Distinguishing mirror from glass by the human visual system and its modelling

（ヒト視覚系による鏡・ガラス材質の識別とそのモデリング）

January 2019

Doctor of Philosophy (Engineering)

Hideki Tamura

田村　秀希

Toyohashi University of Technology

別紙４－１（課程博士（英文））

Date of Submission（month day, year）: January 11th, 2019

Department of Computer Science and Engineering	Student ID Number	D123352	Supervisors
Applicant's name	Hideki Tamura		

Supervisors	Shigeki Nakauchi Michiteru Kitazaki

Abstract (Doctor)

Title of Thesis	Distinguishing mirror from glass by the human visual system and its modelling (ヒト視覚系による鏡・ガラス材質の識別とそのモデリング)

Approx. 800 words

The human visual system effortlessly recognizes various objects made from many kinds of materials, such as steel, wood, and plastic. It is easy to infer their physical, functional, and multisensory properties at a glance. This ability, which we use involuntarily, is called "material perception," and is broadly studied in various research fields to understand an important aspect of the visual system. One challenging case is to distinguish a mirror (a perfect specular surface such as polished metal) from glass (a transparent and refractive medium) because their appearances are totally derived from their surroundings. Just changing the surroundings or the object shape dramatically alters the image. Thus, discrimination is hard and complex, and remains poorly understood. In this thesis, we investigated how the visual system distinguishes mirrors from glass materials, and clarified what visual cues contribute to this task.

First, we developed various models (classifiers), which were designed to mimic the visual system and trained to distinguish mirror from glass using over 750,000 images rendered by computer graphics. Then, we compared the performance of humans and the models, including thousands of feedforward neural networks and other models based on "hand-engineered" image features. For randomly selected images, humans and all models performed with high accuracy, and therefore correlated highly with one another. To tease the models apart, a series of human psychophysics defined "diagnostic" images were used to decouple the true material class and the class perceived by humans. We used these images for a large-scale and organized optimization to find a neural network that behaves like humans do. The best network was relatively shallow, and none of the models correlated better than 0.6 with human responses despite an extensive and systematic search. These findings imply the existence of three important gaps between humans and the networks, such as the feedback architecture, training objective, and task generality.

Next, we clarified visual static and dynamic cues that contribute to distinguishing mirror from glass as image and video stimuli, respectively. A new image editing method that we proposed modulates the luminance and color saturation profiles along the trajectory from the object contour to its center. These two kinds of pixel information altered material appearances between the mirror and glass in each other, suggesting that they contributed as static cues. Additionally, as dynamic cues, we found that the motion ratio between the direction of object rotation and its opposite direction determined the extent of material appearance between transparent and specular reflective objects. Our model based on dynamic cues sufficiently identified the different materials.

Moreover, we tested two optical illusions involved in mirror and glass to understand material perception in the aspect of a bridge between a physical property and our perception. We focused on the glare illusion to test the relationship between brightness enhancement and self-luminosity related to specular highlights of the materials. This was robust across stimulus intensities ranging from dark to light with subjective gray, white, and luminous appearances. We also discovered a new illusion in which a rotating glass prism was perceived as being made of a mirror, and simultaneously, its direction of rotation was also misperceived. This finding suggests that the interaction between shape, surface properties, and illumination strongly affects our material and motion judgements.

Even though we only focused on two specific materials that are common and basic, they have entirely different physical properties and enrich our lives. Our approach can expand the possibilities for other materials or optical properties to help us understand the visual system more deeply. Thus, this thesis clarified various aspects of distinguishing mirror from glass and provided further related challenges.

Acknowledgments

I would like to thank my supervisor Shigeki Nakauchi for many discussions, vast knowledge, and a good work atmosphere, and for giving me a chance to work at the Visual Perception and Cognition Laboratory. From my undergraduate days, I spent seven years in his great lab, and I greatly treasure this period of my life.

I would like to thank my external supervisors: Roland W. Fleming for his warm and kind mentoring and for giving me great opportunities to work at Giessen for nine months in total; Hidehiko Komatsu and Shin'ya Nishida for their fruitful comments and discussions in conferences, my examinations of promotion, and my internship.

I would like to thank our lab's staff, Tetsuto Minami, Yuki Kawai, and ex-lab staff, Kowa Koida, Hiroshi Higashi, Takehiro Nagai, Kanae Miyazawa, Yusuke Tani, and Taeko Ishibashi for their professional and scientific advice and administrative support. I would like to also thank all of our lab and ex-lab members at Toyohashi and Fleming's lab members at Giessen for helping me out in my work and stay.

Finally, I am very grateful for my family members and girlfriend for their patience and support.

