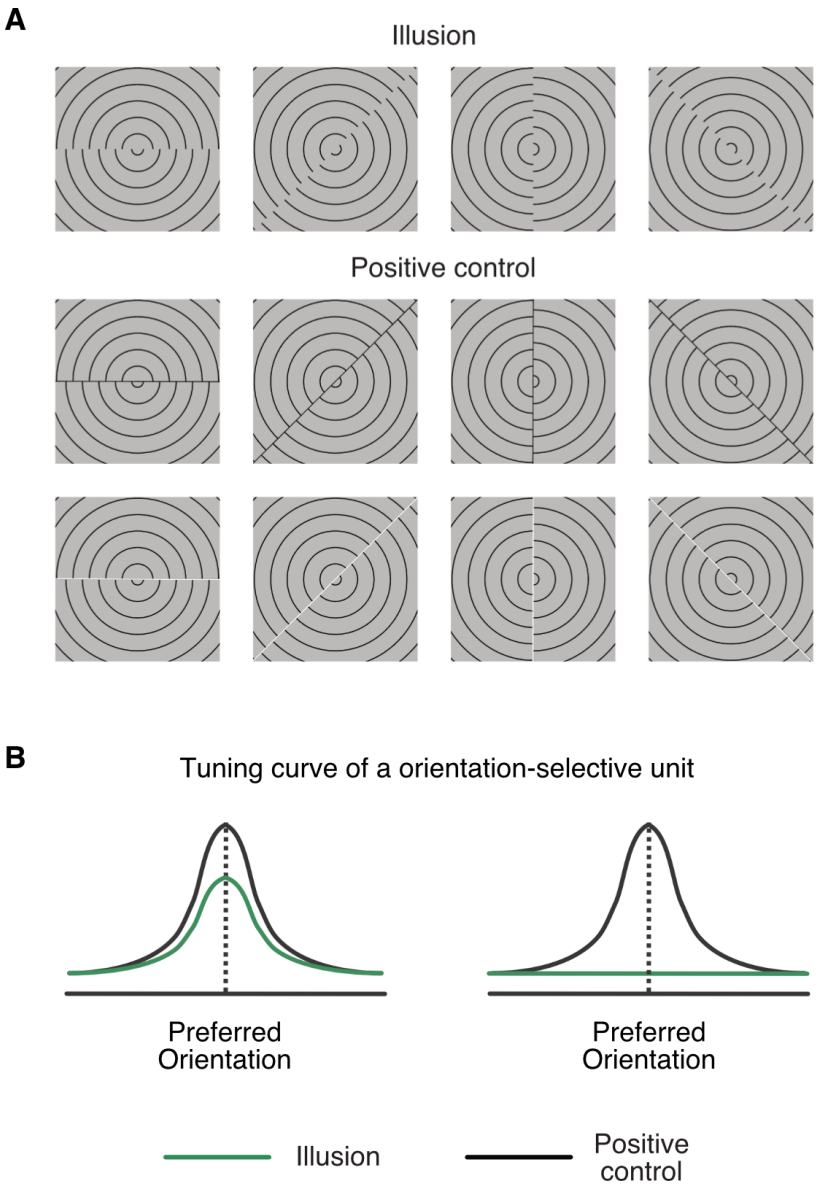

Fig. 6.1 Illusory line stimuli and predictions. (A) Example images of the Illusion (top), positive control with black central line (middle), positive control with white central line (bottom). From left to right, the orientations of the induced illusory line are 0, 45, 90, 135 degrees. (B) Predictions of the two conditions: if a DNN unit responds similarly to illusory line and real line orientations (left) or if a DNN unit does not respond to illusory line similarly to real line (right).

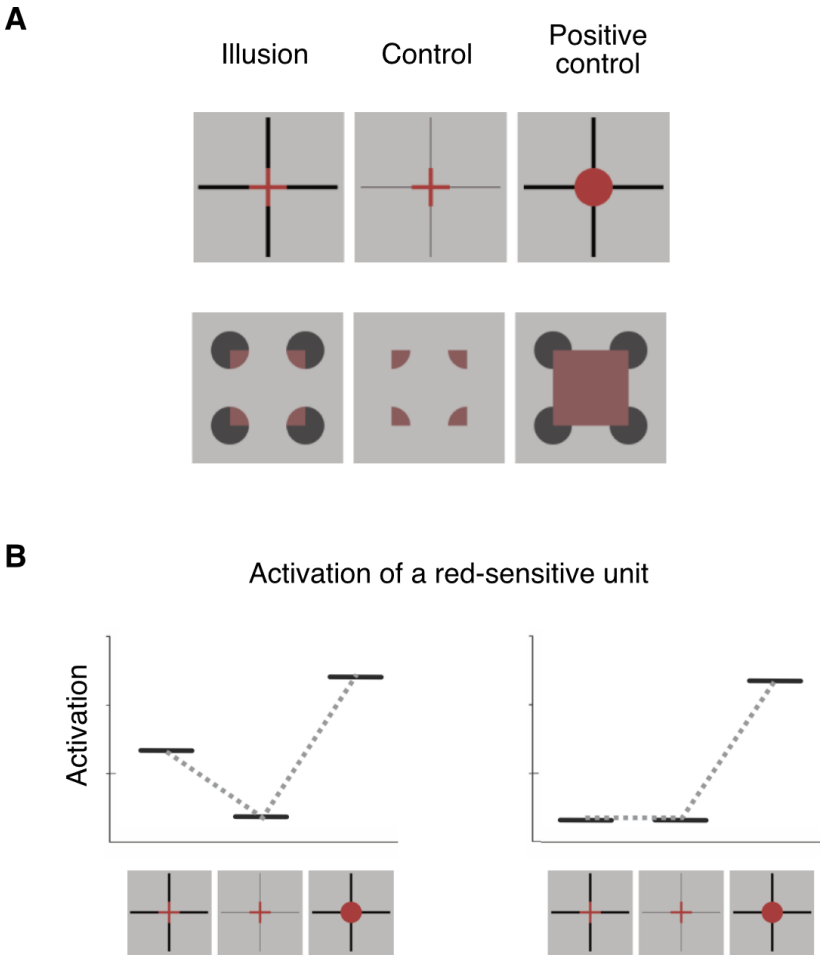

Fig. 6.2 Illusory color stimuli and predictions. (A) Example images of the illusion (top), control (middle), and positive control (bottom) conditions for neon color spreading: the Ehrenstein (middle) and Varin (bottom) configurations. (B) Predictions for two scenarios: a red-sensitive DNN unit responds similarly to illusory and real colors (left) or a DNN unit does not respond to illusory colors (right).

6.1.2 Unit selection methods

The procedure of unit selection conditioning on a specific pair of phases is as follows: first, we generated 24 images that comprised a real central line of 12 different orientations (from 0 to 180 degrees in a step of 15 degrees) and two types of backgrounds with concentric circles corresponding to each of the phases; second, from units that respond to the image center ("center-responsive units"), we selected those with maximum positive activation at the preferred orientation (0, 45, 90, or 135 degrees) for both backgrounds; third, we ranked units based on how much higher the activation of the preferred orientation was than the average activation of the other orientations. In the verification stage, we visualized the tuning curves of the top 5% center-responsive units in each layer, when positive control images of different orientations were used as the input of the DNN (Figure 6.3). The unit activation to a positive control image was averaged from activations to the original image and the swapped version (swapping two concentric patterns while fixing the central line). We randomly generated 50 pairs of phases and repeated the above selection and test procedure 50 times. Note that different sets of units could be selected and visualized depending on the phases of the inducing patterns.

It is worth noting that DNN units – especially in higher layers – may be broadly tuned to features other than orientation or have a large receptive field covering inducer lines. The variation in unit activations may not solely depend on the difference in the orientation of the central line. The following settings were used to address these problems. First, with the concentric inducing pattern in illusory images, the orientation of inducer lines and the orientation difference between illusory and inducer lines remained unchanged while the orientation of the central line varied. Second, the orientation selectivity was measured in the presence of concentric patterns similar to those in illusory images, instead of using images with only a single central line. Third, the concentric curves were made continuous (circle) to avoid picking units responding to terminators of inducer lines. Fourth, to avoid picking position-selective units, the range was restricted to center-responsive units: the center unit of each channel in convolutional layers (except conv1; $3 \times 3 = 9$ units in the channel center) and all units in fully connected layers.

The unit selection was performed for Ehrenstein and Varin, respectively (Figure 6.4A). First, we generated 24 selection-stage images, which comprised of a disk or square surface (image region of interest) with uniform color of three levels of luminance (0.3, 0.5, 1; relative to the luminance of gray and based on measurements of the display in the scanner) and four levels of saturations (0, 0.3, 0.7, 1) in two different configurations (same as that in illusion or control; thick or thin black lines for Ehrenstein, and with or without Pacman for Varin). The surface was red when the saturation was positive and gray when the saturation

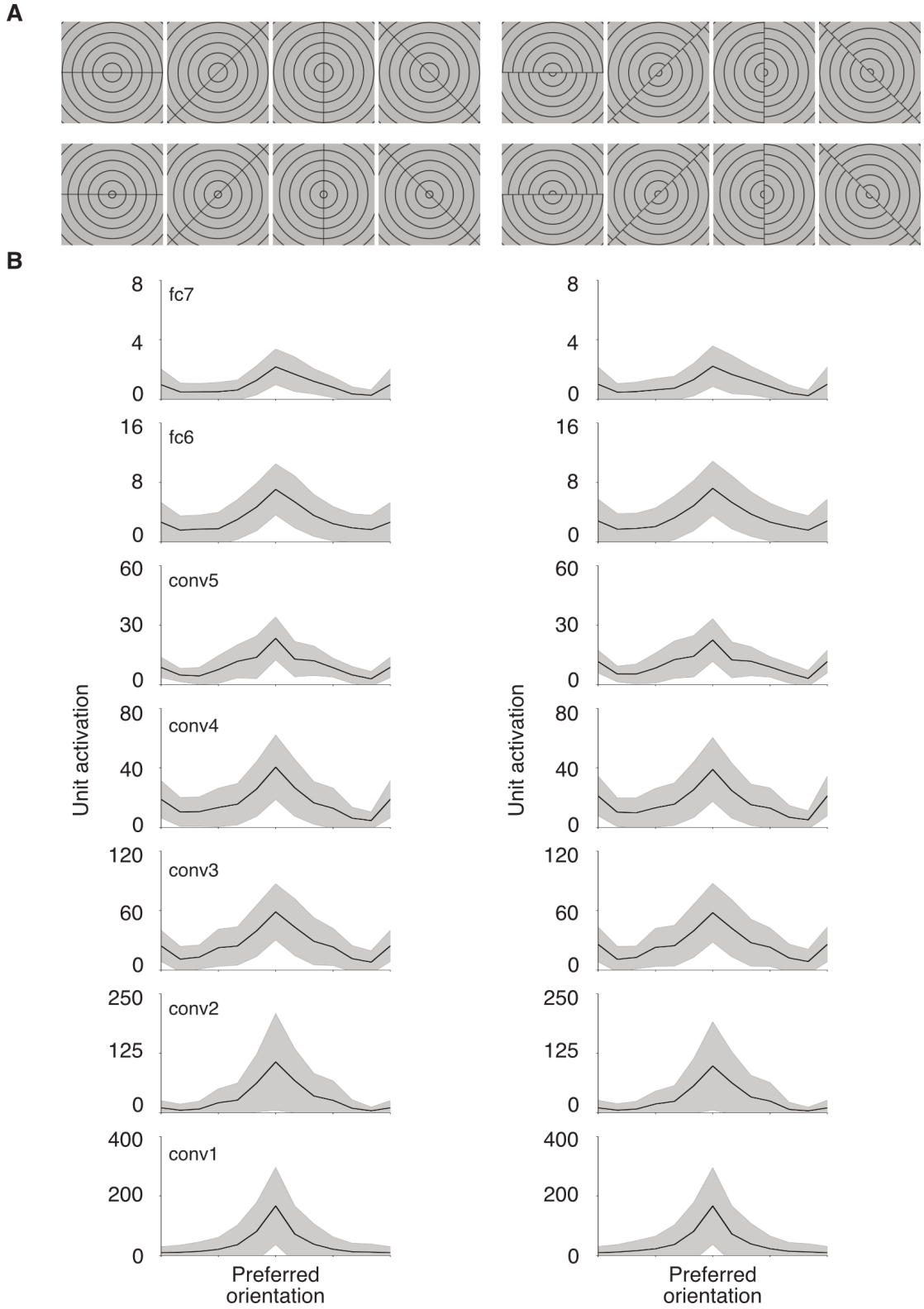

Fig. 6.3 Verification of orientation-selective units. (A) Example images used to identify (left) and verify (right) orientation-selective units, respectively. The two rows in the left or right panel constitute a pair of background phases. (B) Selected orientation-selective units have higher activations to higher color saturations. Black lines represent the median value and shaded areas represent the interquartile range of units.

was equal to zero. Then, we selected units that always had higher activations to the red surface than the gray surface of the same luminance, even when the configurations were different. We took the intersection among the sets of selected units from different levels of luminance and saturation, because it was uncertain at which level of luminance and saturation DNN would represent the illusory color if any. We further ranked the selected units by the average activation difference between the red and gray across different levels of luminance and saturation.

The activations of the top 5% units in each layer were visualized after normalization by subtracting the averaged activation from the images with a gray surface (Figure 6.4B). To confirm the success of unit selection, we expanded the set of selection-stage images with more saturation levels (from 0 to 1 in a step of 0.1) and plotted the unit activations as the saturation value increased. For each saturation level, we averaged the unit activation across three luminance levels and two types of configurations.

6.1.3 Results

We visualized the tuning curves of the top 5% center-responsive units in each layer, when illusory or positive control images of different orientations were used as the input of the DNN (Figure 6.5). The unit activation to an illusory or positive control image was averaged from activations to the original image and the swapped version (swapping two concentric patterns while fixing the central line). Each unit's tuning curve was normalized by subtracting the minimum activation value across all orientations. For the positive control images, the tuning curves exhibited distinct peaks at the units' preferred orientation, while the illusory images did not elicit such sharp peaks.

6.1.4 Results

As for illusory color, the activations of the top 5% units in each layer were visualized after normalization by subtracting the averaged activation from the images with a gray surface (Figure 6.6). The nearly identical activations for illusion and control conditions suggest that the units do not respond to illusory color. This minimal difference cannot be attributed to the units' lack of sensitivity, as most units demonstrated either equal activation for control and illusion conditions or slightly higher activation for the control condition, rather than the reverse.

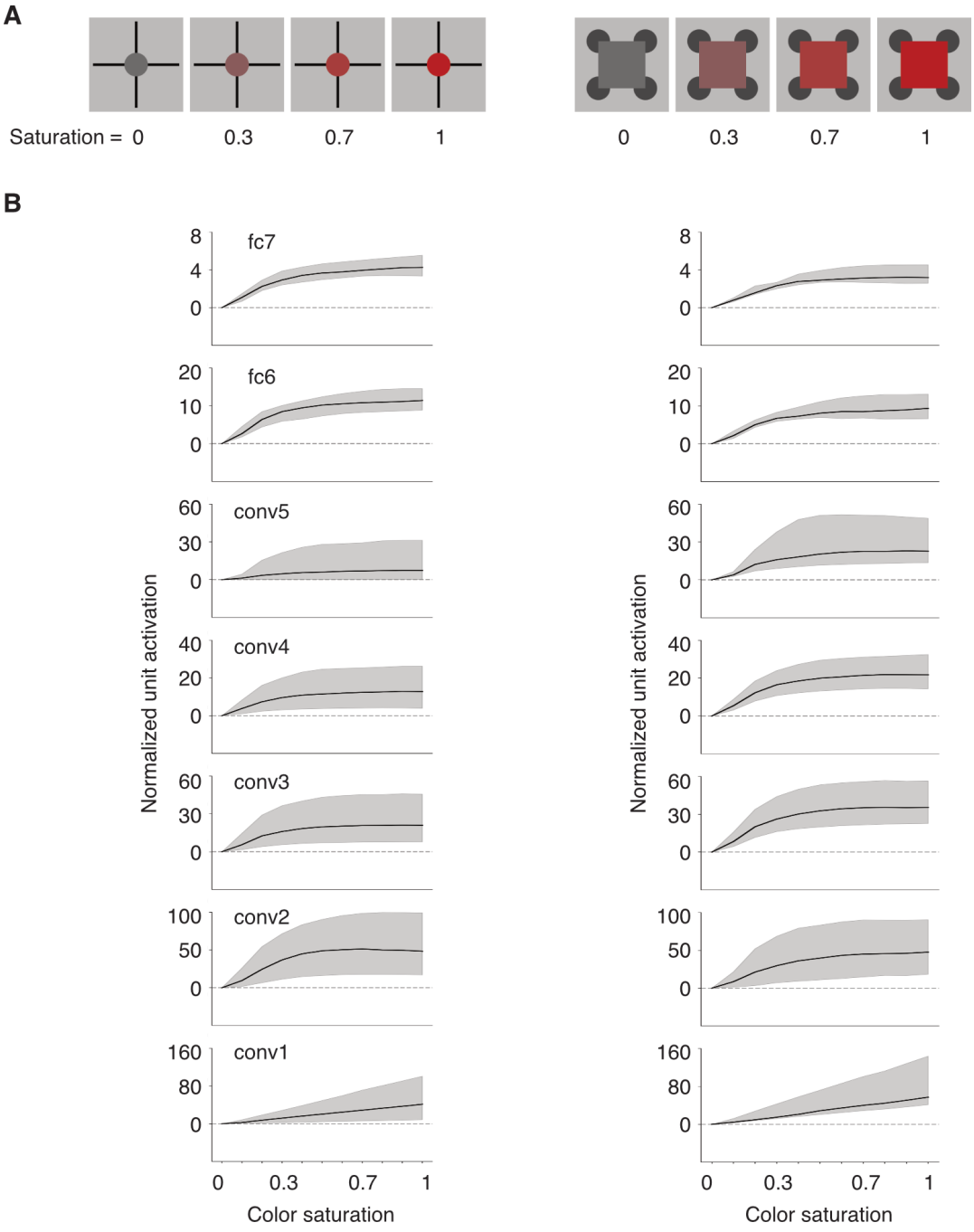

Fig. 6.4 Verification of red-sensitive units. (A) Example images used to identify red-sensitive units. Units were analyzed for Ehrenstein (left) and Varin (right), respectively. (B) Selected color-sensitive units have higher activations to higher color saturations. Black lines represent the median value and shaded areas represent the interquartile range of units.

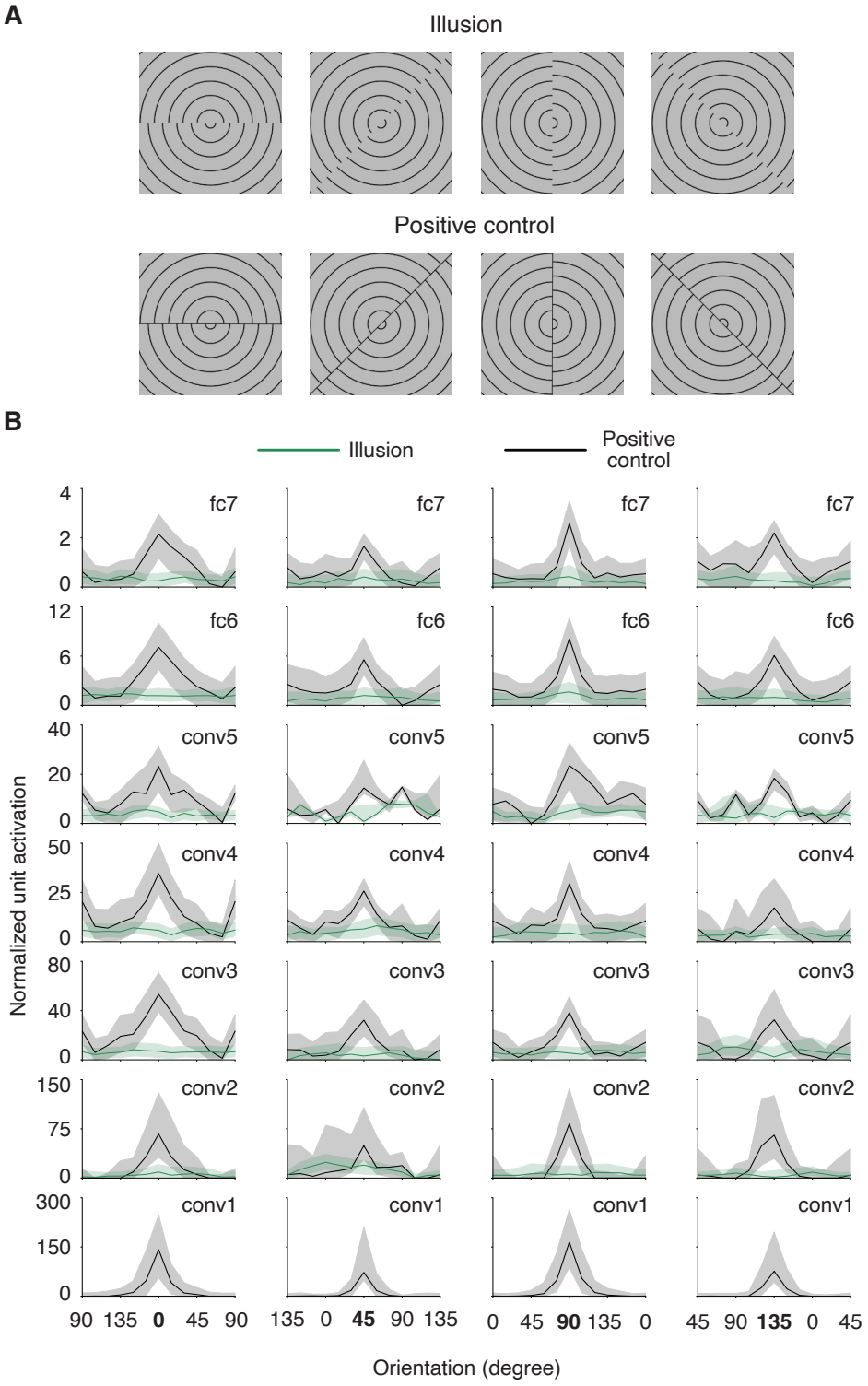

Fig. 6.5 Tuning curves of orientation-selective units. Each unit's tuning curve was normalized by subtracting the minimum activation value across all orientations. Lines represent the median activation, and shaded areas represent the interquartile range of the units pooled across different background phases. The tuning curves exhibited sharp peaks at the units' preferred orientation for the positive control images (black) but not for the illusory images (green). Figure from Cheng et al. (2023).