Contents

List of Figures

Chapter 1

Introduction

1.1 Material perception (“shitsukan”)

We encounter various objects made of many kinds of materials in our daily life. For example, we use a laptop made of steel, sit on a wooden chair, and eat an ice cream sundae from a glass cup. We effortlessly recognize them and infer their physical, functional, and multisensory properties at a glance. This ability is called material perception, that is, perception of objects’ features and physical properties (Anderson, 2011; Fleming, 2014, 2017). It is also known as “shitsukan” in Japanese, a word that broadly expresses both the physical and mental states of an object (Komatsu and Goda, 2018).

Material perception helps us make a decision about the most appropriate way to approach before touching an object. For example, when we look at a flask filled with liquid on an unstable surface, we may grasp it delicately (see Figure 1.1). We can use all five senses to perceive object properties in the external world. Here, this thesis discusses how we distinguish various materials using our vision.

1.2 Three physical factors for material perception

Our eyes obtain a pattern of light rays as visual information from the external world. Images are constructed with three physical factors: three-dimensional shape, illumination, and optical properties (Figure 1.2). This complex information is converted to a two-dimensional retinal image and our brain recognizes it as shitsukan information. When we change an optical property of a silver-like material to that of glass, the pattern of light rays actually changes dramatically. However, we simply perceive changes in the material. Moreover, when we change the illumination to another source, the pattern of light rays also changes dramatically, yet our perception of the target material remains stable. Thus, the visual system somehow makes material recognition easy using certain heuristics (Fleming, 2012) or efficient rules.

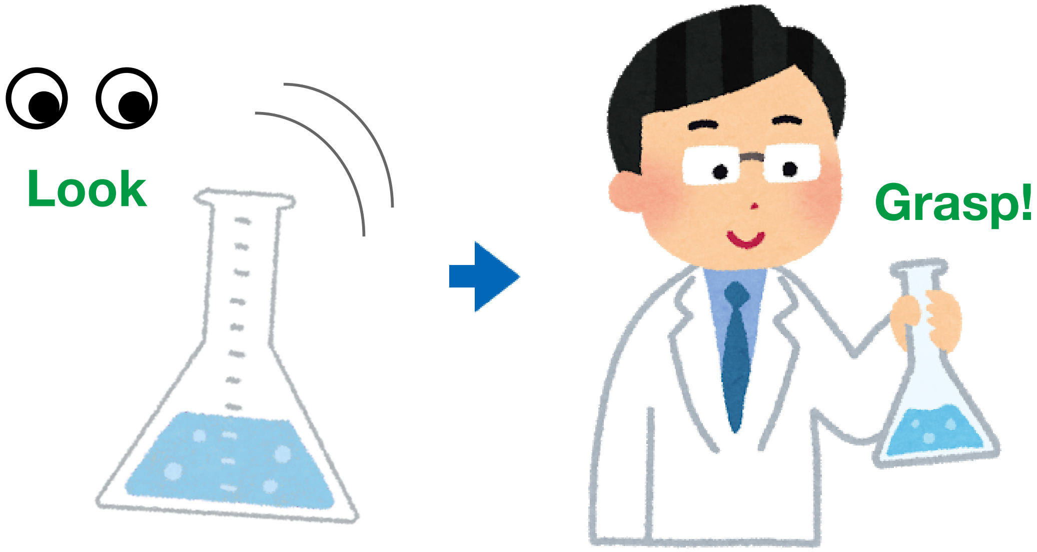


Figure 1.1: We approach appropriately before making contact with an object, using visual information

1.3 Why mirror and glass?

1.3.1 Mirror and glass

One of the main streams of material perception studies is to clarify how the visual system distinguishes various materials. Previous studies reported that the visual system comprehensively categorizes various materials, such as wood, metal, and glass, using static images (Hiramatsu et al., 2011; Fleming et al., 2013; Sharan et al., 2014; Nagai et al., 2015; Tanaka and Horiuchi, 2015). This is important to our survival and the function is known as "material recognition" or "material categorization."

Here, we especially focus on two specific materials "mirror" and "glass" (Kim and Marlow, 2016; Tamura et al., 2018). How does the visual system distinguish them? Mirror material is a specular reflected material, such as a well ground smooth surface metal (Fleming et al., 2004). In contrast, glass material is a thick transparent and refracted material, such as common glass or plastic (Fleming et al., 2011; Schlüter and Faul, 2014, 2016). In this thesis, these materials are physically defined as follows:

- An object has a smooth surface
- Mirror material a Bidirectional Reflectance Distribution Function model, which is perfectly specular reflection (R=1, T=0)

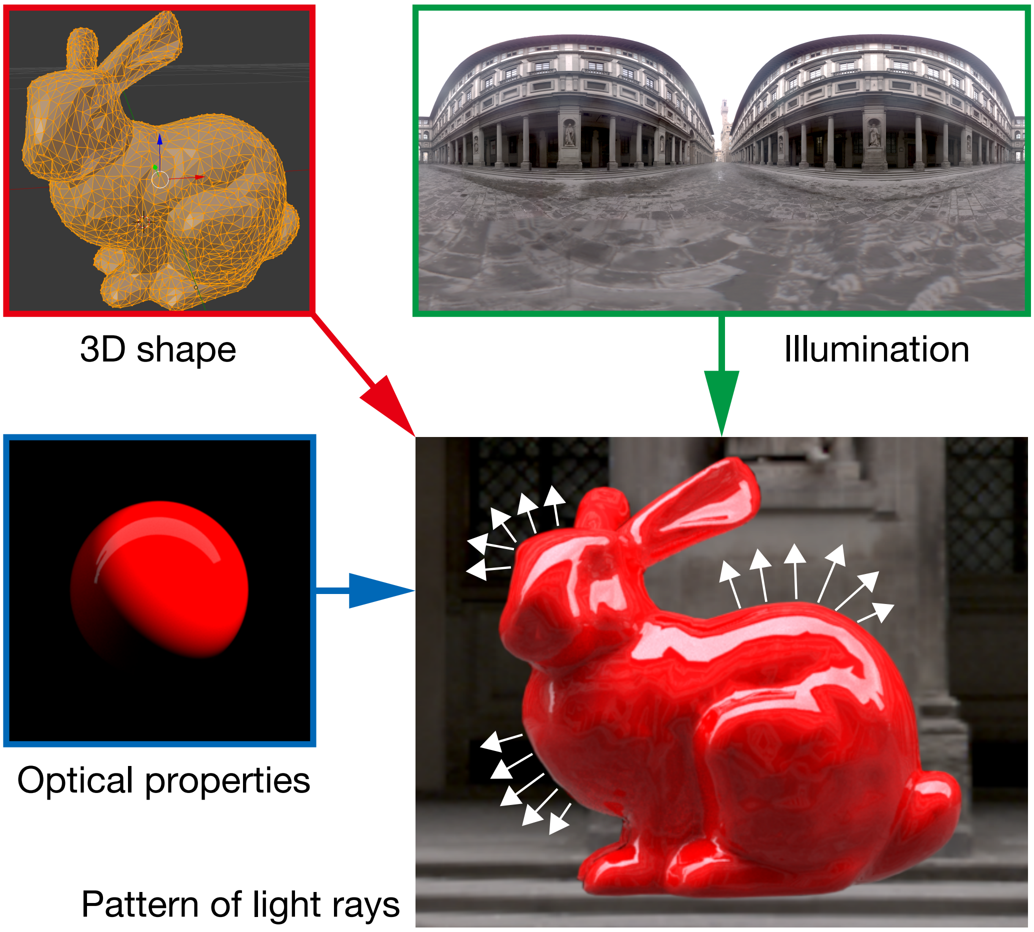


Figure 1.2: Three physical factors and an image with shitsukan

Three physical factors and an image with shitsukan. Three physical factors (3D shape, illumination, and optical properties) make a complex pattern of light rays. This figure is adapted from Fig. 2 in Komatsu and Goda (2018).

- Glass material has a physical property, in which refractive index is 1.5 (R=0.96, T=0.04)

As Figure 1.3, a relationship between an incoming light, reflectance, and transmittance are theoretically defined with the following formulas.