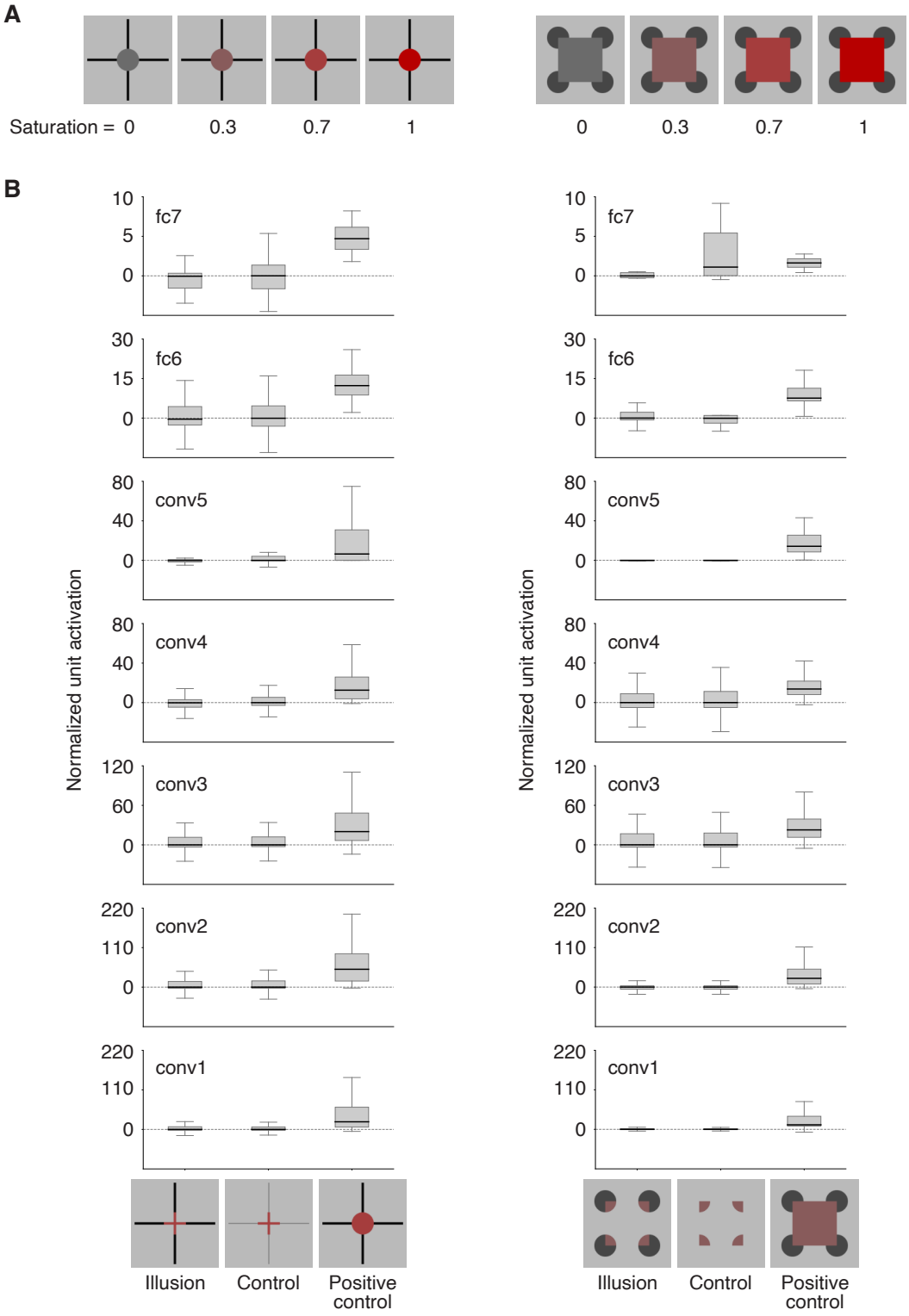

Fig. 6.6 Normalized responses of units selective for color. Median values are represented by black lines, and interquartile ranges are shown as shaded areas. If units were responsive to illusory color, similar activation levels would be expected for illusion and positive control images, exceeding the control condition. However, the near-identical activation for illusion and control suggests that units do not respond to illusory color. This subtle difference was not attributable to unit insensitivity, as most units exhibited either identical activation for control and illusion or higher activation for control. Figure from Cheng et al. (2023).

6.2 Robustness to the choice of generator

In the context of my research project or study, I primarily employed GAN as the generator for reconstruction. This choice was made before the data collection described in Chapter 4. In addition to the primary GAN-based generator, I expanded my exploration to include generators based on two other techniques: diffusion methods and pixel optimization (Figure 6.7). This indicates a comprehensive approach to examining different generative models to make sure that the observations can be qualitatively replicated.

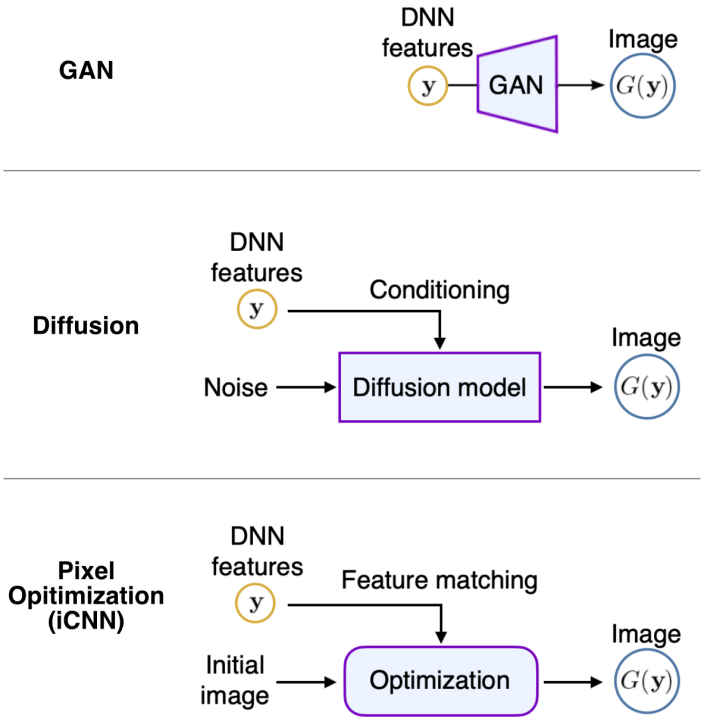

Fig. 6.7 Illustration of image generation with DNN features as the input. Three different generators are shown: Generative Adversarial Network (GAN) generator (top), conditional diffusion model (middle), and pixel optimization (bottom). Both GAN and diffusion models necessitate pre-training on extensive natural image datasets.

6.2.1 Image generator

Diffusion

Unlike GANs that use an adversarial approach, diffusion models are based on a different principle, inspired by the physical process of diffusion. Diffusion models have gained prominence for their ability to produce high-quality, diverse samples and have shown remarkable success,

particularly in generating complex data such as images and audio. The core idea behind diffusion models is derived from non-equilibrium thermodynamics, particularly the concept of diffusion, which is a process that seeks to reduce concentration gradients (differences) within a system. In the context of machine learning, diffusion is used as an analogy for gradually adding noise to the data until it turns into a completely random distribution. The model then learns to reverse this process, effectively denoising to generate samples from the learned data distribution. A diffusion model is typically trained in two phases. In the forward phase, it adds noise to the data step by step over several iterations, gradually transforming the data into a Gaussian distribution. In the reverse phase, the model learns to generate data by reversing the noise process, starting from random noise and sequentially reducing the noise to construct data samples. This reverse process is where the generation of new, synthetic samples occurs. Diffusion models consist of two main components: the forward diffusion process and the reverse diffusion (or denoising) process. The forward process is usually fixed and known, often a simple Markov chain that adds Gaussian noise at each step. The reverse process, which is learned, is more complex and aims to model the distribution of the data at each step of the reverse chain. In this research, we adapted and trained a conditional diffusion model, modifying the architectural framework based on a preceding study (Dhariwal and Nichol, 2021). Specifically, the model's training was directed towards reconstructing original input images, predicated on the rectified outputs from the fc6 layer of the CaffeNet model. We utilized the structural design of the class-conditional ImageNet 64x64 model, as delineated in the available repository (https://github.com/openai/improved-diffusion/), which incorporates a conditioning vector of 768 dimensions. To accommodate the 4096-dimensional fc6 feature vector extracted from CaffeNet, as discussed in the "DNN image features" section, we integrated an additional fully-connected layer with input and output dimensions of 4096 and 768, respectively. This layer was instrumental in channeling the fc6 features into the diffusion model. For the training corpus, approximately 1.2 million natural images sourced from ImageNet (45) were employed, each resized to a uniform 64x64 pixel dimension. The training was conducted under a linear noise schedule, with a total of 4000 diffusion steps established. The model underwent an extensive training regime, spanning 1 million steps, with a batch size of 128 and a learning rate set at 0.0001. This meticulous training approach was aimed at effectively conditioning the diffusion model to generate high-fidelity images based on the intricate fc6 feature mappings.

Pixel optimization (iCNN)

In deep image reconstruction research (Shen et al., 2019a), the authors employed this technique to transform decoded features from multiple layers of a Deep Neural Network

(DNN) into a coherent image, with the relevant code accessible via https://github.com /KamitaniLab/DeepImageReconstruction. The approach involved optimizing the pixel values of an input image to align its image features with the corresponding decoded features, effectively facilitating the accurate reconstruction of visual data. With the decoded features, the reconstruction process starts with an initial image. This image is typically random or a grayscale to begin with and serves as the starting point for the iterative process. The core of the iterative reconstruction is an optimization loop. In this loop, the initially generated image is input into a pre-trained deep neural network, which extracts the image features at multiple layers. These features are then compared to the decoded features from the brain activity. The difference between the two sets of features forms the basis for optimizing the image. Using backpropagation, the pixel values of the generated image are adjusted to reduce the difference, effectively making the image's features more similar to the decoded brain features. To ensure that the reconstructed images are not only feature-accurate but also visually recognizable and noise-free, regularization techniques are often applied. This might involve imposing natural image priors or other constraints to guide the optimization toward more realistic images.

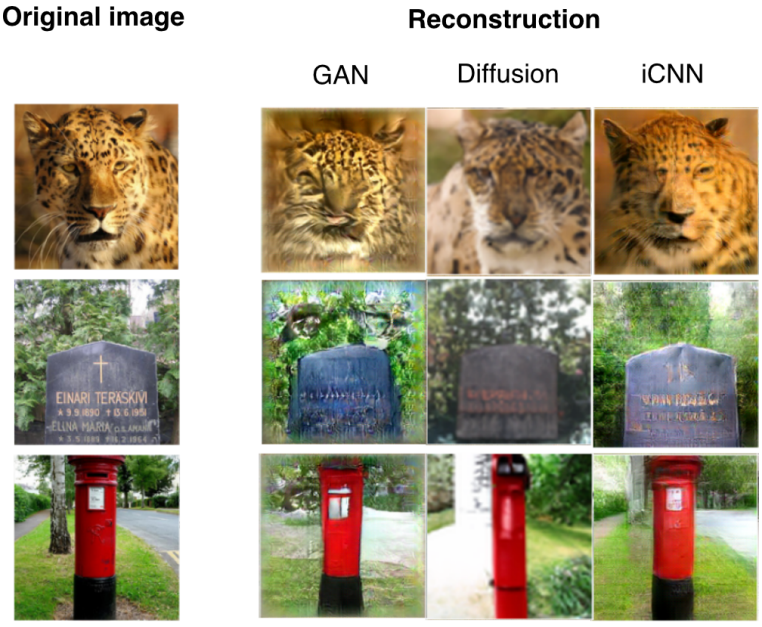

Fig. 6.8 Testing generators with DNN features of original images. Different generators produced reconstructions similar to the original images with different styles.

6.2.2 Results

While the three types of generators—GAN, diffusion methods, and pixel optimization—each produced images with their own unique characteristics or "flavors," they all provided qualitatively similar results regarding the visual features this study was interested in (Figure 6.9 and 6.10). Similar to the results obtained using GAN, reconstructions from stimulus features This suggests that despite the operational differences between the models, they were comparably effective at generating or reconstructing images. Essentially, the illusory features that were consistent with our subjective experiences in reconstructions from brain activity were robust among reconstruction methods.

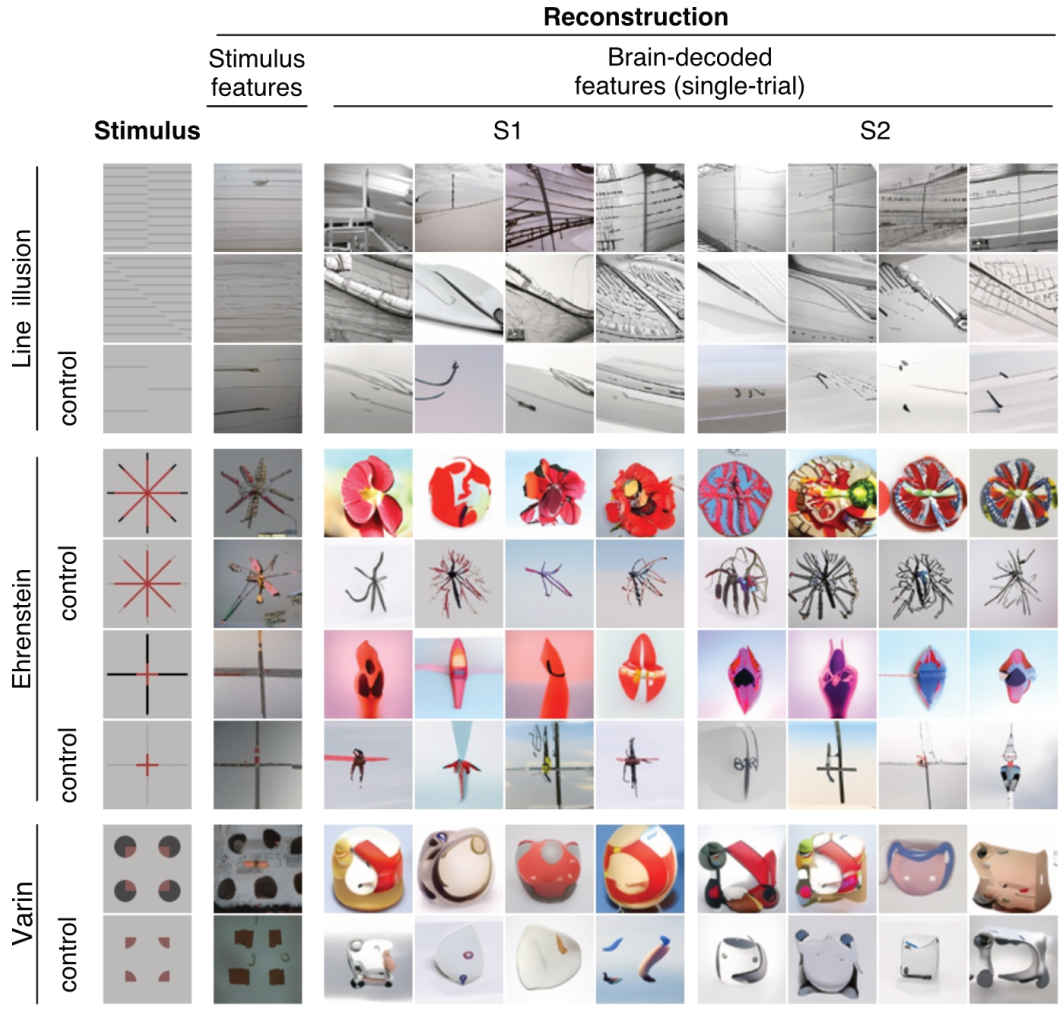

Fig. 6.9 Single-trial reconstructions of illusory and control images using a conditional diffusion model for two representative participants (S1, S2). Reconstructions are shown for both stimulus features and brain-decoded features obtained from fMRI signals in the entire visual cortex (VC). Figure from Cheng et al. (2023).

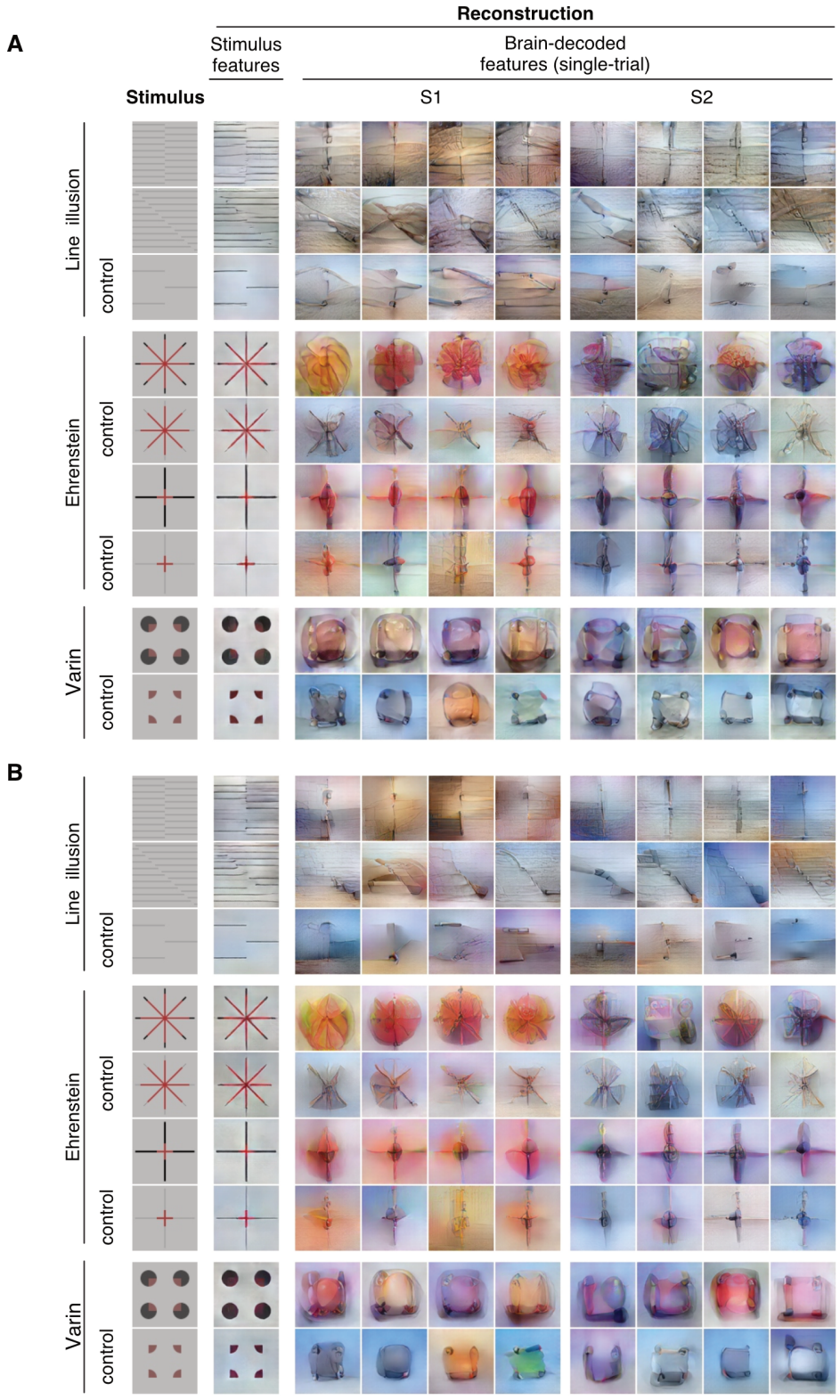

Fig. 6.10 Single-trial reconstructions of illusory and control images using pixel optimization (iCNN) for two representative participants (S1, S2). Reconstructions are shown for both stimulus features and brain-decoded features obtained from the same fMRI trials as in Figure 6.4, focusing on the whole visual cortex (VC). (A) CaffeNet. (B) VGG19. Panel A from Cheng et al. (2023).

6.3 Discussion

By examining the responses of individual DNN units to illusory stimuli, we demonstrate that the reconstructed illusory features are not a result of the DNNs' pre-existing representations but rather reflect the neural representations of these illusions in the visual cortex.

The lack of activations to illusory features in the DNN units suggests that the DNNs used in our study do not inherently represent illusory features. This finding is consistent with previous studies showing that standard feedforward DNNs trained on object recognition tasks do not exhibit the same perceptual biases and illusions as humans (Fan and Zeng, 2023). However, it is important to note that our analysis only examined a limited set of DNNs and that other DNNs with different architectures or training regimes may exhibit different properties. The extension of our analysis to a broader range of DNNs, including those specifically designed to capture perceptual phenomena, is an interesting direction for future research.

The robustness of our reconstruction results to the choice of generator module highlights the generalizability of our approach. By comparing the reconstructions obtained using GANs, diffusion models, and pixel optimization, we demonstrate that the reconstructed illusory features are consistent across different generative models, despite differences in their training procedures and architectures. This finding suggests that the success of our approach is not dependent on the specific choice of generator module but rather on the effectiveness of the linear decoding models in mapping brain activity to DNN features.

However, it is important to note that the choice of generator module can still have a significant impact on the quality and style of the reconstructed images. While the three types of generators used in our study produced qualitatively similar results regarding the reconstructed illusory features, they each had their own unique characteristics and limitations. GANs are known for their ability to generate high-quality and diverse images but can suffer from mode collapse and training instability (Goodfellow et al., 2014). Diffusion models have shown remarkable success in generating complex and realistic images but can be computationally expensive and require large amounts of training data (J. Ho et al., 2020). Diffusion models also tend to remember the training data and bring unwanted artifacts to reconstructions. Pixel optimization methods, such as the iCNN approach used in our study, can directly optimize the pixel values of the reconstructed image but may be sensitive to the weights of different layers, which have varying representational correspondences with visual areas (Shen et al., 2019a).

In conclusion, our analysis of the DNN and generator modules provides a deeper understanding of the mechanisms underlying the successful reconstruction of illusory contours and neon color spreading from brain activity patterns. By demonstrating the lack of illusory

tuning properties in the DNN units and the robustness of our reconstruction results to the choice of generator module, we strengthen the validity and generalizability of our approach. The insights gained from this analysis can inform the development of more advanced and specialized models for visual image reconstruction and the investigation of other types of perceptual phenomena using neuroimaging and computational techniques.

Chapter 7

Reconstruction of illusory contour along the visual hierarchy

This chapter explores the reconstruction of illusory contours, specifically line illusions, from brain activity patterns across different regions of the visual cortex hierarchy. By harnessing fMRI data from early visual areas like V1-V4 as well as higher-order areas such as the lateral occipital complex (LOC), fusiform face area (FFA), and parahippocampal place area (PPA), we aim to elucidate how the perception of illusory contours is represented and processed along the visual hierarchy. The content of this chapter is based on sections "RESULTS: Quantitative analyses of illusory lines across multiple brain areas" and "Supplementary Materials" of Cheng et al. (2023).

7.1 Reconstruction

Reconstructed images can be obtained by harnessing fMRI activity data from individual visual areas within the visual cortex, specifically V1-V4, LOC, FFA, and PPA , as detailed in section 4.3. Representative reconstructions from these areas for illusory contour stimuli are presented. In Figure 7.1, the results indicate that the earlier visual areas, namely V1–V3, generally provide accurate reconstructions of both illusory and inducer lines. On the other hand, V4 and the higher visual areas tend to exhibit less precise localization of illusory lines and a diminished quality in the reconstruction of inducer lines. These findings underscore the varying capabilities and roles of different visual cortical areas in the representation of illusion stimuli.

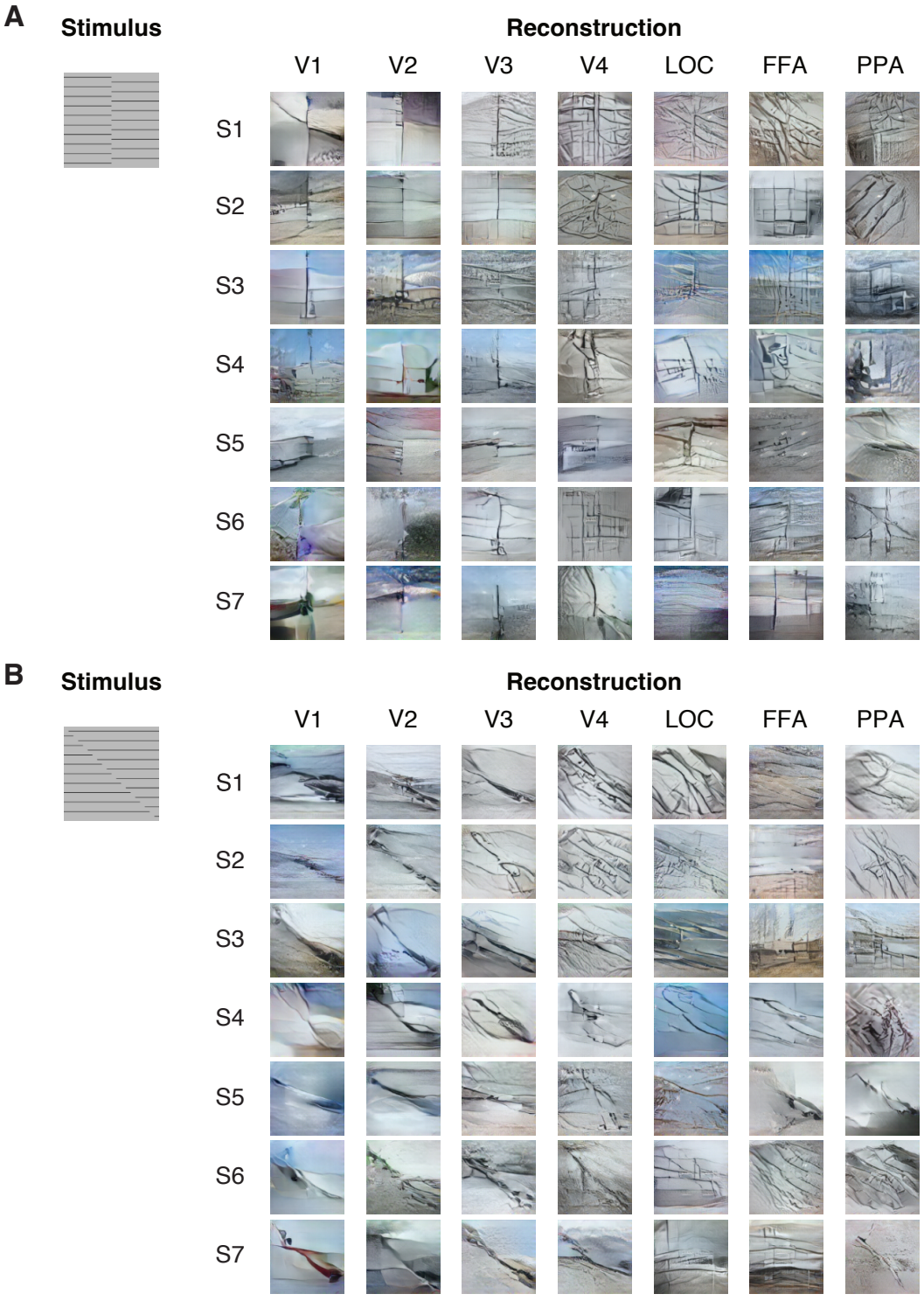

Fig. 7.1 Single-trial reconstructions of line illusion from different visual areas for each subject (using different trials from Figure 3). (A) 90°-difference configuration. (B) 45°-difference configuration. Figure from Cheng et al. (2023).