$$R = \left\{ \frac{n_t - n_i}{n_t + n_i} \right\}^2 \tag{1.1}$$

$$T = \frac{4 n_t n_i}{(n_t + n_i)^2} \tag{1.2}$$

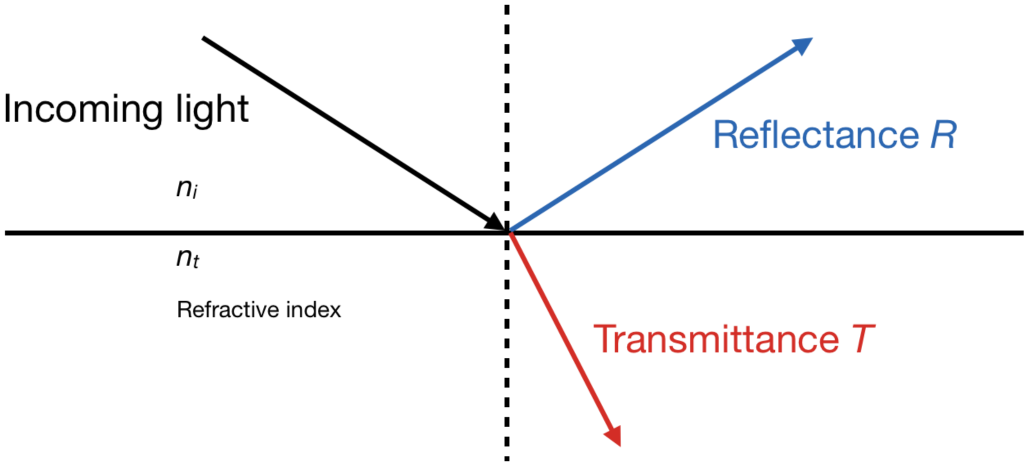


Figure 1.3: Relationship between incoming light, reflectance, and transmittance

When incoming light travels from a medium with refractive index n_i to another medium with refractive index n_t, the reflectance R and transmittance T of the output light are expressed by the relationship as mentioned above. For example, when incoming light travels from air ($n_i = 1$), to common glass($n_t = 1.5$), R and T are 0.04 and 0.96, respectively.

The appearance of both a mirror and glass derive strongly from their surroundings. Changing the surroundings directly changes their appearance. For example, mirrors create a distorted and reflected image of the surrounding world. In contrast, glass materials create an image, which is transmitted, refracted, and reflected internally multiple times before emerging as incident light again. The cues for distinguishing a mirror from glass are somehow generalized to deal with many varieties of images; in other words, all situation that we would meet. Therefore, a perceptual border between mirror and glass in our perception is sometimes ambiguous even when the physical properties of the materials are totally different.

1.3.2 An ambiguity between mirror and glass

We use an example to illustrate the ambiguity between mirrors and glass in Figure 1.4. When we are asked which materials comprise part of the door (within the green area) in Figure 1.4, it is difficult to answer correctly quickly. In particular, within the yellow area inside the green area, we do not rapidly judge whether the image on the surface is reflected from the bookshelf on the front side, or the image shows the bookshelf on the rear side through the window. Although this example is shown as a special difficult case from daily life, there is such an ambiguity when distinguishing a mirror from glass.

On the other hand, Figure 1.5 shows examples of actual target images (stimuli) in this study.

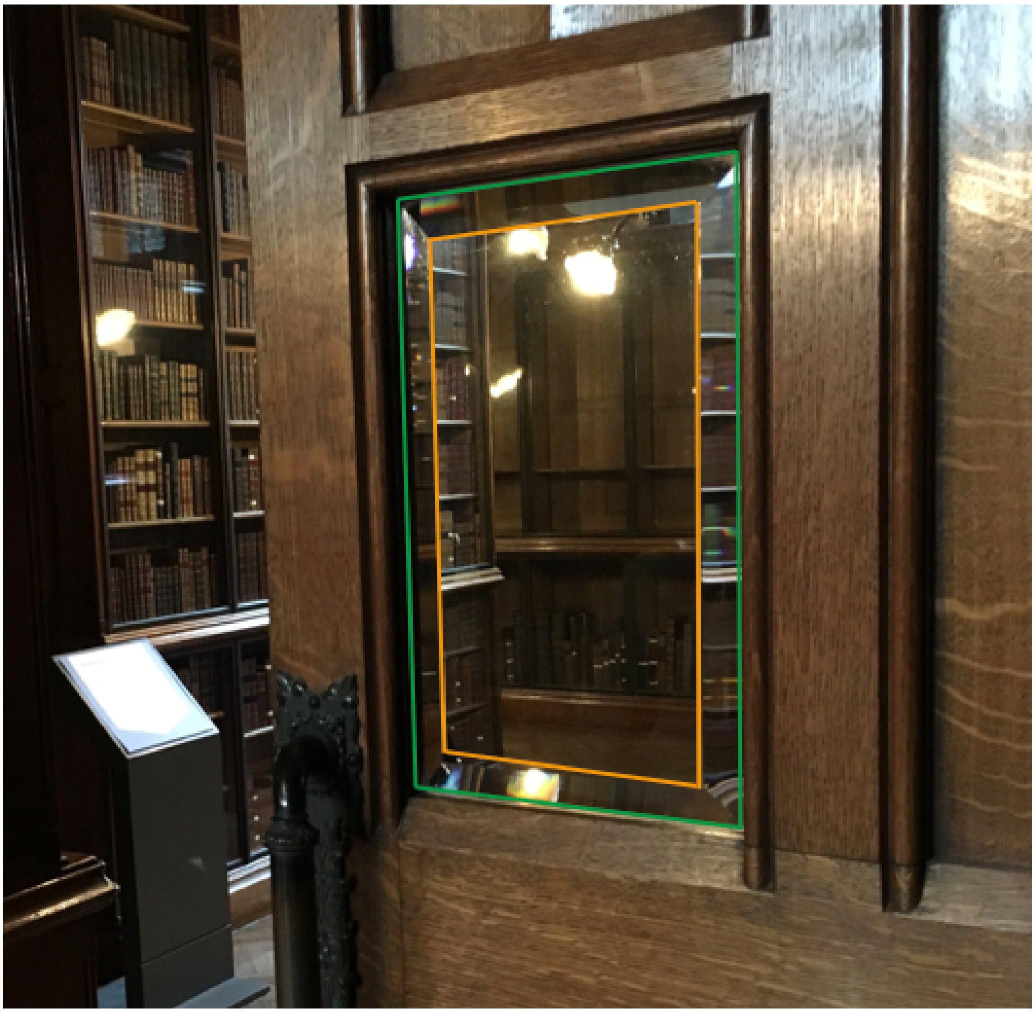


Figure 1.4: An example of ambiguity in distinguishing a mirror from glass in the material used in the door