7.2 Evaluating Subjective Components in Reconstructions

To quantitatively evaluate the subjective components of illusory contours in the reconstructions, we developed a method to detect the principal orientation within reconstructed images. This allows us to determine if the prominent orientation matches the illusory contour orientation or the inducer orientation, providing a way to assess the representation of the illusion across visual areas.

7.2.1 Method

To quantify the reconstructed illusory lines, the principal line orientation within each individual trial reconstruction was determined employing the Radon transform (Radon, 1917). This method involved calculating the Radon projections across the line regions bisecting the image's center for every possible orientation. A distinct line presence would induce a notable fluctuation in the projection values along parallel line regions corresponding to the line's specific orientation. Consequently, the orientation exhibiting the maximal variance was designated as the principal orientation for each reconstruction. This methodological approach allowed for a systematic and quantitative analysis of the line orientations within the reconstructed illusory lines. In our study, each single-trial reconstructed image of the line illusion was subjected to a comprehensive evaluation, both on a global scale encompassing the entire image and on a local scale focusing on designated image regions.

To facilitate a detailed comparison, we identified the principal orientation for each image or specified region of interest (Figure 7.2). Our objective was to ascertain the greater resemblance between the principal orientation and either the illusory or inducer orientations, employing cosine similarity as our comparative measure. The principal orientation was pinpointed using a technique commonly applied in texture analysis (Jafari-Khouzani and Soltanian-Zadeh, 2005). This involved converting the images to grayscale followed by the application of a Radon transform to discern linear trends. Specifically, we selected the largest disk area, denoted as A, in the image region and projected it onto a line space. This projection was accomplished by summing the pixel intensities along each line within A:

$$R(r,\theta) = \sum_{(x,y)\in A} I(x,y)\delta(r - x\cos\theta - y\sin\theta) \tag{7.1}$$

where each line is parameterized by its distance from the center and orientation. The intensity of a pixel at a given location is represented by $I(x,y)$, and it contributes to the sum if it lies on the line being considered.

For each orientation, the variance of the projections was computed across lines intersecting a small, centrally located disk region within the image. The rationale is that a pronounced black line at a specific orientation would notably reduce the projection value, leading to a marked variance in projections among neighboring lines of the same orientation. By calculating this variance across various orientations, we were able to determine the principal orientation as the one exhibiting the largest variance. To ensure precision, only lines within a close proximity (no more than five pixels in distance) to the region's center were considered for variance calculations, as this restricted range was anticipated to include the illusory line. Through this meticulous approach, we aimed to accurately delineate and analyze the illusory and non-illusory regions within the reconstructed images.

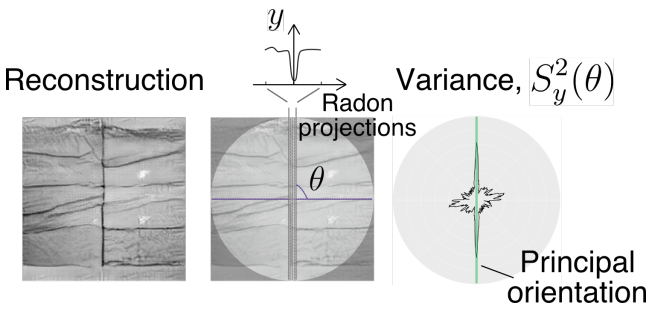

Fig. 7.2 Identification of principal orientation using Radon transform. The orientation exhibiting the highest variance in Radon projections across line positions was designated as the principal orientation within an image. Figure from Cheng et al. (2023).

In this investigation, each participant was considered as an independent replication unit, prompting the majority of statistical analyses to be executed on an individual subject basis. Prior to initiating the experimental tests, we predetermined the sample size of the test data, specifying the number of trials per image/condition.

For the analysis of principal orientations, we conducted one-sided z tests to ascertain if the proportion closer to illusory was significantly higher in one condition compared to another. To attain a statistical power of 80% and detect a substantial effect size (Cohen's $h = 0.8$) at a 0.05 significance level, a minimum of 20 samples for each condition was necessary. Consequently, we conducted 20 trials for every stimulus image per participant. The results presented in this thesis reflect the aggregated statistical tests across different stimulus configurations, meaning each condition for statistical comparison comprised more than 20 trials. Specifically, the number of trials that met the inclusion criteria for comparing local illusory versus non-illusory regions, the counts were 234, 232, 240, 240, 226, 240, 240 for subjects S1–S7 respectively.

7.2.2 Results

Distribution of principal orientations

The principal orientations were aggregated by combining single-trial reconstructions from the whole visual cortex (VC) for stimulus images, specifically those with a 90° (45°) difference between illusory and inducer orientations, across 275 (551) trials from seven subjects that met the exclusion criteria (Figure 7.3). For 90°- difference configurations, the orientation distribution exhibited a bimodal peak corresponding to the illusory and inducer orientations, with a notable 61.1% of principal orientations aligning more closely with the illusory orientation. Similarly, configurations with a 45° difference also demonstrated a high proportion of orientations closer to the illusory ones. The results of individual subjects were consistent with that of the pooling results (Figure 7.4).

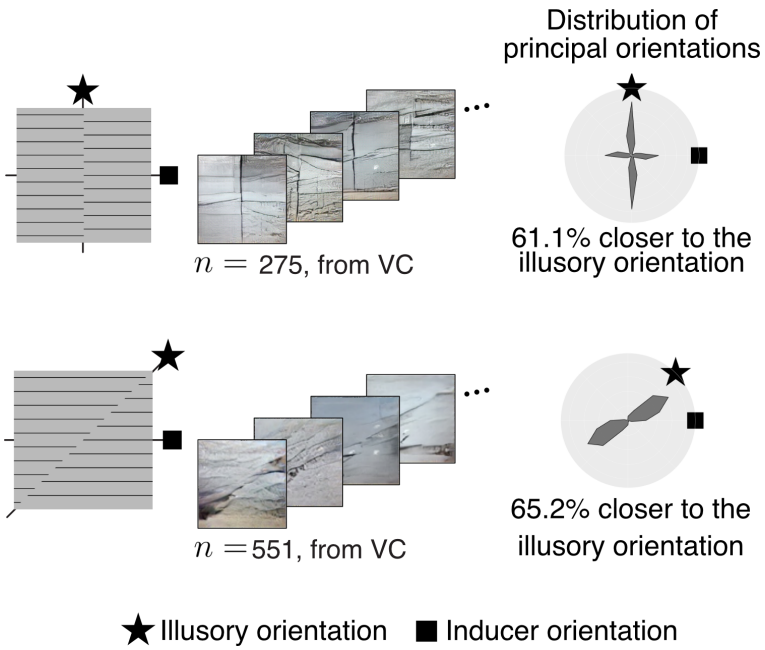

Fig. 7.3 Principal orientation distribution in reconstructions from VC. Results are derived from single-trial reconstructions aggregated across seven subjects. Data for all 90°- (top) or 45°- (bottom) difference configurations are combined, yielding a total of 420 samples; bin size = 15°. Figure from Cheng et al. (2023).

Further exploration was conducted into the influence of the number of inducer lines on the strength of illusory perception. The orientation distributions (Figure 7.5) reflected a gradual weakening of illusory lines concurrent with the reduction in inducer lines. The proportion of orientations closer to the illusory orientation was notably decreased from 9 to 3 lines.

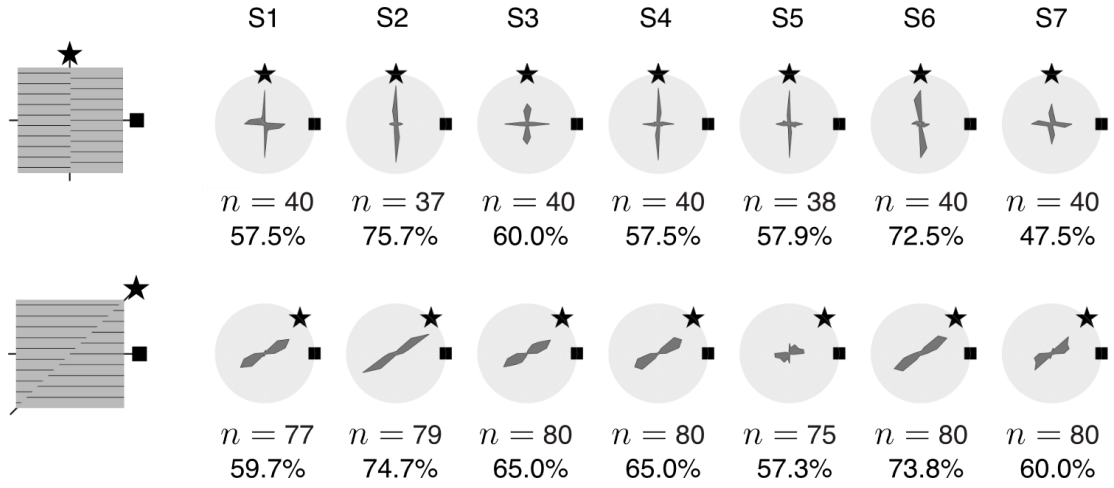

Fig. 7.4 Principal orientation distribution in single-trial reconstructions from VC for individual subjects. Data for all 90°- (top) or 45°- (bottom) difference configurations are aggregated for each subject, resulting in n samples; bin size = 15°. The proportion of principal orientations closer to the illusory orientation is indicated below the sample size. Figure from Cheng et al. (2023).

These findings validate that the reconstructions are reflective of the illusory modifications influenced by the quantity of inducer lines.

Utilizing fMRI activity captured from distinct regions within the visual cortex, we have successfully obtained reconstructions of visual stimuli. The distributions of principal orientations, compiled across configurations with a 90°- and 45°-differences from individual subjects (Figure 7.6 and 427.7), suggest that areas V1–V3 were typically consistent in generating accurate reconstructions of both the illusory and the inducer lines.

Quantifying global and local presence of illusory orientation

Perceptions of the illusory line typically manifest at the adjoining segments of inducer gratings. Nonetheless, the principal orientation detection method previously employed does not adequately represent the specific locality associated with this illusory perception. To scrutinize the localized manifestation of illusory orientation within reconstructions, we embarked on an analysis of local image regions, discerning the principal orientation distinctly within 1) illusory regions, anticipated to exhibit the illusory line, and 2) non-illusory regions, expected to display solely inducer lines (top panel of Figure 7.8). To demarcate these regions within each stimulus image, we isolated and cropped the four most prominent disks adjacent to the central lines of the image. Out of these, two disks encapsulated illusory lines while the remaining two did not. We computed proportions that were closer to the illusory for each local region, alongside the global region from prior analysis. The closer-to-illusory

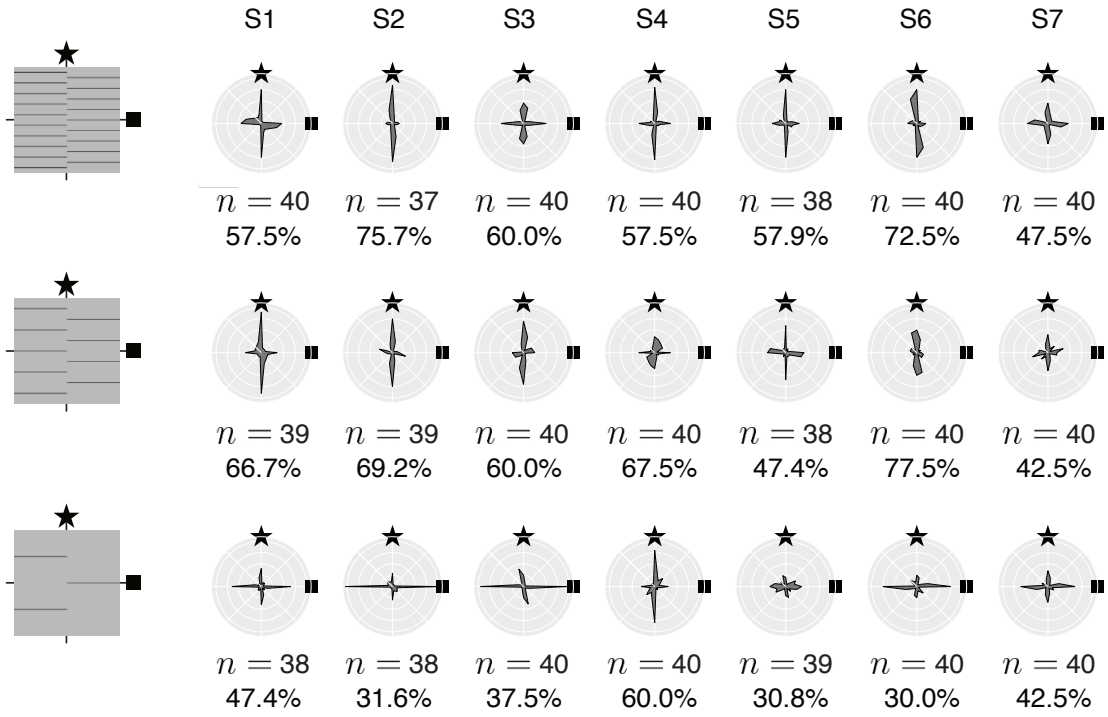

Fig. 7.5 Comparison of reconstructions with varying numbers of inducer lines (VC, individual subjects). Data for all 19- (top), 9- (middle), or 3- (bottom) line configurations are aggregated for each subject, yielding n samples; bin size = 15°. The proportion of principal orientations closer to the illusory orientation is specified below the sample size. Figure from Cheng et al. (2023).

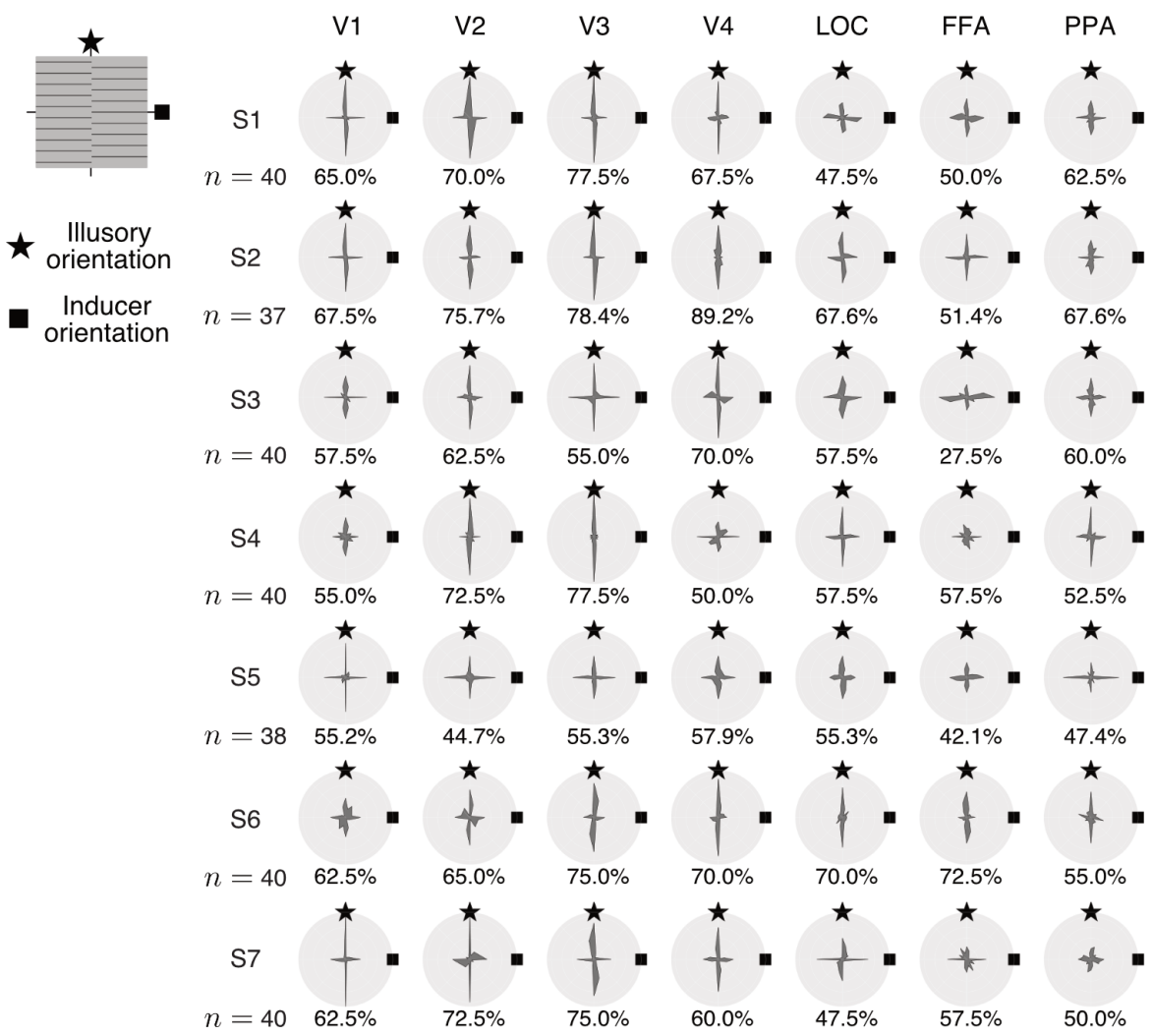

Fig. 7.6 Comparison of reconstructions from different visual areas for 90°-difference configurations. Results are based on single-trial reconstructions aggregated for each subject, yielding *n* samples; bin size = 15°. The proportion of principal orientations closer to the illusory orientation is indicated below each polar plot. Figure from Cheng et al. (2023).

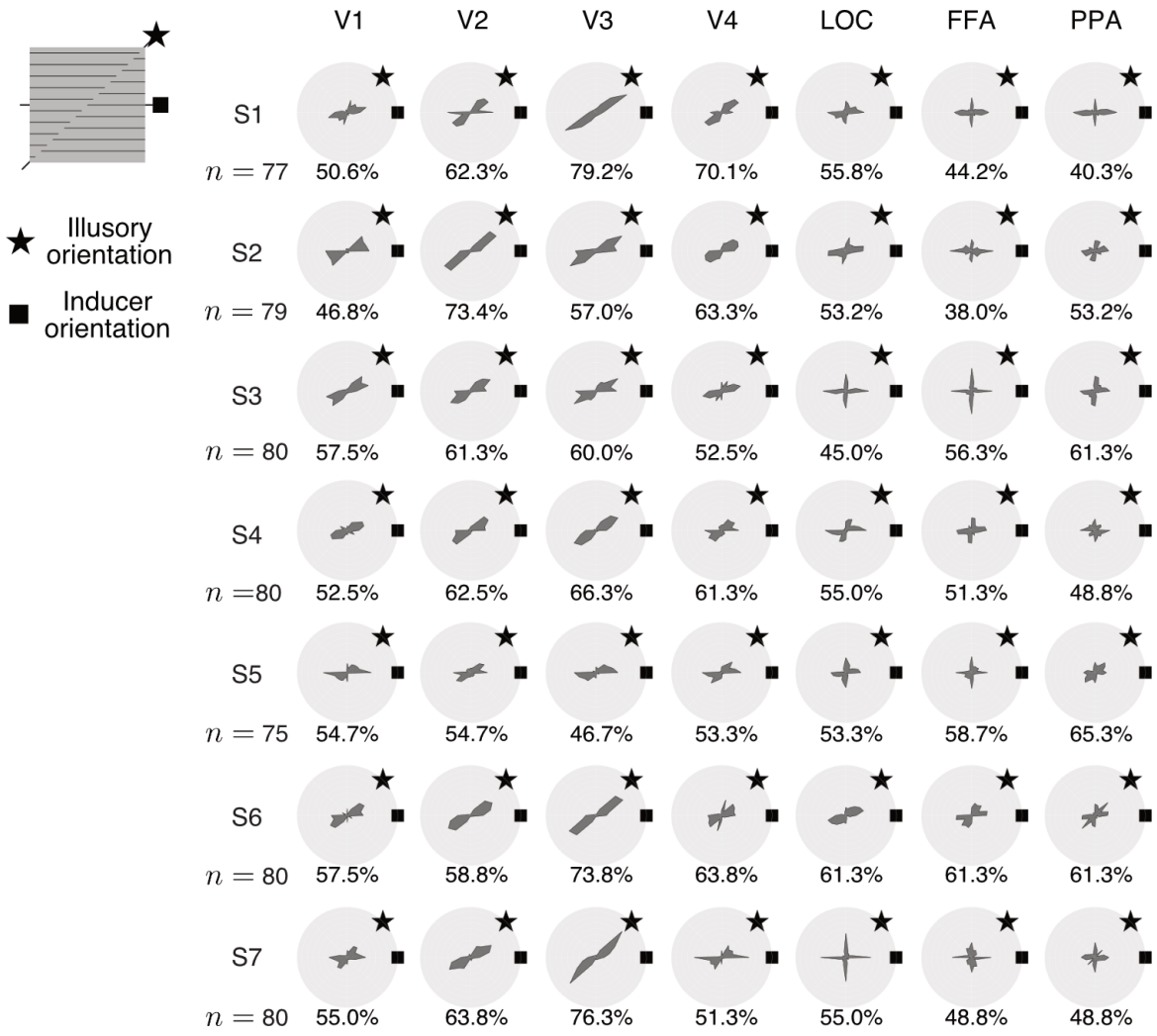

Fig. 7.7 Comparison of reconstructions from different visual areas for 45°-difference configurations. Results are based on single-trial reconstructions aggregated for each subject, yielding n samples; bin size = 15°. The proportion of principal orientations closer to the illusory orientation is indicated below each polar plot. Figure from Cheng et al. (2023).

proportions in local illusory regions bore resemblance to those in the global region. However, these proportions in local non-illusory regions were markedly lower, as evidenced by one-sided z tests for proportions ($p < 0.01$ in 7 out of 7 subjects). The similar tendency was obtained for 90°- and 45°-difference configurations, respectively (Figure 7.9).

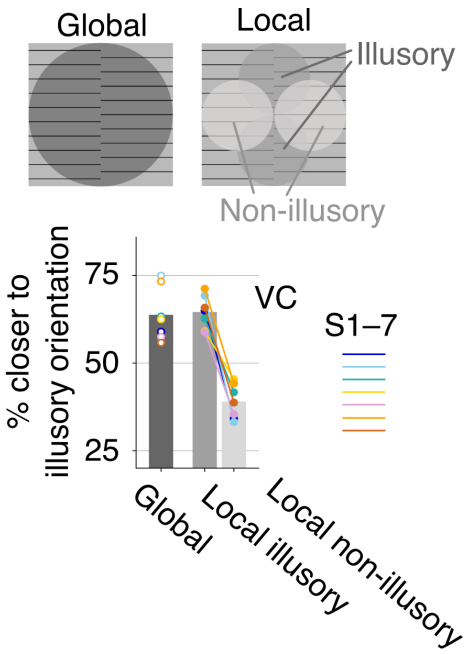

Fig. 7.8 Global and local presence of illusory orientation in reconstructions from VC. The proportion of principal orientations closer to the illusory orientation than the inducer orientation is shown for global and local image regions in reconstructions aggregated across all subjects and configurations. Individual subjects are represented by color circles and lines. Solid circles indicate comparisons with statistically significant differences at the individual level. Figure from Cheng et al. (2023).

The strength of the illusory orientation component for the global regions was highest around V2 to V4 (Figure 7.10). The difference between local illusory and non-illusory regions, which indicates the consistency with the local illusory percept, was substantial in V1 to V3 (one-sided z tests for proportions; $p < 0.05$ in five out of seven subjects at V1; all seven subjects at V2, V3, and V4; six out of seven at LOC; four out of seven at FFA; and three out of seven at PPA). The overall global and local trends in reconstructions of illusory images were found to be similar to the trends observed for the positive control images, although the strength of the illusory line was weaker than that of the real line in V1 to V3 (Figure 7.11). The strength of illusory orientation components (indicated by global closer-to-illusory proportions) in V1 to V4 and LOC appears to reflect the strength of subjective line perception, which was manipulated by the number of lines (Figure 7.12).

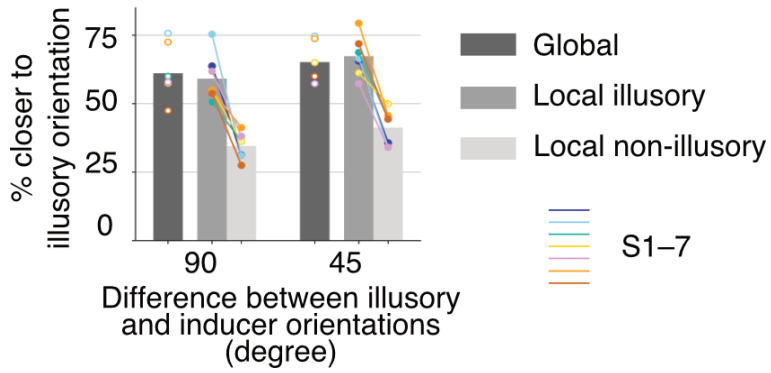

Fig. 7.9 Proportion of principal orientations closer to the illusory orientation than the inducer orientation for 90°- and 45°-difference configurations. Individual subjects are represented by color circles and lines. Solid circles indicate comparisons with statistically significant differences at the individual level. Figure from Cheng et al. (2023).

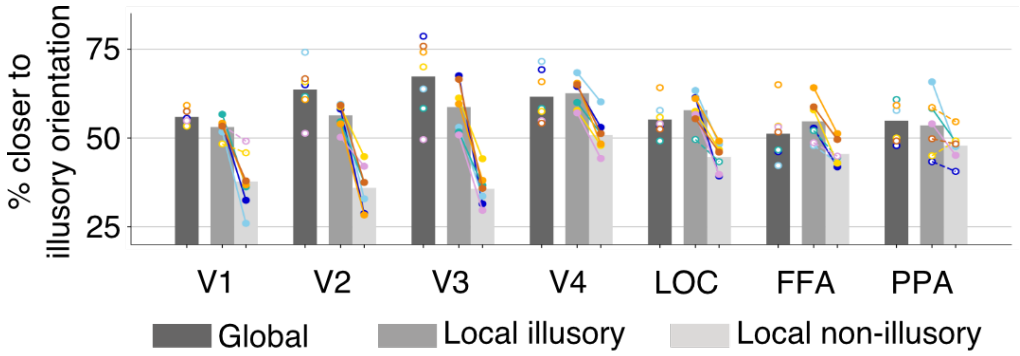

Fig. 7.10 Comparison of reconstructions from different visual areas. The proportion of principal orientations closer to the illusory orientation than the inducer orientation is calculated by aggregating all subjects and configurations. Individual subjects are represented by color circles and lines. Solid circles indicate comparisons with statistically significant differences at the individual level. Figure from Cheng et al. (2023).

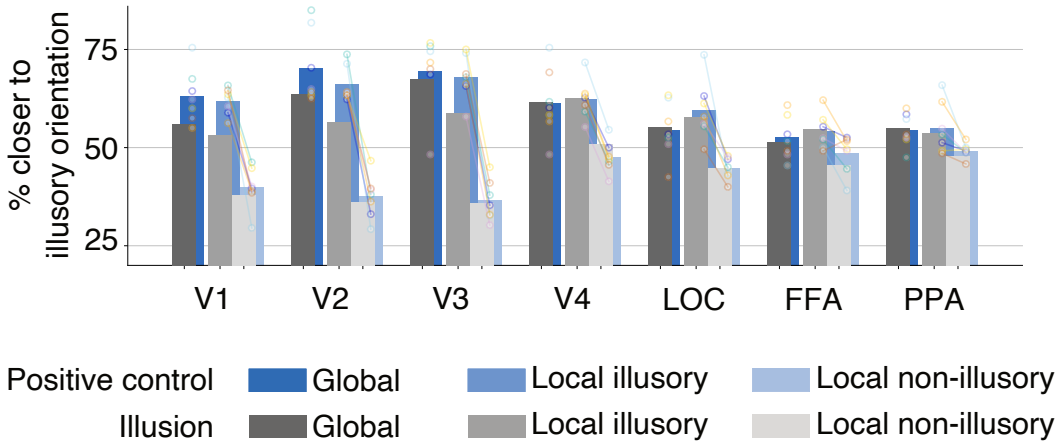

Fig. 7.11 Comparison between positive control and illusion conditions. The proportion of principal orientations closer to the illusory orientation than the inducer orientation is shown for different visual areas (aggregated across all subjects and configurations; gray bars are identical to those in Fig. 7.10). Individual subjects for the positive condition are represented by color circles and lines. Solid circles indicate comparisons with statistically significant differences at the individual level. Figure from Cheng et al. (2023).

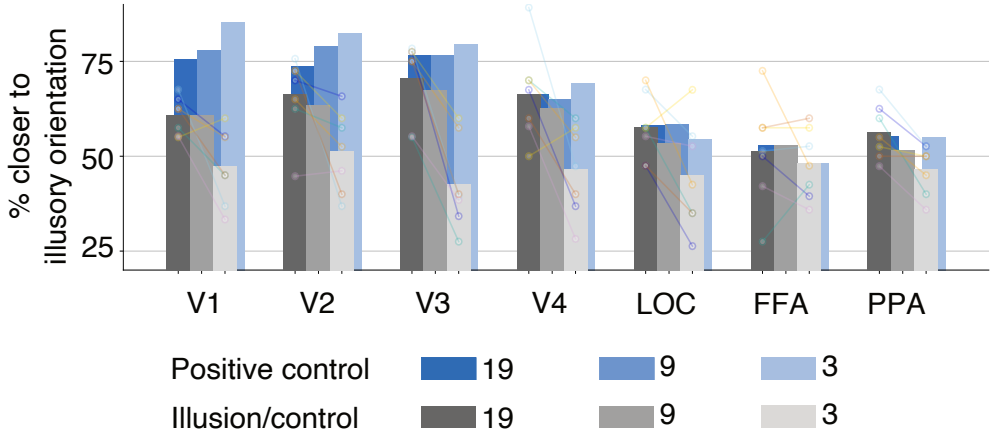

Fig. 7.12 Comparison between different numbers of inducer lines. The proportion of principal orientations closer to the illusory orientation than the inducer orientation in global regions is shown for different visual areas. Individual subjects for illusion or control conditions are represented by color circles and lines. Figure from Cheng et al. (2023).

7.3 Discussion

The evaluation results for reconstructions of line illusions show significant differences between two types of local image regions, decided by whether illusory lines were expected to be reconstructed or not, in V1 to V4 and LOC for most subjects. Moreover, the diminishing strength of illusory reconstructions with fewer inducer lines suggests that the early visual areas integrate contextual information like the number of inducers in constructing the final illusory percept. These results are consistent with earlier studies (von der Heydt and Peterhans, 1989; Grosof et al., 1993; Larsson and Amunts, 1999; Mendola, Dale, Fischl, Liu, and Tootell, 1999; Ramsden et al., 2001; M. M. Murray et al., 2004; Halgren et al., 2003; Sáry et al., 2007; Knebel and Murray, 2012b; Pan et al., 2012b; Cox et al., 2013; Pak et al., 2020), demonstrating the involvement of low-level areas and the lateral occipital complex (LOC) in processing both illusory lines.

Besides providing evidence for the 'where' problem of illusory contour processing, our results further revealed the exact form of 'what' is represented by each visual area. The early visual areas like V1-V3 can accurately represent both the illusory contours and inducer lines that give rise to the illusion. This suggests that these low-level areas are capable of representing the basic visual features that constitute the illusion. However, higher areas like V4 and beyond show a diminished ability to localize and represent the illusory contours precisely, even though they can still represent the illusory and inducer lines to some degree. This dissociation between lower and higher visual areas points to the hierarchical processing of illusory contours, indicating that the local representations of illusory contours are formed in early visual areas.

The comparisons between evaluation results for illusion and positive conditions suggest that both illusory and real lines are represented similarly throughout the visual areas; in other words, they share similar neural representations across areas. One interesting observation is that the actual strength of the illusory line was weaker than that of the real line in V1–V3, which could be related to the difference in perceptual appearances between illusory and real lines.

Overall, these results demonstrate how illusory contour is represented along visual hierarchy, with low-level areas more consistent with our overall illusory percepts.

Chapter 8

Reconstruction of neon color spreading along the visual hierarchy

This chapter investigates how the visual system reconstructs the perception of neon color spreading illusions like the Ehrenstein and Varin patterns. By generating reconstructions from activity patterns across the visual hierarchy, we can assess where and how the illusory color filling-in effects are represented. This can shed light on the underlying neural mechanisms that give rise to these vivid subjective color experiences from relatively simple line configurations. The content of this chapter is based on sections "RESULTS: Quantitative analyses of illusory color across multiple brain areas" and "Supplementary Materials" of Cheng et al. (2023).

8.1 Reconstruction

We generated reconstructions of neon color spreading illusions like the Ehrenstein and Varin configurations from activity patterns in different visual areas along the hierarchy. Representative examples of these reconstructions are shown, highlighting differences between illusory, control, and positive control conditions across areas. In the illusion condition of the Ehrenstein configuration, reconstructions across the visual areas displayed red regions analogous to those observed in the positive control condition. Notably, the lower visual areas more accurately represented the dimensions of the red regions and inducer lines (as shown in Figure 8.1 and 8.2). Conversely, within the control condition, designed to diminish or negate the illusory color spreading by introducing a gap in line width, the reconstructions predominantly lacked the color, despite both the illusion and control stimuli comprising identical red regions. Observations from reconstructions pertaining to the Varin illusion condition revealed similarities to the positive control condition primarily in mid-to-higher

visual areas, characterized by expansive red regions (Figure 8.3). At the lower visual areas (V1–V3), the color presence was minimal. However, V2 and V3 demonstrated the capability to faithfully reconstruct the actual color in the positive control condition. The inducer regions were generally inadequately reconstructed across both the illusion and positive control conditions, even in the lower areas. While reconstructions from these lower areas hinted at square-like outlines, they did not replicate the illusory square shape, instead seeming to encompass the entire stimulus area. In contrast, the control condition showed almost a complete absence of color; however, square outlines were more distinctly reconstructed from lower areas, aligning with the illusory squares but devoid of color spreading.

8.2 Evaluating Subjective Components in Reconstructions

To quantify the extent of reconstructed neon color filling-in effects, we employed a regression analysis approach. This involved modeling the reconstructions based on regressors for the expected illusory color surfaces and the inducing stimulus configurations. The regression coefficient for the illusory surface provided a metric to evaluate the representation of subjective color spreading across visual areas and conditions.

8.2.1 Method

In order to quantify the extent of color diffusion, regression analysis was conducted on the pixel color values within each reconstructed image (Figure 8.4). This analysis entailed generating redness (saturation) maps derived from the original RGB values of both reconstructed and stimulus images. Furthermore, redness maps were crafted for anticipated illusory surface regions. The redness map profile obtained from a reconstruction was then modeled based on those corresponding to the expected illusory surface region and the stimulus. It is important to note that the regressor for the illusory surface was uniformly applied across both the illusory and corresponding control conditions to facilitate comparative analysis. The coefficient for the illusory surface 1 was employed as the metric for evaluating the reconstruction of illusory color. These regression coefficients were computed for every individual trial (reconstruction) and subsequently aggregated across various configurations (incorporating differing sizes and numbers of lines for Ehrenstein patterns) within each participant and brain region.To assess the phenomenon of color filling-in, we conducted a regression analysis to approximate a reconstructed image through the amalgamation of the illusory surface and the inducing stimulus, introducing an additive error term for accuracy. Furthermore, utilizing the actual color surface from the positive control image, we executed a supplementary regression analy-

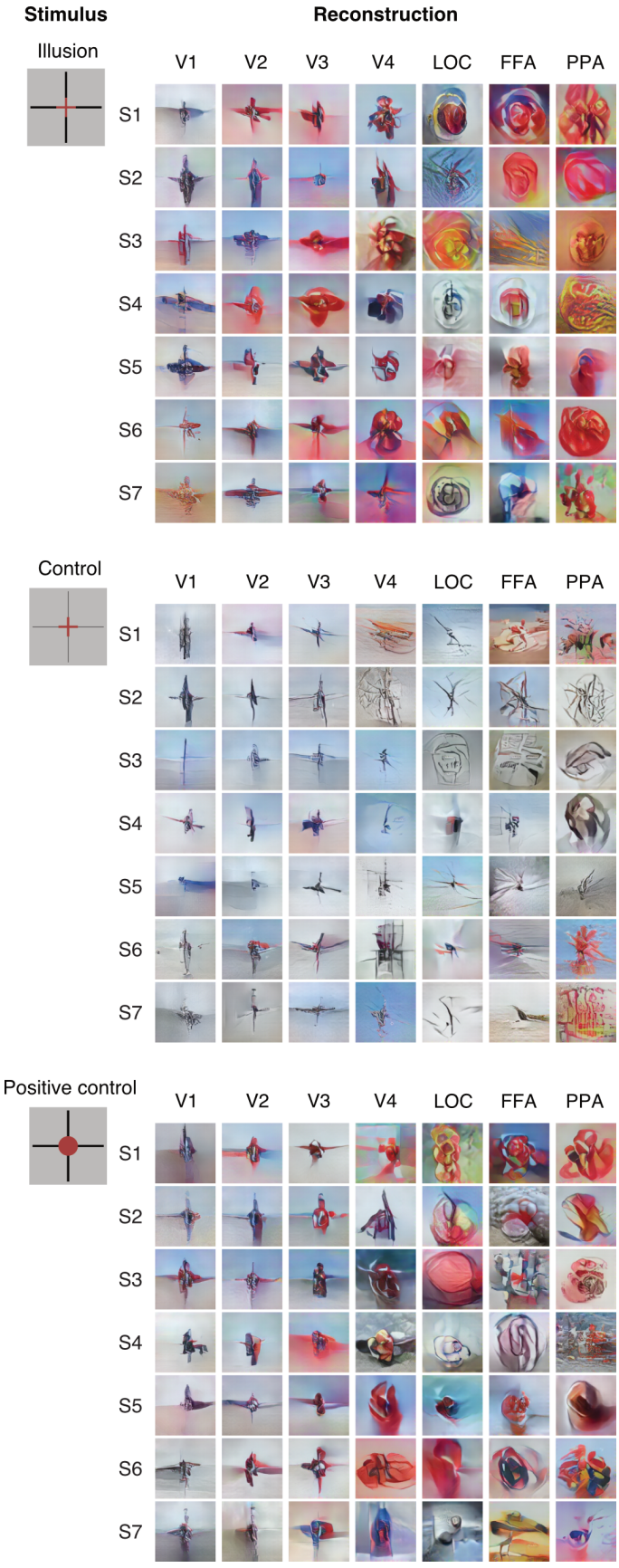

Fig. 8.1 Single-trial reconstructions of small-size neon color spreading (Ehrenstein) from different visual areas. Representative reconstructions of the illusion (top), control (middle), and positive control (bottom) conditions are shown for each subject. Figure from Cheng et al. (2023).

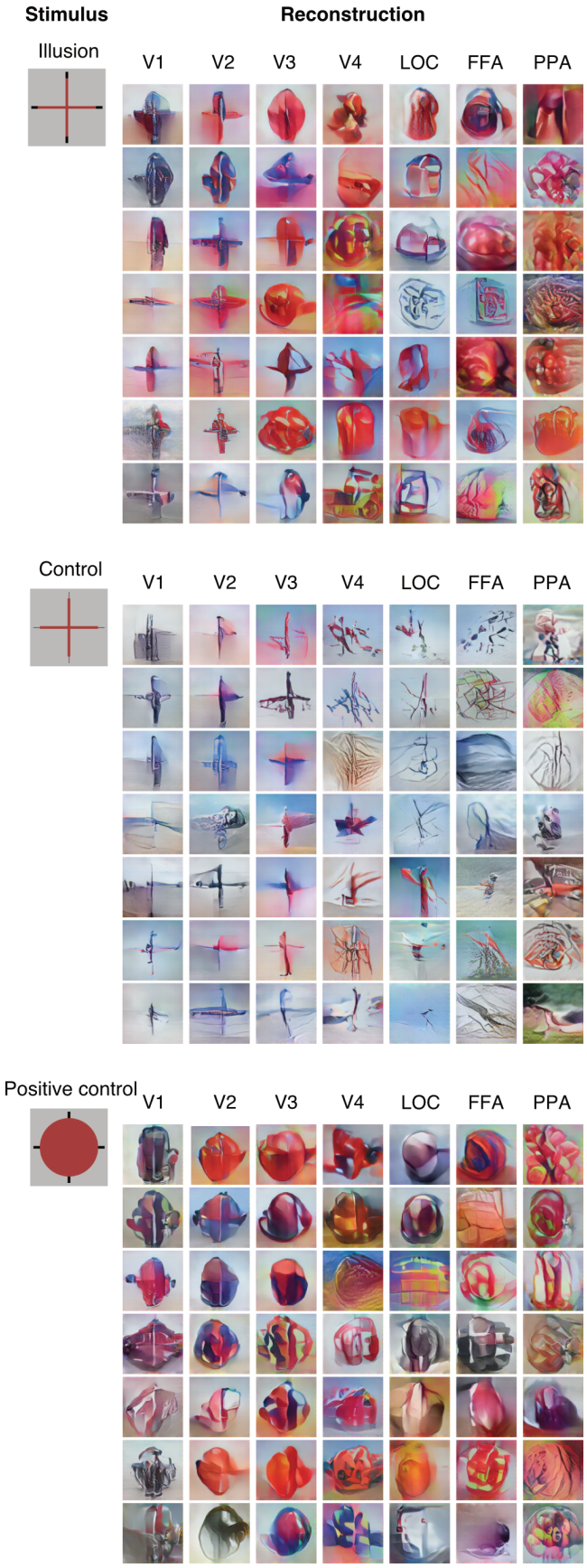

Fig. 8.2 Single-trial reconstructions of large-size neon color spreading (Ehrenstein) from different visual areas. Representative reconstructions of the illusion (top), control (middle), and positive control (bottom) conditions are shown for each subject. Figure from Cheng et al. (2023).

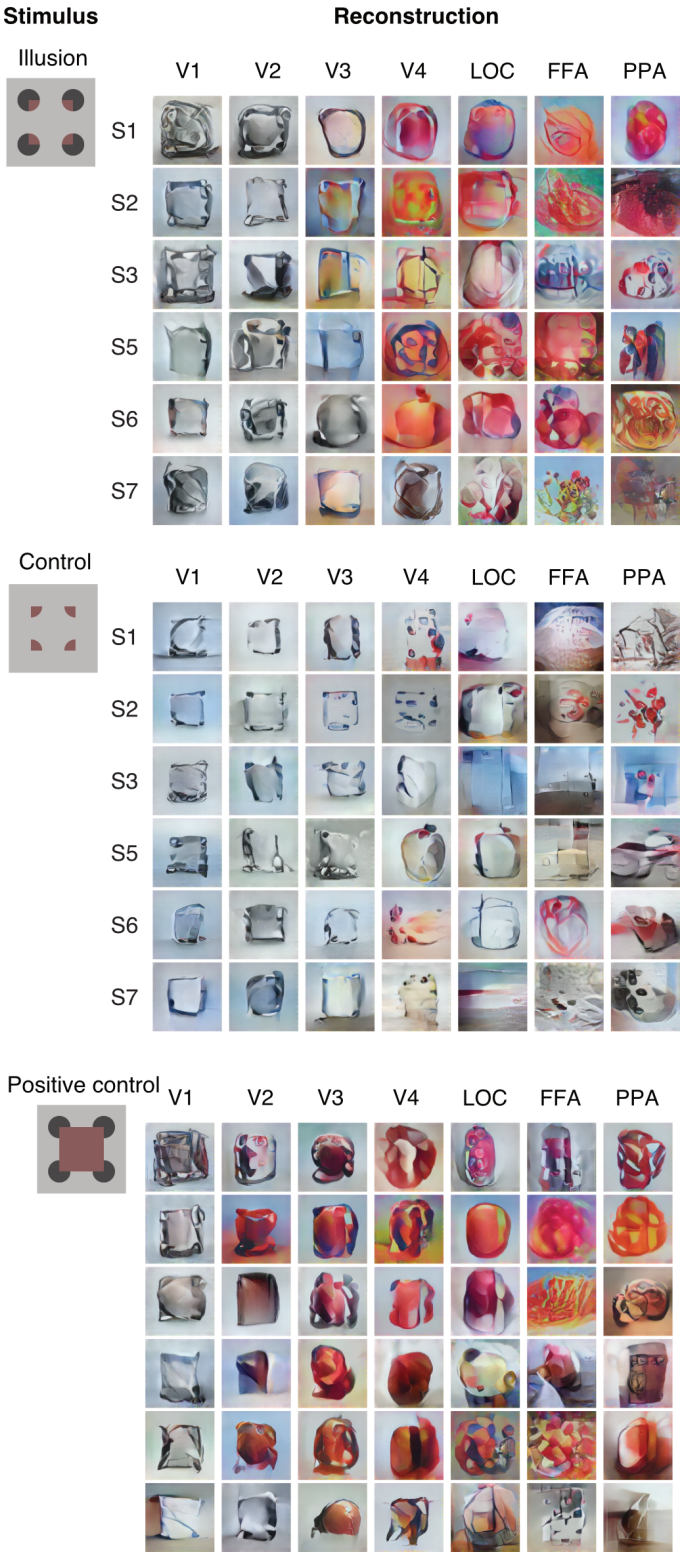

Fig. 8.3 Single-trial reconstructions of neon color spreading (Varin) from different visual areas. Representative reconstructions of the illusion (top), control (middle), and positive control (bottom) conditions are shown for each subject. Figure from Cheng et al. (2023).

sis, deliberately omitting the inducing stimulus regressor to facilitate a direct comparison between the illusory and real colors (Figure 8.5).