It is difficult to distinguish whether the image is from the front or behind the door. This picture was taken by the author at the John Rylands Library (Manchester, UK) on August 9th, 2018.

The target shape is not so simple in this context, which is similar to problems occurring in real life. Reflectance, transmittance, and refraction of light from natural illumination have complex interrelations and are input to our retina as two-dimensional visual information (image). The visual system estimates real three-dimensional information from the image, and recognizes their surface properties.

It is quite unlikely that we process all image information because we recognize the surface property or material of the target object at a glance. We usually perceive mirror and glass effortlessly and the visual system may use certain efficient methods. Our aim is to identify the information, in other words,

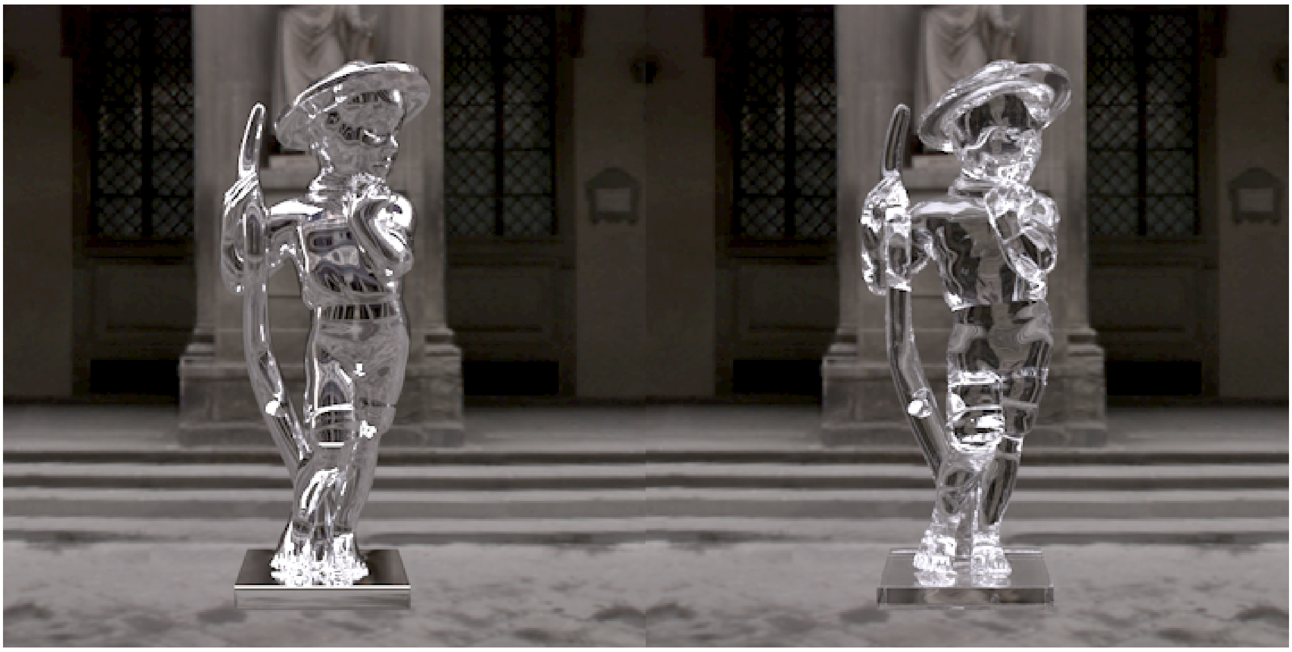


Figure 1.5: Mirror and glass objects

The left and right are made from mirror and glass materials, respectively.

what kinds of important cues are used for recognizing these materials. Thus, we hypothesized that the visual system properly uses certain visual cues to distinguish mirror from glass without complex information processing derived from reflection, transmittance, and refraction of the light.