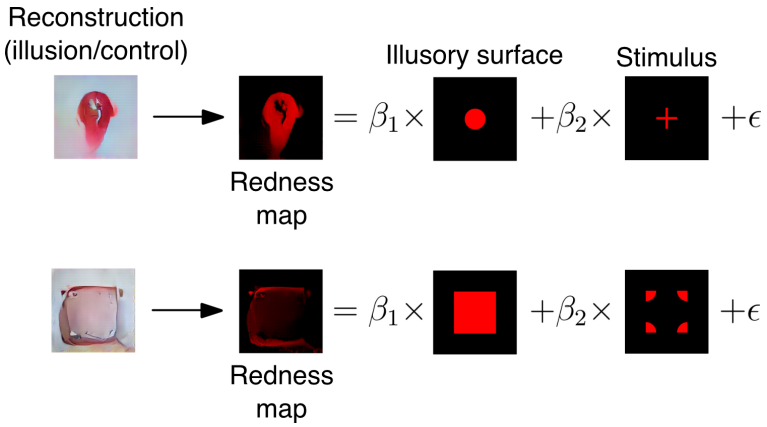

Fig. 8.4 Schematic of regression analysis for comparing the illusion and control conditions. The redness map of a reconstructed image for Ehrenstein (top) and Varin (bottom) was fitted by those of the illusory surface (expected region of color filling-in) and the stimulus. Figure from Cheng et al. (2023).

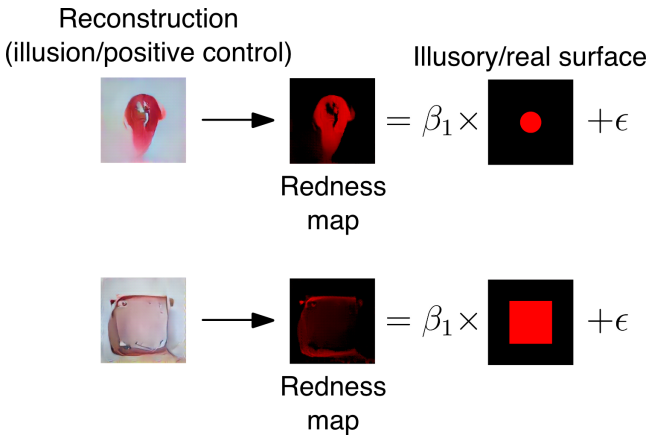

Fig. 8.5 Schematic of regression analysis for comparing the illusion and positive control conditions. The redness map of a reconstructed image for Ehrenstein (top) and Varin (bottom) was fitted by that of the illusory or real surface. Figure from Cheng et al. (2023).

Prior to conducting the regression analysis, we generated redness maps from the reconstructions, alongside the respective illusory/real surfaces and the inducing stimulus. These maps were created by transforming the RGB images to HSV color space using version 4.5.2 of the openCV library, followed by the extraction of saturation (S) values specifically for red pixels. Pixels not identified as red were assigned a value of zero. Red pixels were defined as those with hue values (H) ranging between 0° to 10° or 160° to 180°.

The regression models for the redness map of each single-trial reconstruction were formulated as follows:

$$y_i = \varepsilon_i + \beta_1 x_{i,1} + \beta_2 x_{i,2} \quad \text{(illusion versus control)} \tag{8.1}$$

$$y_i = \varepsilon_i + \beta_1 x_{i,1} \quad \text{(illusion versus positive control)} \tag{8.2}$$

where y_i denotes the value of the i-th pixel in the redness map of a reconstruction of size $h \times w$, ε_i represents the error term, and $x_{i,1}$ and $x_{i,2}$ are the values of the i-th pixel in the redness maps of the illusory/real surface and the inducing stimulus respectively. β_1 and β_2 are coefficients corresponding to the illusory/real surface and the inducing stimulus. These coefficients were determined by minimizing the sum of squared errors across all pixels within each reconstruction.

This methodological approach allowed for a detailed quantitative analysis of the color filling-in effect, providing insights into the relative contributions of illusory surfaces and inducing stimuli in the perception of color.

We utilized one-sided t tests on individual trial/reconstruction-derived illusory surface coefficients. To achieve the desired statistical power of 80% and discern a large effect size (Cohen's d = 0.8) at a 0.05 significance level, at least 20 samples per condition were required. Accordingly, we performed 20 trials for each stimulus image per participant. The actual count of trials meeting the exclusion criteria for comparing illusion versus control conditions were 77, 74, 80, 80, 75, 80, 80 (n_1) and 79, 75, 80, 80, 77, 80, 80 (n_2) for Ehrenstein across subjects S1–S7; and 20, 20, 20, 19, 20, 20 (n_1) and 20, 20, 20, 19, 20, 20 (n_2) for Varin across subjects S1–3 and S5–7.

8.2.2 Results

It was observed that the illusory surface coefficient, a measure indicative of the strength of illusory perception, generally exhibited an increase in the illusion condition compared to the control across various visual cortical areas (Figure 8.6). Specifically, for the Ehrenstein configuration, significant increases in the illusory surface coefficient were noted across multiple brain regions, with notable consistency in VC, V2, V3, V4, LOC, and FFA (one-sided t tests in individual subjects, $p < 0.05$ in 7/7 subjects at VC, 5/7 at V1, 7/7 at V2, V3, V4, LOC, and FFA, and 6/7 at PPA). The observations still hold for configurations of different sizes and line numbers (Figure 8.7). These areas, particularly V2-V4 and higher, demonstrated robust effects of the illusion for individual trial results (Figure 8.8). Conversely, for the Varin configuration, while there was a general increase in the illusory surface coefficient in the illusion condition, especially in mid-to-higher visual areas, the

effects were less consistent across subjects compared to those observed with the Ehrenstein pattern (Figure 8.9; one-sided t tests in individual subjects, $p < 0.05$ in 3/6 subjects at VC, 2/6 at V1, V2, V3, and 3/6 at V4, 4/6 at LOC, 3/6 at FFA, 1/6 at PPA).

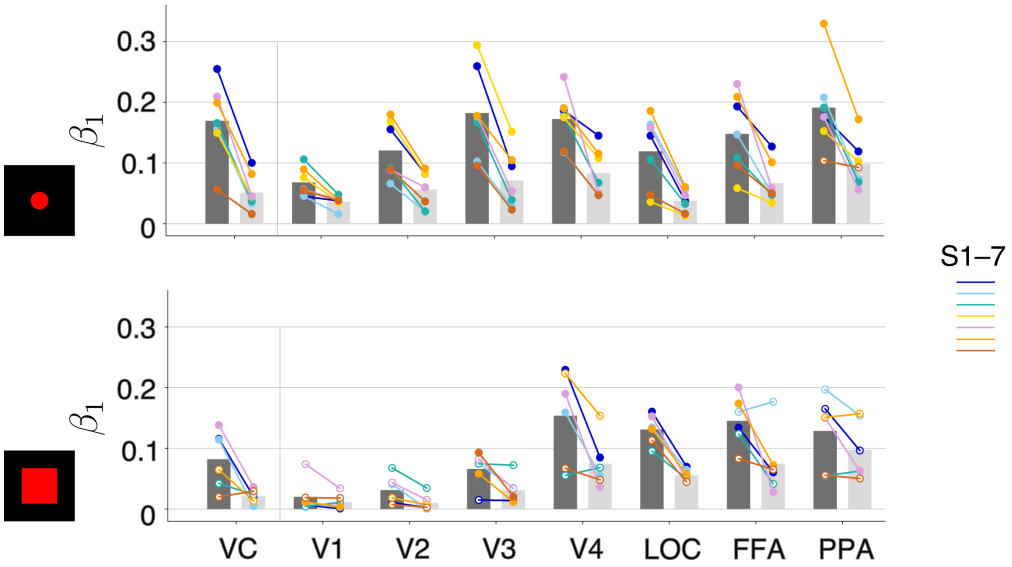

Fig. 8.6 Comparison of the illusory surface coefficient values between illusion and control conditions. Results for all configurations (sizes and numbers of lines) and seven subjects are aggregated for Ehrenstein (top). Results for six subjects are aggregated for Varin (bottom). Individual subjects are represented by color lines. Solid circles indicate comparisons with statistically significant differences at the individual level. Figure from Cheng et al. (2023).

Furthermore, regression analyses were performed to fit the profile of the redness map from reconstructions to the illusory or real color surface for both illusion and positive control conditions. For the Ehrenstein illusion, surface coefficients were comparable between the illusion and positive control conditions across brain areas, whereas for Varin, there was a trend of lower surface coefficients in the illusion condition in comparison to the positive control, particularly noted in low-to-mid brain areas (Figures 8.10). Interestingly, larger-sized Ehrenstein configurations tended to demonstrate lower illusory surface coefficients under the illusion condition as opposed to the positive control condition in lower-to-mid brain areas (Figures 8.11). These findings suggest that the intensity of illusory color reconstruction may be contingent on the spatial extent of the perceptual filling-in as well as the specific stimulus configurations employed.

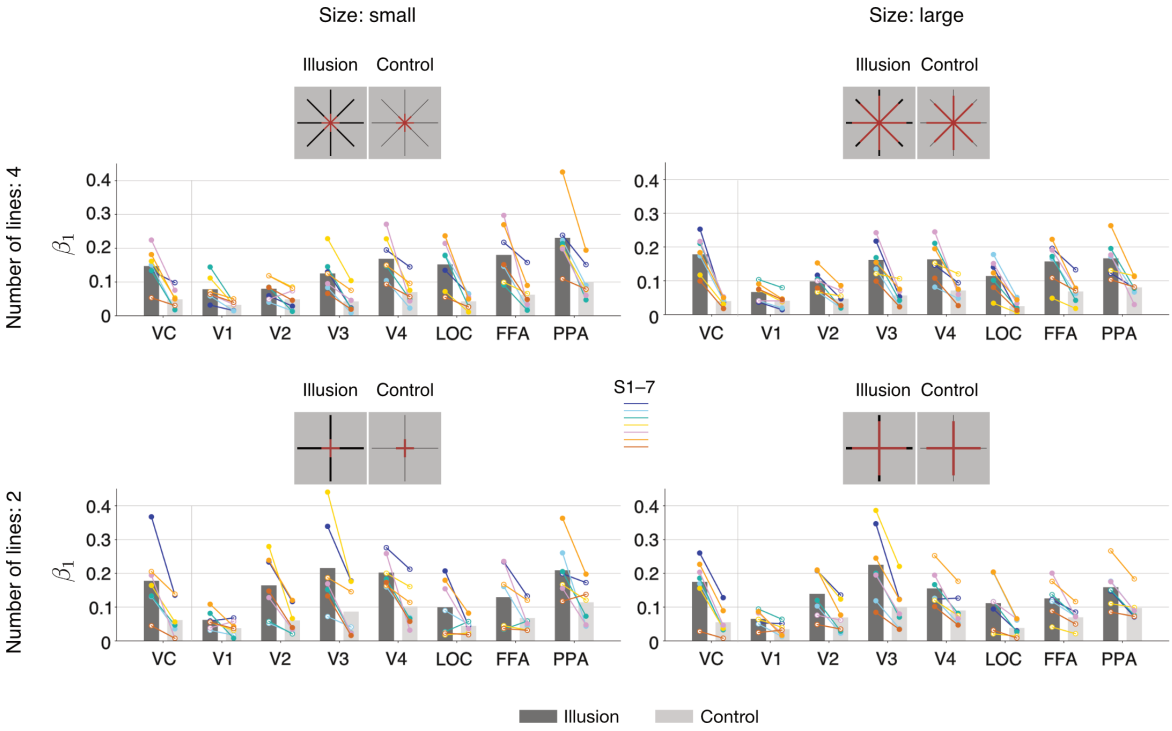

Fig. 8.7 Comparison of the illusory surface coefficient values between illusion and control conditions for different sizes and numbers of lines (Ehrenstein). Results are based on single-trial reconstructions from VC and specific visual areas. Individual subjects are represented by color circles and lines. Solid circles indicate comparisons with statistically significant differences at the individual level. Figure from Cheng et al. (2023).

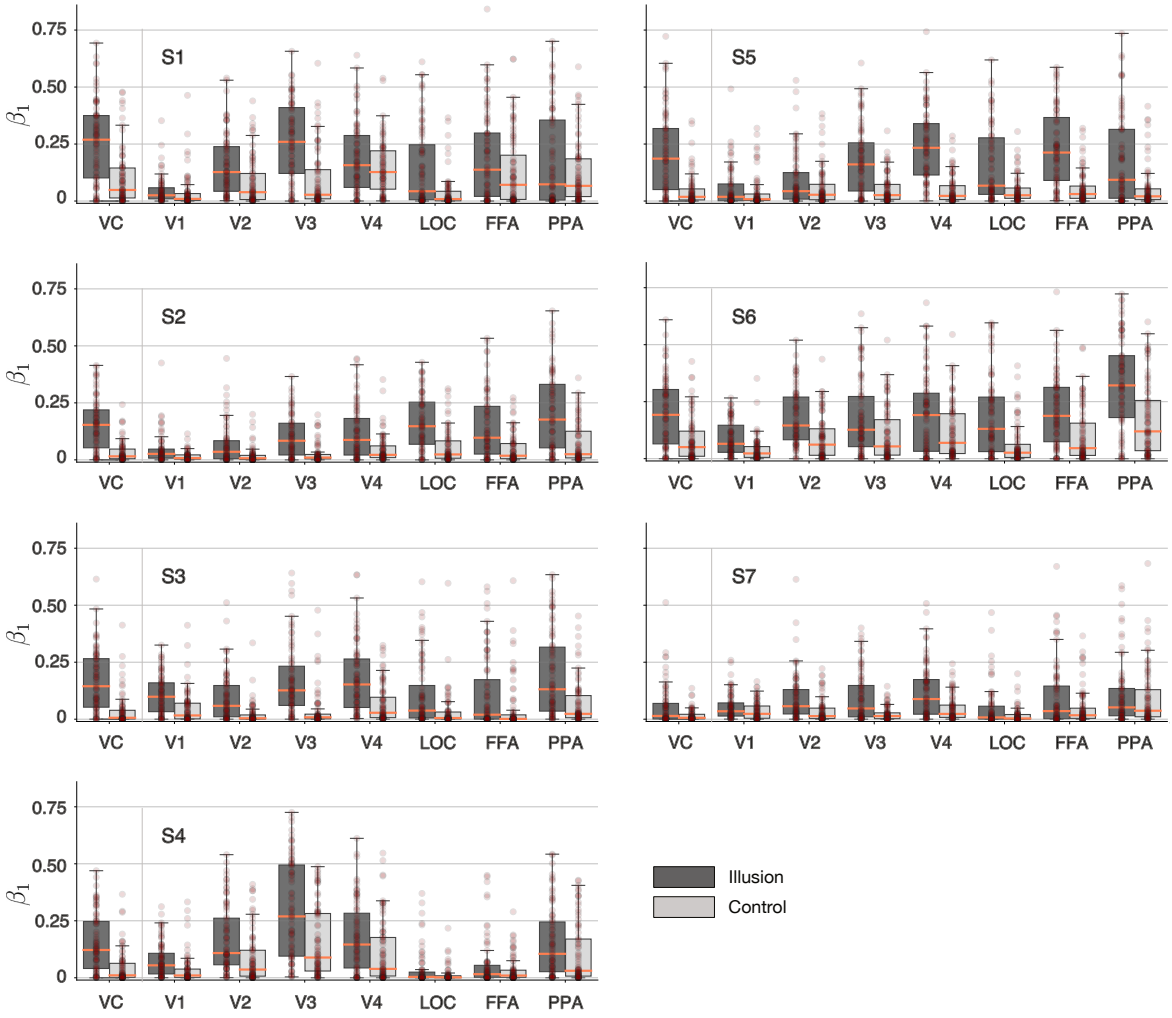

Fig. 8.8 Comparison of the illusory surface coefficient values between illusion and control conditions of Ehrenstein for individual subjects. Results are based on single-trial reconstructions from VC and specific visual areas. Individual trials are represented by dots (opacity indicates dot density). Median values are shown by coral lines, and interquartile ranges are depicted by shaded areas of boxplots. Figure from Cheng et al. (2023).

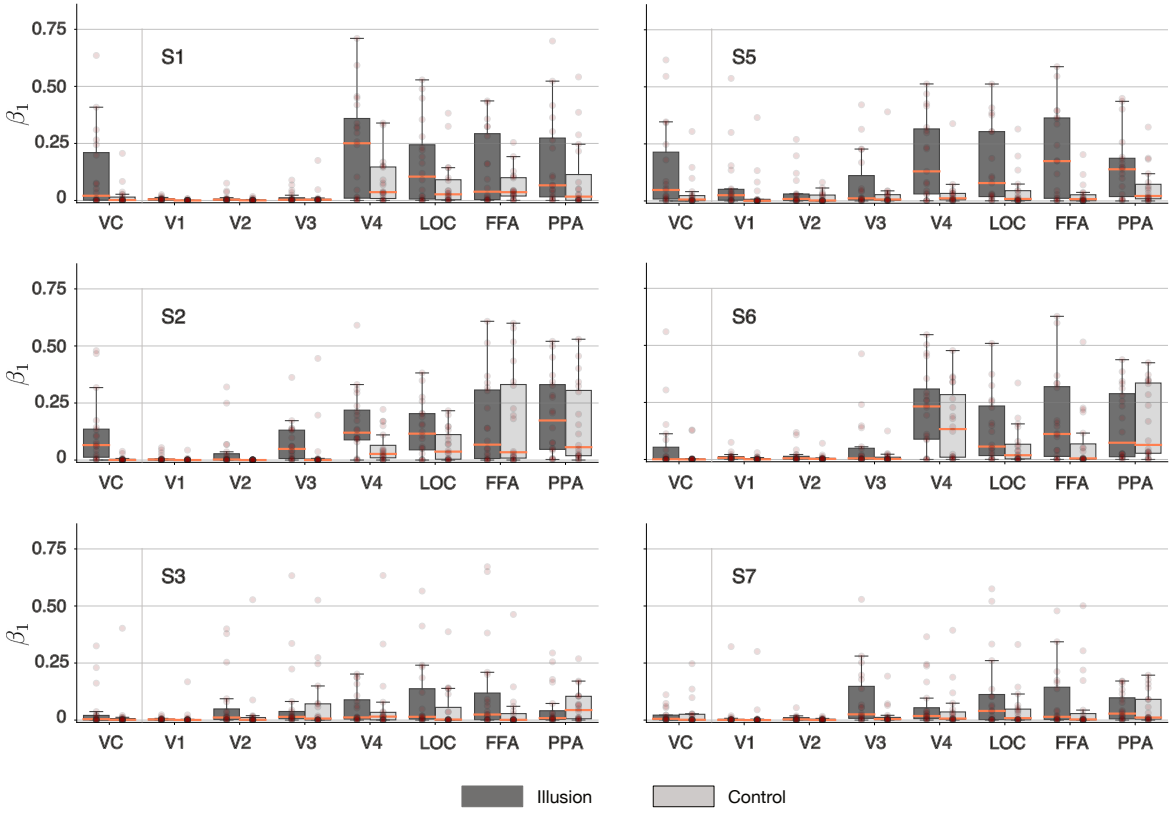

Fig. 8.9 Comparison of the illusory surface coefficient values between illusion and control conditions of Varin for individual subjects. Results are based on single-trial reconstructions from VC and specific visual areas. Individual trials are represented by dots (opacity indicates dot density). Median values are shown by coral lines, and interquartile ranges are depicted by shaded areas of boxplots. Figure from Cheng et al. (2023).

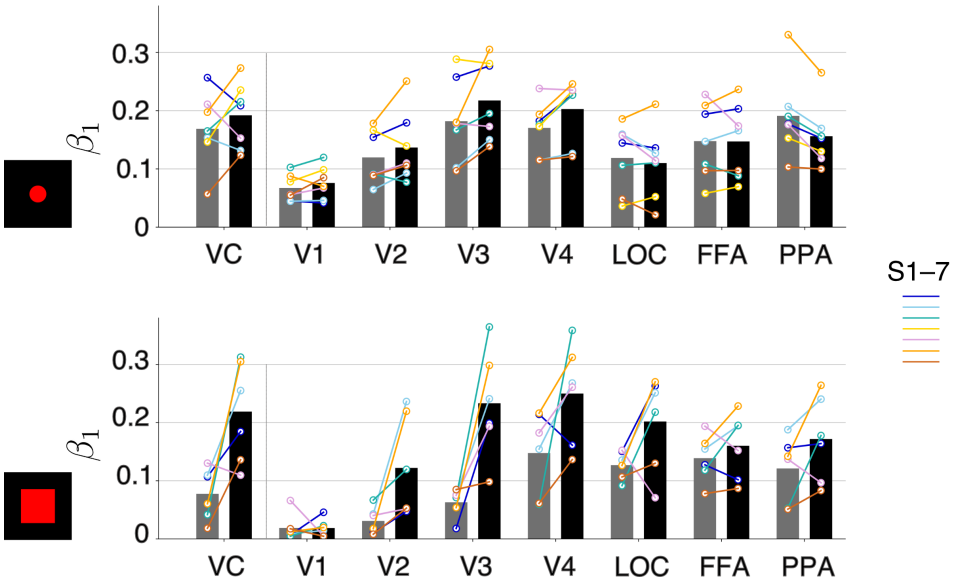

Fig. 8.10 Comparison of the illusory surface coefficient values between illusion and positive control conditions. Results for all configurations (sizes and numbers of lines) and seven subjects are aggregated for Ehrenstein (top). Results for six subjects are aggregated for Varin (bottom). Individual subjects are represented by color lines and dots. Figure from Cheng et al. (2023).

8.3 Discussion

The reconstructions and quantitative analyses reveal key insights into how neon color spreading is processed along the visual hierarchy. For the Ehrenstein configuration, both lower areas like V2–V4 and higher regions like LOC and FFA show robust reconstruction of the illusory color surfaces compared to control conditions without filling-in. This highlights that the impression of color filling-in, despite no physical color surfaces, may be affected by the computations spanning multiple stages of visual processing.

Interestingly, the lower areas more faithfully reconstruct the precise spatial extents and boundaries of the illusory color surfaces, while higher regions represent the surfaces more diffusely. This suggests a transition from precisely encoding the low-level visual features that induce color filling-in to a more integrated representation of the subjective color experience at higher levels. Furthermore, the modulation of reconstructed color strength by spatial factors like the size and line number of the inducing configurations points to the influence of contextual integration processes.

The Varin configuration results show different representations of color filling-in compared to the Ehrenstein configuration. Here the lower areas struggle to reconstruct the illusory color despite having robust color reconstructions in the positive control conditions. The filling-in

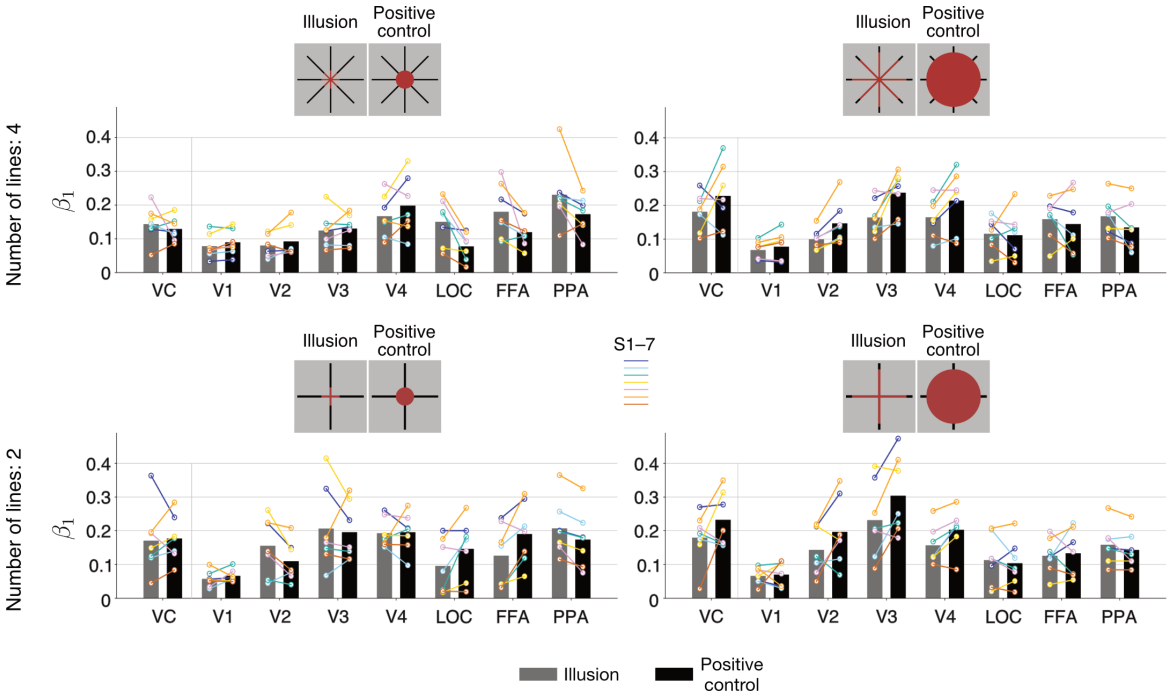

Fig. 8.11 Comparison of the illusory surface coefficient values between illusion and positive control conditions for different sizes and numbers of lines (Ehrenstein). Results are based on single-trial reconstructions from VC and specific visual areas. Individual subjects are represented by color circles and lines. Figure from Cheng et al. (2023).

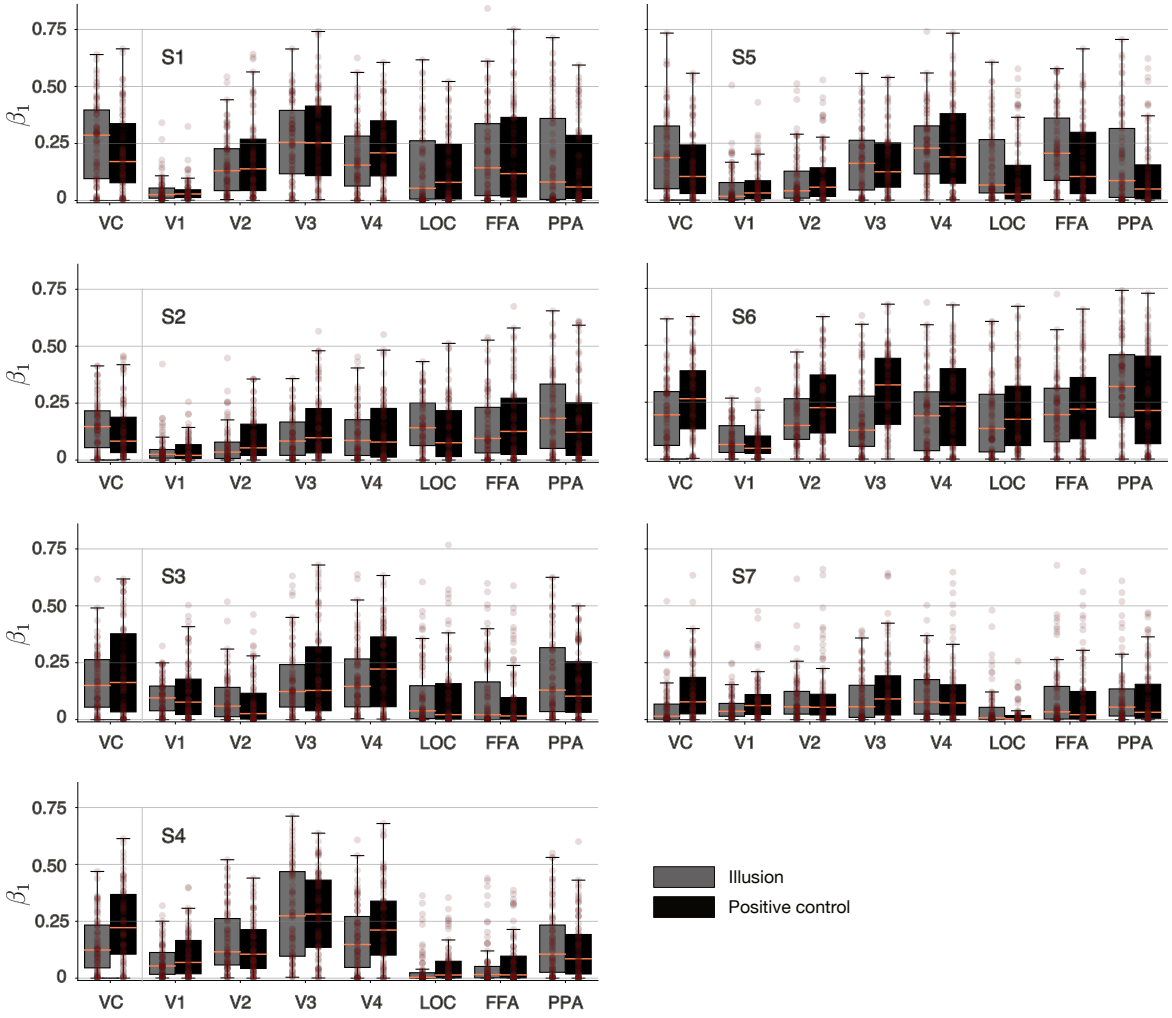

Fig. 8.12 Comparison of the illusory surface coefficient values between illusion and positive control conditions of Ehrenstein for individual subjects. Results are based on single-trial reconstructions from VC and specific visual areas. Individual trials are represented by dots (opacity indicates dot density). Median values are shown by coral lines, and interquartile ranges are depicted by shaded areas of boxplots. Figure from Cheng et al. (2023).

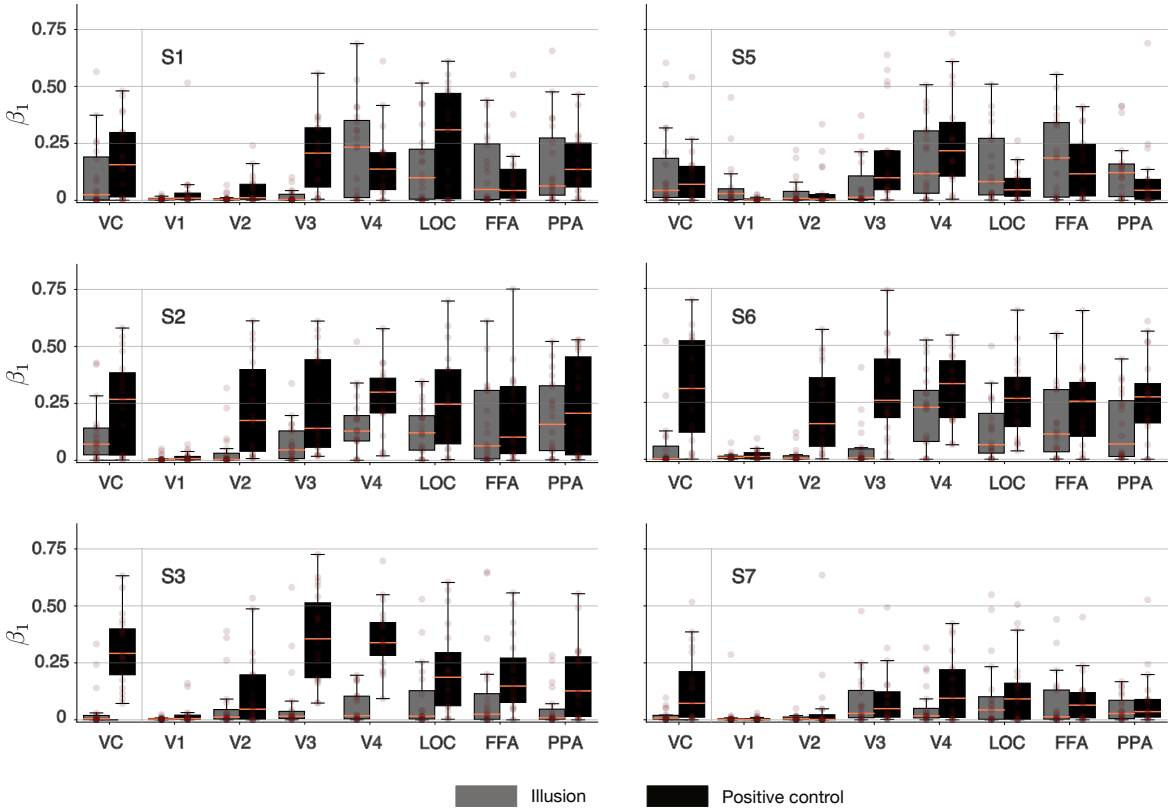

Fig. 8.13 Comparison of the illusory surface coefficient values between illusion and positive control conditions of Varin for individual subjects. Results are based on single-trial reconstructions from VC and specific visual areas. Individual trials are represented by dots (opacity indicates dot density). Median values are shown by coral lines, and interquartile ranges are depicted by shaded areas of boxplots. Figure from Cheng et al. (2023).

effects only become pronounced from V4 onwards, potentially implicating more higher-order processes.

Computational models suggest that neon color spreading and related phenomena arise from lateral interactions and top-down feedback mechanisms. Although the Ehrenstein and Varin configurations may share these mechanisms, their different inducer arrangements could result in varying strengths of lateral and top-down interactions. In the Varin configuration, the peripheral location of the red wedges might impede color diffusion from the periphery to central vision via lateral connections. At the same time, the top-down feedback from higher brain regions may not be strong enough to activate neurons in the lower areas effectively.

Overall, the results suggest that neon color spreading arises from hierarchical visual computations, with contributions from lower areas, especially for more local, boundary-dependent filling-in effects, as well as higher areas for representing the subjective color percepts. However, the results cannot disambiguate explanations of neural mechanisms about how color filling-in arises. It could be through bottom-up processing with lateral connections, first arising in V4, or top-down signals from higher areas where predictions are generated.

Chapter 9

General discussion

In this chapter, we provide an integrative discussion of our findings, highlighting the strengths and limitations of our approach, and proposing future directions for research. We revisit our methodology, discuss the implications of our results for understanding the neural mechanisms underlying illusory perception, and explore the broader implications for theories of consciousness.

By synthesizing our findings and situating them within the broader context of psychophysics, computational neuroscience and the study of consciousness, we aim to contribute to the ongoing endeavor of unraveling the mysteries of the mind and brain.

9.1 Potential confounds and alternative accounts

We have shown reconstructions of illusory percepts as images from single-trial brain activity by leveraging a computational model that learns the mapping between brain activity and the perception of non-illusory stimuli.

Success of reconstructing illusory percepts

To reconstruct illusory percepts, we employed a two-part model comprising DNN feature decoders and an image generator. The DNN feature decoders were designed to predict the activations of units in a DNN from brain activity. We trained these decoders on fMRI data collected while participants viewed a diverse set of natural images, including objects, materials, and scenes. The image generator was pre-trained using a large dataset of natural images without the involvement of fMRI data. We applied the trained decoders and generator to the fMRI data induced by illusory images and (positive) control images. The test results showed that the illusory features were exhibited in the reconstructions from fMRI activity of the visual cortex.

However, there are potential confounds and alternative accounts for the reconstruction results:

- Does DNN inherently represent illusory features?
  Answer: No. To test if our DNN inherently represents illusory features, We analyzed the responses of individual DNN units (section 6.1). Our results have shown that DNN units do not respond to illusory features in a similar way to real features (Fig. 6.5 and 6.6).

- Does our reconstruction pipeline create spurious lines and color?
  Answer: No. Feeding stimulus features into the generator did not lead to recognizable illusory features in reconstructions ("Stimulus feature" panel in Fig. 5.3).

- Does the presence of noise lead to spurious lines or color?
  Answer: No. To mimic the noise in brain-decoded features, the noises sampled from the unit-wise empirical distributions from decoded features of non-illusory trials were added to stimulus features. The reconstructions with stimulus feature plus noise did not exhibit illusory components ("Stimulus feature plus noise" panel in Fig. 5.3).

- Are our reconstruction results limited to a particular combination of model components?
  Answer: No. Qualitatively similar results were obtained by using different DNNs or generators (Fig. 6.9 and 6.10). Basically we can replace DNN and generator modules as long as they satisfy key requirements. The choice of DNN involves a trade-off in the degree to which it mimics brain processing. For the generator, it should be able to accurately generate the original images given the DNN features of the corresponding images.

- Is our reconstruction model's output space constrained by the training images?
  Answer: No. Our model demonstrated the ability to generalize beyond the specific stimuli used during training, successfully reconstructing illusory percepts that emerged from novel artificial shapes (section 5.2.2). This generalization capability suggests that the model has learned to capture the fundamental principles and rich information underlying the neural code, rather than merely memorizing the mapping between the training images and their corresponding brain activity patterns.

Dependency of reconstruction results on brain areas

We trained and tested separate reconstruction models for individual brain areas. To train a

decoder for a pair of DNN units and brain areas, we chose at most 500 voxels from an area as input. The total number of voxels was different for brain areas, ranging from hundreds to thousands. Some areas only had fewer than 500. The appearances of reconstructed images varied across different brain areas. Especially, higher cortical areas exhibited a more diffuse representation of illusory colors compared to lower areas.

However, there are potential confounds and alternative accounts for the dependency of reconstruction results on brain areas:

- Does feature prediction accuracy explain varied reconstructions across brain areas for the same attribute?
  Answer: No. The target features (fc6) were similarly well predicted from individual brain areas (Fig. 3.2 and Fig. 5.1A).

- Does expansive receptive field size explain the diffusive appearance of color in reconstructions from higher-level brain areas?
  Answer: No. The reconstructions of control images from higher areas did not show diffusive color (Fig. 8.1).

9.2 Revisit methodology

Our reconstruction framework bridges the gap between internal brain representations and their external manifestations in the physical world. This section provides a general discussion of the methodology from the perspective of externalization and scientific exploration of perceptual experiences.

9.2.1 Strengths

Advantages over "outer" psychophysical methods

Our methodology offers two key advantages over traditional psychophysical methods, enabling more comprehensive and efficient externalization and scientific exploration of perceptual experiences.

First, our method provides more direct access to perceptual experiences. Traditional psychophysical studies rely on behavioral responses, which are not direct measures of visual representations due to the complex processes involved in decision-making and motor command generation (Figure 9.1). These responses are influenced by cognitive factors such as decision criteria and biases, making it challenging to determine the origin of the response,

especially in the case of illusory stimuli where there is no correct answer. In contrast, our approach utilizes passive viewing experiments, eliminating the contamination from decision noise and providing more direct access to perceptual experiences.

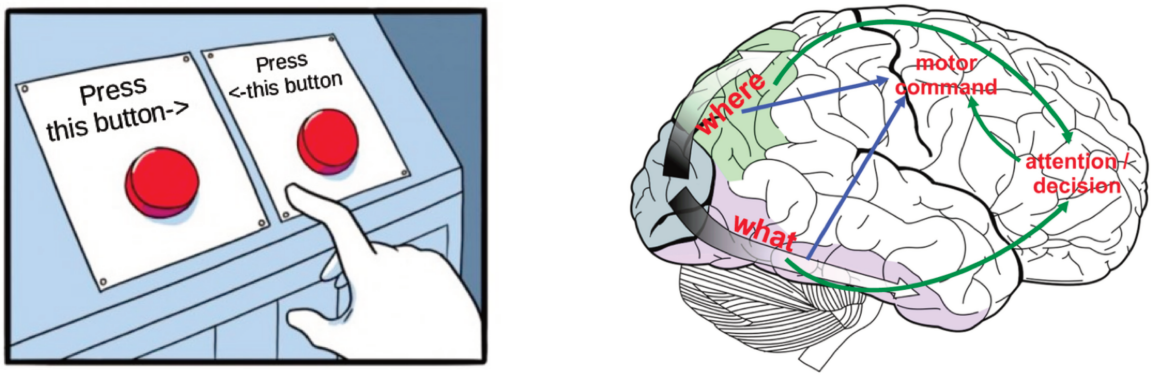

Fig. 9.1 Psychophysical experiments and the underlying complex processes in the brain. Behavioral responses are not direct measures of visual representations but involve complex processes. Figure adapted from Nere et al. (2012).

Second, our method facilitates the translation of holistic visual experiences into tangible data. Psychophysical measurements are often limited to specific properties, requiring additional efforts to design and conduct new experiments when a new visual property of interest emerges. This makes it inconvenient to study holistic experiences. Our decoding model learns the mapping between holistic perception and brain activity in the visual cortex, externalizing perceived content into easily interpreted formats suitable for various quantification analyses.

Advantages over other "inner" psychophysical methods

Our approach enables the externalization of mental contents, surpassing the limitations of conventional methods that primarily focus on testing qualitative hypotheses. Conventional approaches applied to fMRI data can determine which areas are involved in representing illusory features (Figure 9.2A and B). However, they do not reveal how brain activity patterns and their differences explain perceptions and their differences. In contrast, our reconstructions from individual areas reveal the strength of illusory percepts and its shared degree with real stimuli at different processing stages (Figure 9.2C).

Moreover, our approach can be applied to higher brain regions where expansive receptive fields contain both illusory and inducer components. Previous research often categorized widespread brain activity using coarse labels, potentially confounding the results with other factors, such as the properties of inducers. This issue becomes more significant in higher brain regions. By reconstructing spatial configurations in an image, our approach can provide a more accurate mapping between visual features and brain activity.

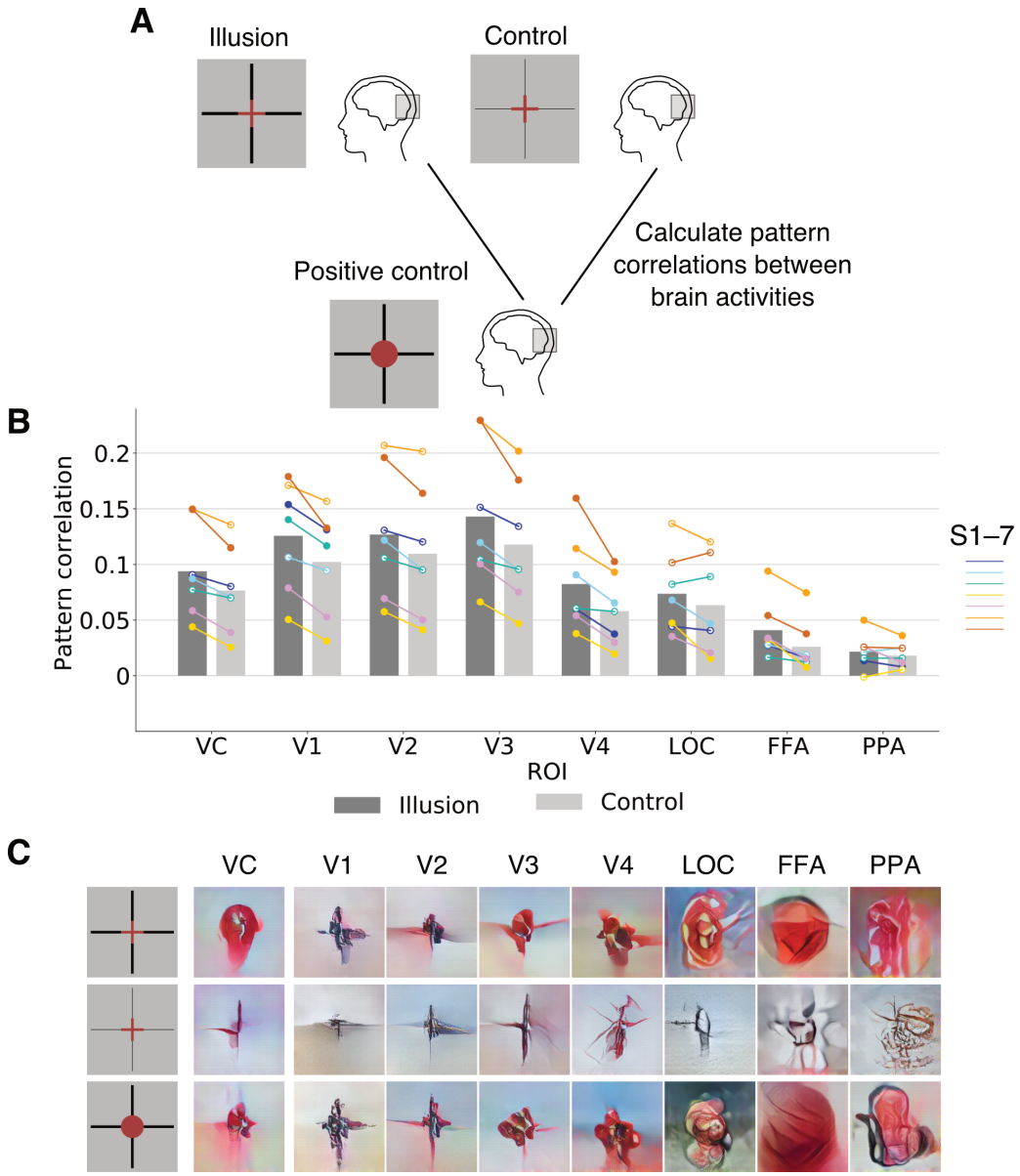

Fig. 9.2 Detection of illusion effect in brain areas. (A) Illustration of examining whether the neural representations of illusion stimuli are more similar to those of positive control stimuli, compared to control stimuli. Pattern correlations between illusion and positive control, or between control and positive control, are calculated. (B) Comparison of pattern correlations between illusion and control across areas. (C) Results using reconstruction method (from Cheng et al. (2023)).

9.2.2 Limitations and future directions

Limitations in illusion types

Our methods are most effective for illusions that exhibit spatially distinct illusory percepts and real stimuli. However, they face limitations in the following scenarios. First, they struggle with more complex or subtle illusions, such as the Café Wall illusion, which contains small squares that induce subtle changes in the perceived orientation (Figure 1.3). Second, it is challenging to study cases where people have different interpretations for the same stimulus without significant spatial changes in their mental image, such as with ambiguous images. One example is the face-vase illusion, which induces the perception of either a face or a vase, depending on how the observer interprets the borders between the black and white regions. To capture such perceptual alternations, a model that designates the ownership of borders, such as a segmentation model, is needed.

Limitations in peripheral reconstructions

Our methods generally show poor reconstructions for peripheral regions of visual stimuli. For example, in the Varin illusion, the black inducers were missing in the reconstructions, indicating a limitation in accurately capturing peripheral visual information. We should ensure that the method can reconstruct peripheral visual information encapsulated in brain activity as faithfully as possible. This aspect may be improved by increasing the diversity of peripheral regions in our training data. Specifically, we can manipulate images and increase the frequency of main objects located in the peripheral regions or allow free-viewing conditions in fMRI experiments. Additionally, we can explore whether attention can enhance the reconstruction performance of peripheral regions.

Improving generalization ability

Our methods use models trained with natural images and generalize them to illusory percepts formed from artificial shapes, requiring good generalization ability. This generalization ability may be partially attributed to the latent feature dimensions of the DNN, which effectively capture image-level information while being compatible with the brain's representations. Decoders were trained on a diverse set of images in which the features of individual dimensions appear frequently enough. The choice of generator also plays a significant role in this context. Recent studies suggest that although both GAN and diffusion models tend to memorize training images, diffusion models may recall more data than GANs, replicating pixel-level details, structures, and styles. As the field of artificial intelligence advances, we will be able to use better DNNs or generators. To improve reconstruction quality, enhancing

the generalization capabilities of both decoders and generators is also essential. This can be achieved by carefully designing the training image dataset or incorporating techniques such as data augmentation, transfer learning, and domain adaptation to bridge the gap between natural and artificial images.

Restricting output space

For the purpose of confirming specific hypothesis, pursuing a general model may not be the most efficient and effective direction. For example, it is worthwhile to try using artificial shapes as training images for both DNN and decoding models. This puts constrains to the output space of reconstruction model and is likely to generate less noisy reconstructed images from brain activity. By tailoring the training data to the specific domain of interest, we can create more specialized models that are better suited to capturing the nuances of illusory percepts. This approach can also help reduce the computational complexity and training time required for the models.

Extending to other materials

Our methodology has the potential for several extensions. It can be adapted to study illusory motion, 3D perception, or auditory illusions by utilizing DNNs and generators trained with diverse data types such as movies, 3D data, and sound data that do not inherently represent illusory features. In this case, our perception can not only be materialized as images but also in a more extended form. This adaptability could also extend to studying other subjective experiences like hallucinations, providing a broader application of our methods in understanding complex perceptual phenomena. However, extending our methodology to these domains may require significant modifications to the architecture of the DNNs and generators, as well as the development of new experimental paradigms and data collection techniques. Collaborative efforts between researchers from different fields, such as neuroscience, psychology, and computer science, will be crucial in addressing these challenges and pushing the boundaries of our understanding of the neural basis of perception.