1.4 Approach

An experiment to test the mirror-glass distinguishing problem using actual objects is not easy to perform because the mirror object completely reflects not only an image of illumination but also that of the camera or the observer on its surface. This strongly biases our perception and we cannot simply compare the two materials. To solve this problem, we used images rendered by computer graphics to accurately simulate the physical appearance of the target objects without the viewer. Use of rendering software has become common recently and is a standard way to create stimuli for material perception studies, such as with natural illumination (e.g., Fleming et al., 2003), a thick transparent object (Fleming et al., 2011), translucency (Fleming and Bülthoff, 2005; Motoyoshi, 2010), with motion (e.g., Doerschner et al., 2011), and liquids (e.g., Kawabe et al., 2015b; van Assen and Fleming, 2016). In this thesis, we rendered all simulated images using a Mitsuba renderer (Jakob, 2010).

We experimentally tested our hypothesis using a combination of human psychophysics (Chapters

2-6), modelling (Chapters 2-4), and machine learning (Chapter 2). We collected behavioral data in which human observers were asked to judge or rate stimuli, and regarded the data as target values. Then, an ideal model was constructed to minimize differences between the target values and those of the model, or to maximize the correlation between them. It is also useful to take a "big data" approach in terms of using a massive number of models designed as the visual system and images of mirror and glass. We performed data mining to obtain reliable cues using deep learning (e.g., Cichy et al., 2016; Guclu and van Gerven, 2015; Hong et al., 2016; Khaligh-Razavi and Kriegeskorte, 2014; Kietzmann et al., 2018; Kriegeskorte, 2015; Kriegeskorte and Douglas, 2018; LeCun et al., 2015; Yamins et al., 2014; Yamins and DiCarlo, 2016).

1.5 Overview

This thesis comprises five studies (Chapter 2-6). First, we present what material perception is, why we need to clarify perception of mirror and glass, and the goal and approaches of this study as an Introduction in this chapter. Next, in Chapter 2, we present how the visual system and models developed by machine learning distinguish mirrors from glass, and discuss the insights from the results. In that chapter, we also propose the use of diagnostic images that perfectly decorrelate image labels and human perception. Then, in Chapters 3 and 4, we present static and dynamic cues for distinguishing mirror from glass. Our main focus in these chapters is what visual cues contribute to distinguishing them. Moreover, in Chapters 5 and 6, we present studies of two optical illusions, which are related to the surface properties of mirror and glass derived from specular highlights. We examined the relationship between brightness enhancement and self-luminosity perception of the glare illusion. We also present the rotating glass illusion in which we discovered a new illusion involving material appearance, perceived shape, and motion. Finally, we summarize the outcomes of the five studies in the final chapter as the conclusion (see also Figure 1.6).