9.3 Implications for neural mechanisms

Our study provides novel insights into the neural mechanisms underlying illusory perception and challenges traditional perspectives on the coupling between boundary detection and color perception. These findings pave the way for further investigations into the neural basis of conscious experience and the development of more comprehensive theories of visual perception and consciousness.

Representation of illusory percepts across brain regions

Our reconstructions from individual areas provide insights into the representation of illusory percepts across different brain regions. For illusory lines, we found that local illusory lines, which closely matched perceptual experiences, were better represented in areas V1 to V3. Central illusory lines tended to be more prominent in higher areas, such as the lateral occipital complex (LOC). Regarding illusory color, our results demonstrate the involvement of multiple areas, including higher areas within the ventral cortex, in its representation.

Interestingly, the neural representations of illusory color varied among different configurations. In the Ehrenstein configuration of neon color spreading, illusory color representations were present from low to high areas, overlapping with real color representations. However, for the Varin configuration, illusory color representations were primarily observed in mid-to-high areas, such as V4, while early areas like V1 and V2 showed a more pronounced representation of real color compared to illusory color.

New findings to be explained by future computational models

These findings challenge traditional perspectives on the neural mechanisms underlying illusory perception. Previous models, such as those proposed by Grossberg and Mingolla (1985) , have predominantly postulated a tight coupling between the perception of lines and colors, suggesting that the neural mechanisms underlying boundary detection and color perception are interconnected and work in unison to create coherent visual experiences.

However, our study presents a more nuanced view. The segregation of illusory line and color representations in different brain areas suggests a decoupling of these processes, indicating a more complex and distributed neural processing framework for visual perception. This is particularly evident in the context of the Varin configuration, where we observed that illusory color perception is not necessarily accompanied by the perception of illusory contours in lower visual areas. This dissociation further supports the idea that different neural substrates are responsible for different elements of visual perception.

Higher cortical areas exhibit a more diffuse representation of illusory colors and contours. This diffusion might indicate that these areas integrate information from various sources and perform more complex processing tasks. However, the precise role of these higher areas in top-down control and their contribution to illusory perception remains unclear. Moreover, in the case of the Varin configuration, lower areas did not clearly exhibit reconstructions of illusory color, suggesting a limited effect from top-down signals. Further research is needed to elucidate the mechanisms underlying the interactions between lower and higher visual areas in the construction of illusory percepts.

Implications for Theories of Consciousness

Our findings have important implications for theories of consciousness. The dissociation between the neural representations of illusory lines and colors challenges the notion of a single, unified neural correlate of consciousness (NCC). Instead, our results suggest that different aspects of conscious experience may be mediated by distinct neural mechanisms distributed across the brain. This highlights the need for more nuanced and dynamic models of consciousness that can account for the complex interactions between different brain areas and the emergence of subjective experiences from these interactions.

Future directions and integrating multiple approaches

To establish a causal relationship between brain activity and subjective percepts, future studies could employ lesion studies or transcranial magnetic stimulation (TMS) to selectively disrupt neural activity in specific brain regions and observe the effects on illusory perception. High-resolution brain measurements, such as layer-specific 7T MRI, have the potential to distinguish bottom-up and top-down signals, providing a more detailed understanding of the neural mechanisms underlying visual processing. Additionally, other modalities, such as magnetoencephalography (MEG), can be employed to investigate the temporal dynamics during the evolution of subjective experiences.

To gain a comprehensive understanding of the neural mechanisms underlying illusory perception, it is crucial to integrate findings from various approaches. Combining insights from neuroimaging, behavioral studies, computational modeling, and theoretical frameworks can help develop a more unified and coherent picture of how the brain processes and constructs subjective visual experiences.

References

Adelson, E. H. (1993). Perceptual organization and the judgment of brightness. *Science*, *262*(5142), 2042–2044.

Aitchison, L., & Lengyel, M. (2017). With or without you: Predictive coding and bayesian inference in the brain. *Current Opinion in Neurobiology*, *46*, 219–227.

Akrami, A., Kopec, C. D., Diamond, M. E., & Brody, C. D. (2018). Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature*, *554*, 368–372.

Amerine, M. A., Roessler, E. B., & Ough, C. S. (1965). Acids and the acid taste. i. the effect of pH and titratable acidity [Copyright 1965 by the American Society for Enology and Viticulture]. *American Journal of Enology and Viticulture*, *16*(1), 29–37.

Anderson, B., & Julesz, B. (1995). A theoretical analysis of illusory contour formation in stereopsis. *Psychological Review*, *102*, 705–743.

Anderson, B., & Winawer, J. (2005). Image segmentation and lightness perception. *Nature*, *434*, 79–83.

Anderson, B., O'Vari, J., & Barth, H. (2011). Non-bayesian contour synthesis. *Current Biology*, *21*(6), 492–496.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183–193.

Banton, T., & Levi, D. M. (1992). The perceived strength of illusory contours. *Perception & Psychophysics*, *52*(6), 676–684.

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication*. MIT Press.

Bartko, S. J., Winters, B. D., Cowell, R. A., Saksida, L. M., & Bussey, T. J. (2007). Perirhinal cortex resolves feature ambiguity in configural object recognition and perceptual oddity tasks. *Learning & Memory*, *14*(12), 821–832.

Beliy, R., Gaziv, G., Hoogi, A., Strappini, F., Golan, T., & Irani, M. (2019). From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, *37*(23), 3327–3338.

Belliveau, J. W., Kennedy, D. N., McKinstry, R. C., Buchbinder, B. R., Weisskoff, R. M., Cohen, M. S., Vevea, J., Brady, T. J., & Rosen, B. R. (1991). Functional mapping of the human visual cortex by magnetic resonance imaging. *Science*, *254*(5032), 716–719.

Bertalmío, M., Calatroni, L., Franceschi, V., Franceschiello, B., Gomez Villa, A., & Prandi, D. (2020). Visual illusions via neural dynamics: Wilson–cowan-type models and the efficient representation principle. *Journal of Neurophysiology*, *123*(5), 1606–1618.

Bindman, D., & Chubb, C. (2004). Brightness assimilation in bullseye displays. *Vision Research*, *44*(3), 309–319.

Blakeslee, B., & McCourt, M. E. (2004). A unified theory of brightness contrast and assimilation incorporating oriented multiscale spatial filtering and contrast normalization. *Vision Research*.

Blakeslee, B., Pasieka, W., & McCourt, M. E. (2005). Oriented multiscale spatial filtering and contrast normalization: A parsimonious model of brightness induction in a continuum of stimuli including white, howe and simultaneous brightness contrast. *Vision Research*, *45*(5), 607–615.

Boyaci, H., Fang, F., Murray, S., & Kersten, D. (2007). Responses to lightness variations in early human visual cortex. *Current Biology*, *17*, 989–993.

Boynton, R. M., & Gordon, J. (1965). Bezold–brücke hue shift measured by color-naming technique [© 1965 Optical Society of America]. *JOSA*, *55*(1), 78–86.

Braddick, O. (1995). Visual perception. seeing motion signals in noise. *Current Biology*, *5*(1), 7–9.

Brainard, D. H. (2006). Bayesian model of human color constancy. *Journal of Vision*.

Bressan, P. (1993). Neon colour spreading with and without its figural prerequisites. *Perception*, *22*(3), 353–361.

Bressan, P., Mingolla, E., Spillmann, L., & Watanabe, T. (1997). Neon color spreading: A review. *Perception*, *26*(11), 1353–1366.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1993). Responses of neurons in macaque MT to stochastic motion signals. *Visual Neuroscience*, *10*(6), 1157–1169.

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience*, *29*(44), 13992–14003.

Brown, H., & Friston, K. J. (2012). Free-energy and illusions: The cornsweet effect. *Frontiers in Psychology*, *3*.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLOS Computational Biology*, *10*(12), e1003963.

Carbon, C.-C. (2014). Understanding human perception by human-made illusions [Publisher: Frontiers]. *Frontiers in Human Neuroscience*, *8*.

Chalk, M., Marre, O., & Tkačik, G. (2018). Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences*, *115*(1), 186–191.

Chen, S., Glasauer, S., Muller, H. J., & Conci, M. (2018). Surface filling-in and contour interpolation contribute independently to Kanizsa figure formation. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(9), 1399–1413.

Cheng, F. L., Horikawa, T., Majima, K., Tanaka, M., Abdelhack, M., Aoki, S. C., Hirano, J., & Kamitani, Y. (2023). Reconstructing visual illusory experiences from human brain activity. *Science Advances*, *9*(46), eadj3906.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181–204.

Clifford, C. W. G., Wenderoth, P., & Spehar, B. (2000). A functional angle on some after-effects in cortical vision. *Proceedings of the Royal Society B: Biological Sciences*, *267*(1454), 1705–1710.

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8), 358–364.

Contini, E. W., Wardle, S. G., & Carlson, T. A. (2017). Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia*, *105*, 165–176.

Cornelissen, F. W., Wade, A. R., Vladusich, T., Dougherty, R. F., & Wandell, B. A. (2006). No functional magnetic resonance imaging evidence for brightness and color filling-in in early human visual cortex. *Journal of Neuroscience*, *26*(14), 3634–3641.

Corney, D., & Lotto, R. B. (2007). What are lightness illusions and why do we see them? [Publisher: Public Library of Science]. *PLOS Computational Biology*, *3*(9), e180.

Cowan, J. D., Neuman, J., & van Drongelen, W. (2016). Wilson–cowan equations for neocortical dynamics. *The Journal of Mathematical Neuroscience*, *6*(1), 1.

Cox, M. A., Schmid, M. C., Peters, A. J., Saunders, R. C., Leopold, D. A., & Maier, A. (2013). Receptive field focus of visual area v4 neurons determines responses to illusory surfaces. *Proceedings of the National Academy of Sciences*, *110*(42), 17095–17100.

Crick, F., & Koch, C. (1995). Are we aware of neural activity in primary visual cortex? *Nature*, *375*(6527), 121–123.

Darlington, T. R., Beck, J. M., & Lisberger, S. G. (2018). Neural implementation of bayesian inference in a sensory-motor behavior. *Nature Neuroscience*, *21*(10), 1442–1451.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural Computation*, *7*(5), 889–904.

De Weerd, P., Desimone, R., & Ungerleider, L. G. (1998). Perceptual filling-in: A parametric study. *Vision Research*, *38*(18), 2721–2734.

de Haas, B., & Schwarzkopf, D. S. (2018). Spatially selective responses to kanizsa and occlusion stimuli in human visual cortex. *Scientific Reports*, *8*(1), 611.

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis.

Doherty, M. J., Campbell, N. M., Tsuji, H., & Phillips, W. A. (2010). The ebbinghaus illusion deceives adults but not young children. *Developmental science*, *13*(5), 714–721.

Donderi, D., & Case, B. (1970). Parallel visual processing: Constant same-different decision latency with two to fourteen shapes. *Perception & Psychophysics*, *8*(5), 373–375.

Dormal, V., Larigaldie, N., Lefèvre, N., Pesenti, M., & Andres, M. (2018). Effect of perceived length on numerosity estimation: Evidence from the müller-lyer illusion. *Quarterly Journal of Experimental Psychology*, *71*(10), 2142–2151.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Dosovitskiy, A., & Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks.

Dresp, B., Lorenceau, J., & Bonnet, C. (1990). Apparent brightness enhancement in the kanizsa square with and without illusory contour formation. *Perception*, *19*(4), 483–489.

Dunn-Rankin, P. (2012). *Scaling methods*. Psychology Press.

Ebbinghaus, H. (1902). *The principles of psychology*. Veit.

Echeveste, R., Aitchison, L., Hennequin, G., & Lengyel, M. (2020). Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. *Nature Neuroscience*, *23*(9), 1138–1149.

Ehrenstein, W. (1941). Über abwandlungen der l. hermannschen helligkeitserscheinung. *Zeitschrift für Psychologie*, *150*, 83–91.

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E.-J., Shadlen, M. N., et al. (1994). Fmri of human visual cortex. *Nature*, *369*(6481), 525–525.

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

Fan, J., & Zeng, Y. (2023). Challenging deep learning models with image distortion based on the abutting grating illusion. *Patterns*, *4*(3), 100695.

Fang, F., Boyaci, H., & Kersten, D. (2009). Border ownership selectivity in human early visual cortex and its modulation by attention. *Journal of Neuroscience*, *29*(2), 460–465.

Fechner, G. T. (1860). Elements of psychophysics [Translated by Herbert Sidney Langfeld (1912)]. In *Elements of psychophysics*.

Fechner, G. T. (1889). *Elemente der psychophysik* (2nd ed., Vols. 2). Breitkopf & Härtel.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47.

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2012). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*, *15*(1), 146–154.

Ffytche, D. H. (2014). What does your theory of hallucinosis make of dream experience? do both reveal the operation of internal image generator mechanisms in the brain? In N. Tranquillo (Ed.), *Dream consciousness: Allan hobson's new approach to the brain and its mind* (pp. 133–136). Springer International Publishing.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences*, *14*(3), 119–130.

Fraser, J. (1908). A new visual illusion of direction. *British Journal of Psychology*, *2*, 307–320.

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

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*, 127–138.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211–1221.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–202.

Gallagher, S., Hutto, D., & Hipólito, I. (2022). Predictive processing and some disillusions about illusions. *Review of Philosophy and Psychology*, *13*(4), 999–1017.

Ganguli, D., & Simoncelli, E. P. (2014). Efficient sensory encoding and bayesian inference with heterogeneous neural populations. *Neural Computation*, *26*(10), 2103–2134.

Gaziv, G., Beliy, R., Granot, N., Hoogi, A., Strappini, F., Golan, T., & Irani, M. (2022). Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, *254*, 119121.

Gerardin, P., Abbatecola, C., Devinck, F., Kennedy, H., Dojat, M., & Knoblauch, K. (2018). Neural circuits for long-range color filling-in. *NeuroImage*, *181*, 30–43.

Gershman, S. J. (2019). The generative adversarial brain. *Frontiers in Artificial Intelligence*, *2*.

Gescheider, G. A. (1997). *Psychophysics: The fundamentals* (3rd). Lawrence Erlbaum Associates.

Gilchrist, A., Kossyfidis, C., Bonato, F., Agostini, T., Cataliotti, J., Li, X., Spehar, B., Annan, V., & Economou, E. (1999). An anchoring theory of lightness perception. *Psychological Review*, *106*(4), 795–834.

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932.

Gomez-Villa, A., Martín, A., Vazquez-Corral, J., Bertalmío, M., & Malo, J. (2020). Color illusions also deceive cnns for low-level vision tasks: Analysis and implications. *Vision Research*, *176*, 156–174.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*.

Gracely, R. H., McGrath, P., & Dubner, R. (1978). Ratio scales of sensory and affective verbal pain descriptors. *Pain*, *5*(1), 5–18.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. John Wiley.

Gregory, R. L. (2006). Bayes window (2). *Perception*, *35*(2), 143–144.

Gregory, R. L. (1997). *Eye and brain: The psychology of seeing* (5th). Princeton University Press.

Grosof, D. H., Shapley, R. M., & Hawken, M. J. (1993). Macaque vi neurons can signal 'illusory'contours. *Nature*, *365*(6446), 550–552.

Grossberg, S. (1994). 3-d vision and figure-ground separation by visual cortex. *Perception & Psychophysics*, *55*, 48–121.

Grossberg, S. (2017). The visual world as illusion: The ones we know and the ones we don't. In A. G. Shapiro & D. Todorović (Eds.), *The oxford compendium of visual illusions* (pp. 90–118). Oxford University Press.

Grossberg, S., & Hong, S. (2006). A neural model of surface perception: Lightness, anchoring, and filling-in. *Spatial Vision*, *19*(2-4), 263–321.

Grossberg, S., & Mingolla, E. (1985). Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. *Psychological Review*, *92*(2), 173–211.

Grossberg, S., & Raizada, R. D. S. (2000). Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex. *Vision Research*, *40*(10), 1413–1432.

Grossberg, S., & Todorovic, D. (1988). Neural dynamics of 1-d and 2-d brightness perception: A unified model of classical and recent phenomena. *Perception & Psychophysics*, *43*(3), 241–277.

Grossberg, S., & Yazdanbakhsh, A. (2005). Laminar cortical dynamics of 3d surface perception: Stratification, transparency, and neon color spreading. *Vision Research*, *45*(13), 1725–1743.

Güçlü, U., & Gerven, M. A. J. v. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., & van Gerven, M. A. J. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in Neural Information Processing Systems*.

Hahn, M., & Wei, X.-X. (2024). A unifying theory explains seemingly contradictory biases in perceptual estimation. *Nature Neuroscience*, *27*, 793–804.

Halgren, E., Mendola, J., Chong, C. D., & Dale, A. M. (2003). Cortical activation to illusory shapes as measured with magnetoencephalography. *NeuroImage*, *18*(4), 1001–1009.

Harrison, W. J., Bays, P. M., & Rideaux, R. (2023). Neural tuning instantiates prior expectations in the human visual system. *Nature Communications*, *14*(1), 5320.

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.

Haynes, J.-D. (2009). Decoding visual consciousness from human brain signals [Publisher: Elsevier]. *Trends in Cognitive Sciences*, *13*(5), 194–202.

Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: Principles, pitfalls, and perspectives. *Neuron*, *87*(2), 257–270.

He, D., Mo, C., Wang, Y., & Fang, F. (2015). Position shifts of fmri-based population receptive fields in human visual cortex induced by ponzo illusion. *Experimental brain research*, *233*, 3535–3541.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Heitger, F., von der Heydt, R., & Kubler, O. (1994). A computational model of neural contour processing: Figure-ground segregation and illusory contours. *From Perception to Action Proceedings of PerAc '94*, 181–192.

Hering, E. (1861). *Beiträge zur physiologie. i. zur lehre vom ortssinne der netzhaut*. Engelmann.

Herrmann, C. S., & Bosch, V. (2001). Gestalt perception modulates early visual processing. *NeuroReport*, *12*(5), 901.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *arXiv preprint*.

Ho, J. K., Horikawa, T., Majima, K., Cheng, F., & Kamitani, Y. (2023). Inter-individual deep image reconstruction via hierarchical neural code conversion. *NeuroImage*, *271*, 120007.

Hong, S. W., & Tong, F. (2017). Neural representation of form-contingent color filling-in in the early visual cortex. *Journal of Vision*, *17*(13), 10.

Horikawa, T., Tamaki, M., Miyawaki, Y., & Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science*, *340*(6132), 639–642.

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*(1), 15037.

Horikawa, T., & Kamitani, Y. (2022). Attention modulates neural representation to render reconstructions according to subjective appearance. *Communications Biology*, *5*(1), 1–12.

Hou, H., Zheng, Q., Zhao, Y., Pouget, A., & Gu, Y. (2019). Neural correlates of optimal multisensory decision making under time-varying reliabilities with an invariant linear probabilistic population code. *Neuron*, *104*(5), 1010–1021.e10.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Huang, G., Chen, D., Li, T., Wu, F., van der Maaten, L., & Weinberger, K. (2018). Multi-scale dense networks for resource efficient image classification. *International Conference on Learning Representations*.

Huang, L., Wang, L., Shen, W., Li, M., Wang, S., Wang, X., Ungerleider, L. G., & Zhang, X. (2020). A source for awareness-dependent figure–ground segregation in human prefrontal cortex. *Proceedings of the National Academy of Sciences*, *117*(48), 30836–30847.

Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580–593.

Hubel, D., Wiesel, T., & STRYKER, M. (1977). Orientation columns in macaque monkey visual cortex demonstrated by the 2-deoxyglucose autoradiographic technique. *Nature*, *269*, 328–330.

Iversen, S. D., & Humphrey, N. (1971). Ventral temporal lobe lesions and visual oddity performance. *Brain Research*, *30*(2), 253–263.

Jafari-Khouzani, K., & Soltanian-Zadeh, H. (2005). Radon transform orientation estimation for rotation invariant texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*, 1004–1008.

Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision*, *6*(11), 13.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*(8), 1020–1026.

Jiahui, G., Feilong, M., Nastase, S. A., Haxby, J. V., & Gobbini, M. I. (2023). Cross-movie prediction of individualized functional topography. *eLife*, *12*, e86037.

Jiang, L. P., & Rao, R. P. N. (2024). Dynamic predictive coding: A model of hierarchical sequence learning and prediction in the neocortex. *PLOS Computational Biology*, *20*(2), 1–30.

Kalar, D. J., Garrigan, P., Wickens, T. D., Hilger, J. D., & Kellman, P. J. (2010). A unified model of illusory and occluded contour interpolation. *Vision Research*, *50*(3), 284–299.

Kamitani, Y., & Shimojo, S. (2003). Global yet early processing of visual surfaces. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 1129–1138). MIT Press.

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

Kanizsa, G. (1955). Margini quasi-percettivi in campi con stimolazione omogenea. *Rivista di Psicologia*, *49*(1), 7–30.

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.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology*, *55*(1), 271–304.

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

Kimchi, R., & Peterson, M. A. (2008). Figure-ground segmentation can occur without attention. *Psychological Science*, *19*(7), 660–668.

Kingdom, F. A. A., & Prins, N. (2016). *Psychophysics: A practical introduction* [Google-Books-ID: 3sHQBAAAQBAJ]. Academic Press.

Kirubeswaran, O. R., & Storrs, K. R. (2023). Inconsistent illusory motion in predictive coding deep neural networks. *Vision Research*, *206*, 108195.

Kitaoka, A., & Ashida, H. (2003). Phenomenal characteristics of the peripheral drift illusion. *Vision*, *15*, 261–262.

Kitaoka, A., & Ishihara, M. (2000). Three elemental illusions determine theZöllner illusion. *Perception & Psychophysics*, *62*(3), 569–575.

Knebel, J.-F., & Murray, M. M. (2012a). Towards a resolution of conflicting models of illusory contour processing in humans. *NeuroImage*, *59*(3), 2808–2817.

Knebel, J.-F., & Murray, M. M. (2012b). Towards a resolution of conflicting models of illusory contour processing in humans. *Neuroimage*, *59*(3), 2808–2817.

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.

Kobayashi, T., Kitaoka, A., Kosaka, M., Tanaka, K., & Watanabe, E. (2022). Motion illusion-like patterns extracted from photo and art images using predictive deep neural networks [2022 The Author(s)]. *Scientific Reports*, *12*(1), 3893.

Kogo, N., Strecha, C., Van Gool, L., & Wagemans, J. (2010). Surface construction by a 2-d differentiation-integration process: A neurocomputational model for perceived border ownership, depth, and lightness in kanizsa figures. *Psychological Review*, *117*(2), 406–439.

Köhler, W., & Fishback, J. (1950). The destruction of the müller-lyer illusion in repeated trials: I. an examination of two theories. *Journal of experimental psychology*, *40*(2), 267.

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.

Kok, P., & de Lange, F. P. (2014). Shape perception simultaneously up- and downregulates neural activity in the primary visual cortex. *Current Biology*, *24*(13), 1531–1535.

Kok, P., & de Lange, F. P. (2015). Predictive coding in sensory cortex. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 221–244). Springer.

Komatsu, H. (2006). The neural mechanisms of perceptual filling-in. *Nature Reviews Neuroscience*, *7*(3), 220–231.

Kourtzi, Z., & Kanwisher, N. (2000). Cortical regions involved in perceiving object shape. *Journal of Neuroscience*, *20*(9), 3310–3318.

Kovács, I., & Julesz, B. (1993). A closed curve is much more than an incomplete one: Effect of closure in figure-ground segmentation. *Proceedings of the National Academy of Sciences*, *90*(16), 7495–7497.

Kriegeskorte, N., & Kreiman, G. (Eds.). (2011). *Visual population codes: Toward a common multivariate framework for cell recording and functional imaging*. The MIT Press.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*.

Króliczak, G., Heard, P., Goodale, M. A., & Gregory, R. L. (2006). Dissociation of perception and action unmasked by the hollow-face illusion. *Brain Research*, *1080*(1), 9–16.

Kusunoki, M., Moutoussis, K., & Zeki, S. (2006). Effect of background colors on the tuning of color-selective cells in monkey area v4. *Journal of Neurophysiology*, *95*(5), 3047–3059.

LARSON-POWERS, N., & Pangborn, R. M. (1978). Paired comparison and time-intensity measurements of the sensory properties of beverages and gelatins containing sucrose or synthetic sweeteners. *Journal of Food Science*, *43*(1), 41–46.

Larsson, J., & Amunts, K. (1999). Neuronal correlates of real and illusory contour perception: Functional anatomy with PET. *European Journal of Neuroscience*.

Layton, O. W., Mingolla, E., & Yazdanbakhsh, A. (2012). Dynamic coding of border-ownership in visual cortex. *Journal of Vision*, *12*(13), 8.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.

Lee, A. C., Yeung, L.-K., & Barense, M. D. (2012). The hippocampus and visual perception. *Frontiers in human neuroscience*, *6*, 91.

Lee, T. S., & Mumford, D. B. (2003). Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*(7), 1434–1448.

Lehar, S. (2003). Directional harmonic theory: A computational gestalt model to account for illusory contour and vertex formation [Publisher: SAGE Publications Ltd STM]. *Perception*, *32*(4), 423–448.

Levinson, M., & Baillet, S. (2022). Perceptual filling-in dispels the veridicality problem of conscious perception research. *Consciousness and Cognition*, *100*, 103316.

Liang, Y., Zhang, M., & Browne, W. (2015). A supervised figure-ground segmentation method using genetic programming, 491–503.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2014). Microsoft COCO: Common objects in context [Computer Vision and Pattern Recognition]. *arXiv preprint arXiv:1405.0312*.

Lindsey, D. T., & Brown, A. M. (2014). The color lexicon of american english. *Journal of Vision*, *14*(2), 17.

Lotter, W., Kreiman, G., & Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception [2020 The Author(s), under exclusive licence to Springer Nature Limited]. *Nature Machine Intelligence*, *2*(4), 210–219.

Lu, Z.-L., & Dosher, B. (2013). *Visual psychophysics: From laboratory to theory*. MIT Press.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

Maertens, M., Pollmann, S., Hanke, M., Mildner, T., & Möller, H. E. (2008). Retinotopic activation in response to subjective contours in primary visual cortex. *Frontiers in Human Neuroscience*, *2*.

Manjunath, B., & Chellappa, R. (1993). A unified approach to boundary perception: Edges, textures, and illusory contours. *IEEE Transactions on Neural Networks*, *4*(1), 96–108.

Manning, C., Morgan, M. J., Allen, C. T., & Pellicano, E. (2017). Susceptibility to ebbinghaus and müller-lyer illusions in autistic children: A comparison of three different methods. *Molecular autism*, *8*, 1–18.

Manookin, M. B., & Rieke, F. (2023). Two sides of the same coin: Efficient and predictive neural coding. *Annual Review of Vision Science*, *9*, 293–311.

Marlow, P. J., Kim, J., & Anderson, B. L. (2017). Perception and misperception of surface opacity. *Proceedings of the National Academy of Sciences*, *114*(52), 13840–13845.

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

Masuda, T., Matsubara, K., Utsumi, K., & Wada, Y. (2015). Material perception of a kinetic illusory object with amplitude and frequency changes in oscillated inducer motion. *Vision Research*, *109*, 201–208.

Mattingley, J. B., Davis, G., & Driver, J. (1997). Preattentive filling-in of visual surfaces in parietal extinction. *Science*, *275*(5300), 671–674.

Mendola, J. D., Dale, A. M., Fischl, B., Liu, A. K., & Tootell, R. B. H. (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.

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.

Metelli, F. (1985). Stimulation and perception of transparency. *Psychological Research*, *47*(4), 185–202.

Miller, J., & Bauer, D. W. (1981). Irrelevant differences in the "same"-"different" task. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(1), 196–207.

Millidge, B., Seth, A., & Buckley, C. L. (2022). *Predictive coding: A theoretical and experimental review* (arXiv:2107.12979). arXiv.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-a., Morito, Y., Tanabe, H. C., Sadato, N., & Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders [Publisher: Elsevier]. *Neuron*, *60*(5), 915–929.

Morgan, M., & Casco, C. (1990). Spatial filtering and spatial primitives in early vision: An explanation of the zöllner–judd class of geometrical illusion. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *242*(1303), 1–10.

Mukamel, R., & Fried, I. (2012). Human intracranial recordings and cognitive neuroscience. *Annual Review of Psychology*, *63*, 511–537.

Müller-Lyer, F. C. (1889). Optische urteilstäuschungen. *Archiv für Physiologie Suppl.*, 263–270.

Munker, H. (1970). *Farbige gitter, abbildung auf der netzhaut und übertragungstheoretische beschreibung der farbwahrnehmung* [Habilitationsschrift]. Ludwig-Maximilians-Universität.

Münsterberg, H. (1894). *Pseudoptics*. Milton Bradley.

Murray, M. M., Foxe, D. M., Javitt, D. C., & Foxe, J. J. (2004). Setting boundaries: Brain dynamics of modal and amodal illusory shape completion in humans. *Journal of Neuroscience*, *24*(31), 6898–6903.

Murray, M. M., & Herrmann, C. S. (2013). Illusory contours: A window onto the neurophysiology of constructing perception [Publisher: Elsevier]. *Trends in Cognitive Sciences*, *17*(9), 471–481.

Murray, M. M., Wylie, G. R., Higgins, B. A., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). The spatiotemporal dynamics of illusory contour processing: Combined high-density electrical mapping, source analysis, and functional magnetic resonance imaging. *Journal of Neuroscience*, *22*(12), 5055–5073.

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.

Nakayama, K., Shimojo, S., & Ramachandran, V. S. (1990). Transparency: Relation to depth, subjective contours, luminance, and neon color spreading. *Perception*, *19*(4), 497–513.

Narain, D., Remington, E. D., De Zeeuw, C. I., & Jazayeri, M. (2018). A cerebellar mechanism for learning prior distributions of time intervals. *Nature Communications*, *9*, 469.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fmri. *NeuroImage*, *56*(2), 400–410.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*(6), 902–915.

Nere, A., Olcese, U., Balduzzi, D., & Tononi, G. (2012). A neuromorphic architecture for object recognition and motion anticipation using burst-STDP. *PLOS ONE*, *7*(5), 1–17.

Nessler, B., Pfeiffer, M., Buesing, L., & Maass, W. (2013). Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. *PLOS Computational Biology*, *9*(4), e1003037.

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature*, *341*(6237), 52–54.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641–1646.

Nonaka, S., Majima, K., Aoki, S. C., & Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, *24*(9), 103013.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.

Nour, M. M., & Nour, J. M. (2015). Perception, illusions and bayesian inference. *Psychopathology*, *48*(4), 217–221.

Ogawa, S., Lee, T.-M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, *87*(24), 9868–9872.

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

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, *37*(23), 3311–3325.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*(4), 481–487.

Olshausen, B. A., & Field, D. J. (2005). How close are we to understanding v1? *Neural Computation*, *17*(8), 1665–1699.

Otazu, X., Vanrell, M., & Párraga, C. A. (2008). Multiresolution wavelet framework models brightness induction effects. *Vision Research*, *48*(5), 733–751.

Ozcelik, F., & VanRullen, R. (2023). Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, *13*(1), 15666.

Pak, A., Ryu, E., Li, C., & Chubykin, A. A. (2020). Top-down feedback controls the cortical representation of illusory contours in mouse primary visual cortex. *Journal of Neuroscience*, *40*(3), 648–660.

Pan, Y., Chen, M., Yin, J., An, X., Zhang, X., Lu, Y., Gong, H., Li, W., & Wang, W. (2012a). Equivalent representation of real and illusory contours in macaque v4. *Journal of Neuroscience*, *32*(20), 6760–6770.

Pan, Y., Chen, M., Yin, J., An, X., Zhang, X., Lu, Y., Gong, H., Li, W., & Wang, W. (2012b). Equivalent representation of real and illusory contours in macaque v4. *Journal of Neuroscience*, *32*(20), 6760–6770.

Pang, Z., O'May, C. B., Choksi, B., & VanRullen, R. (2021). Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *Neural Networks*, *144*, 164–175.

Park, I. M., & Pillow, J. W. (2024). Bayesian efficient coding. *bioRxiv*.

Pearson, J. (2019). The human imagination: The cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, *20*(10), 624–634.

Peterhans, E., & von der Heydt, R. (1989). Mechanisms of contour perception in monkey visual cortex. II. contours bridging gaps. *Journal of Neuroscience*, *9*(5), 1749–1763.

Peters, J. C., Jans, B., van de Ven, V., De Weerd, P., & Goebel, R. (2010). Dynamic brightness induction in v1: Analyzing simulated and empirically acquired fmri data in a "common brain space" framework [Computational Models of the Brain]. *NeuroImage*, *52*(3), 973–984.

Pinna, B., Brelstaff, G., & Spillmann, L. (2001). Surface color from boundaries: A new 'watercolor' illusion. *Vision Research*, *41*(20), 2669–2676.

Pinna, B., & Grossberg, S. (2005). The watercolor illusion and neon color spreading: A unified analysis of new cases and neural mechanisms. *Journal of the Optical Society of America A*, *22*(10), 2207–2221.

Pollack, R. H., & Jaeger, T. B. (1991). The effect of lightness contrast on the colored müller-lyer illusion. *Perception & psychophysics*, *50*, 225–229.

Ponzo, M. (1911). Intorno ad alcune illusioni nel campo delle sensazioni tattili sull'illusione di aristotele e fenomeni analoghi. *Archives Italiennes de Biologie*.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

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

Prinzmetal, W. (1981). Principles of feature integration in visual perception. *Perception & Psychophysics*, *30*(4), 330–340.

Radon, J. (1917). Über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten. *Berichte über die Verhandlungen der Königlich-Sächsischen Akademie der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse*, (69), 262–277.

Ramsden, B. M., Hung, C. P., & Roe, A. W. (2001). Real and illusory contour processing in area v1 of the primate: A cortical balancing act. *Cerebral cortex*, *11*(7), 648–665.

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

Rao, R. P., Gklezakos, D. C., & Sathish, V. (2023). Active predictive coding: A unifying neural model for active perception, compositional learning, and hierarchical planning. *Neural Computation*, *36*(1), 1–32.

Redies, C., Spillmann, L., & Kunz, K. (1984). Colored neon flanks and line gap enhancement. *Vision Research*, *24*(10), 1301–1309.

Redies, C., & Spillmann, L. (1981). The neon color effect in the Ehrenstein illusion. *Perception*, *10*(6), 667–681.

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

Robinson, A. E., Hammon, P. S., & de Sa, V. R. (2007). A filtering model of brightness perception using frequency-specific locally-normalized oriented difference-of-gaussians (FLODOG). *Journal of Vision*, *7*(9), 237.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Saeedi, A., Wang, K., Nikpourian, G., Bartels, A., Totah, N. K., Logothetis, N. K., & Watanabe, M. (2022). Mouse primary visual cortex neurons respond to the illusory "darker than black" in neon color spreading. *bioRxiv*.

Sanger, T. D. (1996). Probability density estimation for the interpretation of neural population codes. *Journal of Neurophysiology*, *76*(4), 2790–2793.

Sáry, G., Chadaide, Z., Tompa, T., Köteles, K., Kovács, G., & Benedek, G. (2007). Illusory shape representation in the monkey inferior temporal cortex. *European Journal of Neuroscience*, *25*(8), 2558–2564.

Sáry, G., Köteles, K., Kaposvári, P., Lenti, L., Csifcsák, G., Frankó, E., Benedek, G., & Tompa, T. (2008). The representation of Kanizsa illusory contours in the monkey inferior temporal cortex. *European Journal of Neuroscience*, *28*(10), 2137–2146.

Schlauch, R. S., & Rose, R. M. (1990). Two-, three-, and four-interval forced-choice staircase procedures: Estimator bias and efficiency. *Journal of the Acoustical Society of America*, *88*(2), 732–740.

Schrauf, M., Lingelbach, B., & WIST, E. R. (1997). The scintillating grid illusion. *Vision research*, *37*(8), 1033–1038.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Schumann, F. (1900). Beiträge zur analyse der gesichtswahrnehmungen. erste abhandlung. einige beobachtungen über die zusammenfassung von gesichtseindrücken zu einheiten. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, *23*, 1–32.

Schwartz, O., Hsu, A., & Dayan, P. (2007). Space and time in visual context. *Nature Reviews Neuroscience*.

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., & van Gerven, M. a. J. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, *181*, 775–785.

Seghier, M. L., & Vuilleumier, P. (2006). Functional neuroimaging findings on the human perception of illusory contours. *Neuroscience and Biobehavioral Reviews*, *30*(5), 595–612.

Sereno, M. I., Dale, A., Reppas, J., Kwong, K., Belliveau, J., Brady, T., Rosen, B., & Tootell, R. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, *268*(5212), 889–893.

Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, *8*(28), eabm2219.

Shapiro, A. G., & Todorović, D. (Eds.). (2017). *The oxford compendium of visual illusions*. Oxford University Press.

Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of Vision*, *14*, 12.

Shen, G., Dwivedi, K., Majima, K., Horikawa, T., & Kamitani, Y. (2019b). End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, *13*.

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019a). Deep image reconstruction from human brain activity [Publisher: Public Library of Science]. *PLOS Computational Biology*, *15*(1), e1006633.

Shimojo, S., Kamitani, Y., & Nishida, S. (2001). Afterimage of perceptually filled-in surface. *Science*, *293*(5535), 1677–1680.

Shine, J. M., Halliday, G. M., Naismith, S. L., & Lewis, S. J. G. (2011). Visual misperceptions and hallucinations in parkinson's disease: Dysfunction of attentional control networks? *Movement Disorders*, *26*(12), 2154–2159.

Shipley, T. F., & Kellman, P. J. (1992). Perception of partly occluded objects and illusory figures: Evidence for an identity hypothesis. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 106–120.

Shipp, S. (2016). Neural elements for predictive coding. *Frontiers in Psychology*, *7*.

Shipp, S. (2024). Computational components of visual predictive coding circuitry. *Frontiers in Neural Circuits*, *17*.

Shirakawa, K., Tanaka, M., Aoki, S. C., Majima, K., & Kamitani, Y. (2023). Critical assessment of generative ai methods and natural image datasets for visual image reconstruction from brain activity [Retrieved from osf.io/nmfc5].

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1216.

Simoncelli, E. P. (2009). Optimal estimation in sensory systems. In *The cognitive neurosciences, iv* (pp. 525–535).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations (ICLR 2015)*, 1–14.

Singletary, N. M., Gottlieb, J., & Horga, G. (2024). The parieto-occipital cortex is a candidate neural substrate for the human ability to approximate bayesian inference. *Communications Biology*, *7*(1), 1–18.

Sohn, H. (2021). Neural implementations of bayesian inference. *Current Opinion in Neurobiology*.

Song, J.-H., & Nakayama, K. (2008). Target selection in visual search as revealed by movement trajectories. *Vision research*, *48*(7), 853–861.

Soriano, M., Spillmann, L., & Bach, M. (1996). The abutting grating illusion. *Vision Research*, *36*(1), 109–116.

Spillmann, L., Fuld, K., & Neumeyer, C. (1984). Brightness matching, brightness cancellation, and increment threshold in the ehrenstein illusion. *Perception*, *13*(5), 513–520.

Sprevak, M. (2024). Predictive coding i: Introduction. *Philosophy Compass*, *19*(1), e12950.

Stanley, D. A., & Rubin, N. (2003). fMRI activation in response to illusory contours and salient regions in the human lateral occipital complex. *Neuron*, *37*(2), 323–331.

Stanley, G. B., Li, F. F., & Dan, Y. (1999). Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *The Journal of Neuroscience*, *19*(18), 8036–8042.

Stevens, J. C., & Marks, L. E. (1965). Cross-modality matching of brightness and loudness [Publisher: National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *54*(2), 407–411.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. John Wiley & Sons.

Stevens, S. S., & Guirao, M. (1963). Subjective scaling of length and area and the matching of length to loudness and brightness [Publisher: American Psychological Association]. *Journal of Experimental Psychology*, *66*(2), 177–186.

Sun, E. D., & Dekel, R. (2021). Imagenet-trained deep neural networks exhibit illusion-like response to the scintillating grid. *Journal of Vision*, *21*(11), 15.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826.

Tadin, D., Park, W. J., Dieter, K. C., Mednick, S. C., Wede, D., & Gibson, D. (2019). Spatial suppression promotes rapid figure-ground segmentation of moving objects. *Nature Communications*, *10*, 2732.

Tallon-Baudry, C., Bertrand, O., Delpuech, C., & Pernier, J. (1996). Stimulus specificity of phase-locked and non-phase-locked 40 hz visual responses in human. *The Journal of Neuroscience*, *16*(13), 4240–4249.

Tan, M., & Le, Q. V. (2020). EfficientNet: Rethinking model scaling for convolutional neural networks.

Taylor, M. M., & Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. *The Journal of the Acoustical Society of America*, *41*(4), 782–787.

Teufel, C., & Fletcher, P. C. (2020). Forms of prediction in the nervous system. *Nature Reviews. Neuroscience*, *21*(4), 231–242.

Thurstone, L. L. (1927). A law of comparative judgment [Publisher: Psychological Review Company]. *Psychological Review*, *34*(4), 273–286.

Torgerson, W. S. (1958). *Theory and methods of scaling*. Wiley.

Troscianko, J., & Osorio, D. (2023). A model of colour appearance based on efficient coding of natural images. *PLOS Computational Biology*, *19*(6), e1011117.

Uka, T., & DeAngelis, G. C. (2003). Contribution of middle temporal area to coarse depth discrimination: Comparison of neuronal and psychophysical sensitivity. *Journal of Neuroscience*, *23*(8), 3515–3530.

van Heusden, E., Harris, A. M., Garrido, M. I., & Hogendoorn, H. (2019). Predictive coding of visual motion in both monocular and binocular human visual processing. *Journal of Vision*, *19*(1), 3.

van Tuijl, H. F. J. M. (1975). A new visual illusion: Neonlike color spreading and complementary color induction between subjective contours. *Acta Psychologica*, *39*(6), 441–IN1.

Victor, J. D., & Conte, M. M. (2012). Local image statistics: Maximum-entropy constructions and perceptual salience [Publisher: Optica Publishing Group]. *JOSA A*, *29*(7), 1313–1345.

Vogels, R., & Orban, G. A. (1990). How well do response changes of striate neurons signal differences in orientation: A study in the discriminating monkey. *Journal of Neuroscience*, *10*(11), 3543–3558.

von der Heydt, R., & Peterhans, E. (1989). Mechanisms of contour perception in monkey visual cortex. I. Lines of pattern discontinuity. *Journal of Neuroscience*, *9*(5), 1731–1748.

von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science*, *224*(4654), 1260–1262.

Walker, E. Y. (2020). A neural basis of probabilistic computation in visual cortex. *Nature Neuroscience*, *23*.

Walker, E. Y. (2023). Studying the neural representations of uncertainty. *Nature Neuroscience*, *26*, Article 1444.

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., & Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in Psychology*, *9*.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A bayesian adaptive psychometric method [Publisher: Psychonomic Society]. *Perception & Psychophysics*, *33*(2), 113–120.

Weber, E. H. (1834). *De pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae [on pulse, breathing, hearing, and touch: Anatomical and physiological notes]*. C. F. Koehler.

Wei, X.-X., & Stocker, A. A. (2012). Efficient coding provides a direct link between prior and likelihood in perceptual bayesian inference. *Advances in Neural Information Processing Systems*, *25*.

Wei, X.-X., & Stocker, A. A. (2015). A bayesian observer model constrained by efficient coding can explain 'anti-bayesian' percepts. *Nature Neuroscience*, *18*(10).

Weidner, R., & Fink, G. R. (2007). The neural mechanisms underlying the Müller-Lyer illusion and its interaction with visuospatial judgments. *Cerebral Cortex*, *17*(4), 878–884.

Weintraub, D. J., & Krantz, D. H. (1971). The poggendorff illusion: Amputations, rotations, and other perturbations. *Perception & Psychophysics*, *10*(4), 257–264.

Weintraub, D. J., Krantz, D. H., & Olson, T. P. (1980). The poggendorff illusion: Consider all the angles. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(4), 718–725.

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, *28*(12), 4136–4160.

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

Westheimer, G. (2008). Illusions in the spatial sense of the eye: Geometrical–optical illusions and the neural representation of space. *Vision Research*, *48*(20), 2128–2142.

White, M. (1979). A new effect of pattern on perceived lightness. *Perception*, *8*, 413–416.

Williams, L. R., & Jacobs, D. W. (1997). Stochastic completion fields: A neural model of illusory contour shape and salience. *Neural Computation*, *9*(4), 837–858.

Wolfe, J. M., Kluender, K. R., Levi, D. M., Bartoshuk, L. M., Herz, R. S., Klatzky, R. L., Lederman, S. J., & Merfeld, D. M. (2006). *Sensation & perception*. Sinauer Sunderland, MA.

Worthington, A. (1969). Paired comparison scaling of brightness judgements: A method for the measurement of perceptual defence. *British Journal of Psychology*, *60*(3), 363–368.

Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications*, *12*(1), 2065.

Yamins, D. L. K., 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*, *111*(23), 8619–8624.

Yan, C., Pérez-Bellido, A., & de Lange, F. P. (2021). Amodal completion instead of predictive coding can explain activity suppression of early visual cortex during illusory shape perception. *Journal of Vision*, *21*(5), 13.

Zarandy, A., Stoffels, A., Roska, T., & Chua, L. (1998). Implementation of binary and gray-scale mathematical morphology on the CNN universal machine. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, *45*(2), 163–168.

Zeki, S. (1983a). Colour coding in the cerebral cortex: The reaction of cells in monkey visual cortex to wavelengths and colours. *Neuroscience*, *9*, 741–765.

Zeki, S. (1983b). Colour coding in the cerebral cortex: The responses of wavelength-selective and colour-coded cells in monkey visual cortex to changes in wavelength composition. *Neuroscience*, *9*, 767–781.

Zhaoping, L. (2005). Border ownership from intracortical interactions in visual area v2. *Neuron*, *47*(1), 143–153.

Zhaoping, L., & Jingling, L. (2008). Filling-in and suppression of visual perception from context: A bayesian account of perceptual biases by contextual influences. *PLOS Computational Biology*, *4*(2), e14.

Zhou, H., Friedman, H. S., & von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *Journal of Neuroscience*, *20*(17), 6594–6611.

Zöllner, F. (1860). Ueber eine neue art von pseudoskopie und ihre beziehungen zu den von plateau und oppel beschrieben bewegungsphaenomenen. *Annalen der Physik*, *186*, 500–525.

Appendix A

Publications

The current work has yielded the following publications and presentation:

A.1 Manuscript

- Cheng, F., Horikawa, T., Majima, K., Tanaka, M., Abdelhack, M., Aoki, S. C., Hirano, J., & Kamitani, Y. (2023). Reconstructing visual illusory experiences from human brain activity. *Science Advances*, 9 (46), eadj3906.

- Ho, J. K., Horikawa, T., Majima, K., Cheng, F., & Kamitani, Y. (2023). Inter-individual deep image reconstruction via hierarchical neural code conversion. *NeuroImage*, 271, 120007.

A.2 Presentation

- Cheng, F., Horikawa, T., Majima, K., Tanaka, M., Abdelhack, M., Aoki, S. C., Hirano, J., & Kamitani, Y. (2024). Deciphering visual representations behind subjective perception using reconstruction methods. Oral presentation for *VSS 2024*, Florida.

- Wang, H., Ho, J. K., Cheng, F., Aoki S. C., & Kamitani, Y. (2023). Inter-individual neural code conversion without paired stimuli. Poster presentation for *Conference on Cognitive Computational Neuroscience*, Oxford.

- Cheng, F., Horikawa, T., Majima, K., & Kamitani, Y. (2022). Reconstruction of line illusion from human brain activity. Poster presentation for *2022 Conference on Cognitive Computational Neuroscience*, San Francisco.

Appendix B

Code availability

The experimental code that support the findings in this thesis is available from the repository:

- Code for illusion reconstruction and evaluation: https://github.com/KamitaniLab/IllusionReconstruction
- Code for Pixel optimization (iCNN) method: https://github.com/KamitaniLab/DeepImageReconstruction