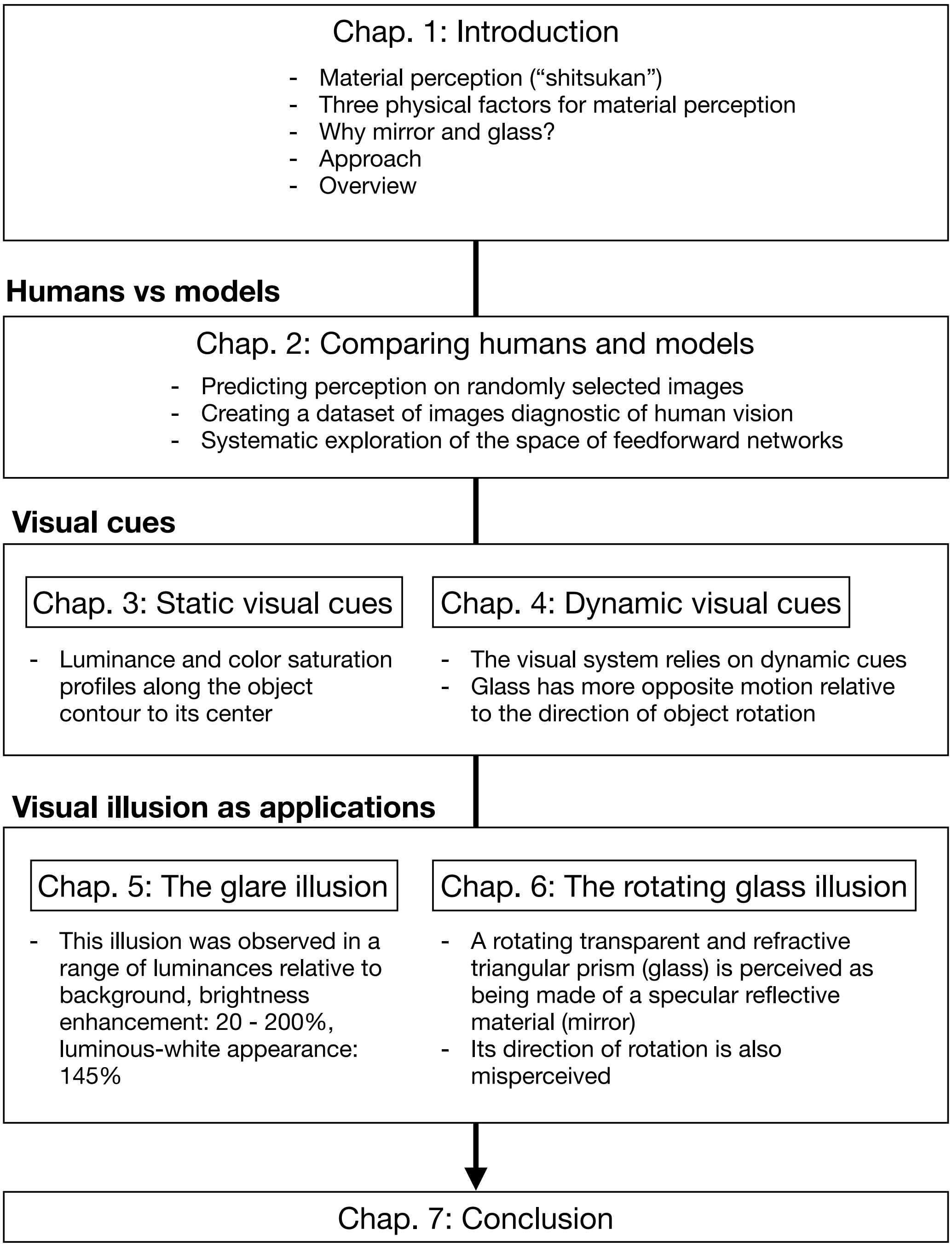

Chap. 1: Introduction
- Material perception (“shitsukan”)
- Three physical factors for material perception
- Why mirror and glass?
- Approach
- Overview
Humans vs models
Chap. 2: Comparing humans and models
- Predicting perception on randomly selected images
- Creating a dataset of images diagnostic of human vision
- Systematic exploration of the space of feedforward networks
Visual cues
Chap. 3: Static visual cues
- Luminance and color saturation profiles along the object contour to its center
Chap. 4: Dynamic visual cues
- The visual system relies on dynamic cues
- Glass has more opposite motion relative to the direction of object rotation
Visual illusion as applications
Chap. 5: The glare illusion
- This illusion was observed in a range of luminances relative to background, brightness enhancement: 20 - 200%, luminous-white appearance: 145%
Chap. 6: The rotating glass illusion
- A rotating transparent and refractive triangular prism (glass) is perceived as being made of a specular reflective material (mirror)
- Its direction of rotation is also misperceived
Chap. 7: Conclusion

Figure 1.6: Flowchart of this thesis

Chapter 2

Comparing humans and models

A similar version of this chapter will be submitted as:

Tamura, H., Prokott, K.E.,& Fleming, R.W. Distinguishing mirror from glass: A 'big data' approach to material perception.

Visually identifying materials is crucial for many tasks, yet material perception remains poorly understood. Distinguishing mirror from glass is particularly challenging as both materials derive their appearance from their surroundings, yet we rarely experience difficulties telling them apart. Here we took a 'big data' approach to uncovering the underlying visual cues and processes, leveraging recent advances in neural network models of vision. We trained thousands of convolutional neural networks on >750,000 simulated mirror and glass objects, and compared their performance with human judgments, as well as alternative classifiers based on 'hand-engineered' image features. For randomly chosen images, all classifiers and humans performed with high accuracy, and therefore correlated highly with one another. To tease the models apart, we created a diagnostic image set for which humans make highly systematic errors, allowing us to decouple accuracy from human-like performance. A large-scale, systematic search through feedforward neural architectures revealed that relatively shallow networks predicted human judgments best, although, surprisingly, no network correlated better than 0.6 with humans. These results cast doubt on recent claims that such architectures are good models of human vision.

2.1 Introduction

Different materials, such as steel, silk, meat or glass, have distinctive visual appearances, and our ability to recognize such materials by sight is crucial for many tasks, from selecting food to effective tool use. Yet, material perception is challenging. The retinal image of a given object is the result of complex interactions between the object's optical properties, 3D shape and the incoming light (Adelson, 2001; Fleming, 2014, 2017; Komatsu and Goda, 2018). Thus, a given material can take on an enormous variety of different appearances depending on the lighting, object shape and viewpoint. At the same time, similar objects with different material properties can create quite similar images in terms of the raw spatial patterns of colour and intensity (Fleming et al., 2003). To succeed at material perception, the visual system must somehow tease apart similar images belonging to different materials, while at the same time grouping together very diverse images belonging to the same material class (Rajalingham et al., 2018; van Assen et al., 2018b). This is a fundamental aspect of biological visual processing, which remains poorly understood.

A particularly challenging case is to distinguish polished mirror-like specular materials ('mirror') from colourless transparent materials ('glass') (Fleming et al., 2011; Kim and Marlow, 2016; Schlüter and Faul, 2014, 2016; Tamura et al., 2018; Tamura and Nakauchi, 2018). Both kinds of material derive their appearance entirely from their surroundings, but through different light transport processes. Mirrors create a distorted reflection of the surrounding world, whereas for glass materials, incident light also enters the material, refracts and may reflect internally multiple times before re-emerging. Yet, in both cases, changing the object shape or surrounding world radically alters the image. As a result, the visual cues we use to distinguish between mirror and glass must generalize well across an enormous variety of images. At the same time, to distinguish the two kinds of material, the visual system must presumably use quite sophisticated image measurements that latch onto often quite subtle differences in the image resulting from the way light interacts with them (Figure A.1A, A.1B).

We reasoned that to work out how the visual system distinguishes mirror from glass, it is useful to take a 'big data' approach in which we embrace the enormous diversity of images of mirror and glass materials that confront the visual system. In particular, using computer graphics we sought to create a dataset of hundreds of thousands of images, which could then be data mined for reliable visual cues. To do the data mining, we turned to deep learning methods.

Over the last five or so years, artificial neural networks (LeCun Jackel, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard et al., 1990; Lecun et al., 1998) have demonstrated significant potential as models of biological vision (e.g., Cichy et al., 2016; Guclu and van Gerven, 2015; Jozwik et al., 2017; Khaligh-Razavi and Kriegeskorte, 2014; Kheradpisheh et al., 2016a; Yamins et al., 2014). See also reviews (Kietzmann et al., 2018; Kriegeskorte, 2015; Kriegeskorte and Douglas, 2018; LeCun et al., 2015; Majaj and Pelli, 2018; Yamins and DiCarlo, 2016). We set out to leverage these advances to gain insights into the visual processes underlying the challenging material perception task of distinguish mirror from glass. Comparisons between human vision and computational models typically use randomly selected images (Ghodrati et al., 2014; Kheradpisheh et al., 2016a,b; Yamins et al., 2014), for which both humans and models achieve high performance. In contrast, our goal was to develop a model that would behave like humans according to more strictly defined criteria. Specifically, we sought a model that could not only predict the successes of human judgments, but also systematic errors, which are presumably the hallmarks of the processes unique to human visual computations. To do this, we created a 'diagnostic' image set that yielded systematic and consistent visual errors (as well as correctly perceived images). However, to our surprise, we found that despite an extensive and guided search, none of the neural networks we investigated correlated better than 0.6 with human judgments.

2.2 Results and Discussion

2.2.1 Predicting perception on randomly selected images

Using computer graphics, we rendered over 750,000 images of mirror and glass objects (Figure 2.1A and 2.4.2) of a wide variety of shapes, naturalistic illuminations and viewpoints. Half were mirror objects (pure specular reflection), the other half were glass (refractive index 1.5). Of these, we randomly selected 500 images of each material class and asked 10 volunteers to rate each object on a five-point scale where 1 indicated that it looked like compellingly like glass, 5 indicated that it looked compellingly like a polished metal, and intervening values indicated different degrees of ambiguous appearance. Figure 2.1B shows a clear bimodal distribution of ratings, with mean ratings of 3.64 for mirror images and 1.75 for glass, with an accuracy of 77.9% correct responses. This suggests that observers are generally quite good at distinguishing mirror from glass.

Before investigating deep learning models in depth, we tested the extent to which relatively simple image measurements could predict perceptual mirror/glass judgments (see Figure A.2A). Specifically, we compared human performance with two 'hand-engineered' image-computable classifiers based on pixel intensity and color histograms ('Color-Hist'), 'mid-level' texture statistics (Portilla and Simoncelli, 2000) ('Port-Sim'). All classifier types were trained on half the dataset (chosen at random) and tested on the other half (all images that were shown to humans were excluded from training and test sets). The Color-Hist classifier used eight pixel-histogram statistics (mean, variance, skewness, and kurtosis of luminance and saturation distributions). 'Port-Sim' used 1,052 features derived—via PCA—from both color and higher-order wavelet coefficient statistics (Portilla and Simoncelli, 2000); see 2.4). These two classifiers were trained to classify mirror and glass with a logistic regression, with the output ranging from zero (glass) to one (mirror). Surprisingly, although based on quite simple image measurements, both of these classifiers achieved accuracies that almost rivaled human performance on the 1,000 images rated by our observers (Figure 2.1C). This suggests that despite the complex optics of reflection and refraction, there are many potential cues that would suffice to perform significantly above chance at distinguishing the two kinds of materials. However, we sought the specific cues that the human visual system relies on. A better test of this is the correlation between the Color-Hist and Port-Sim models and humans on an image-by-image basis. Although the models did correlate significantly with human performance, they did so significantly less well than individual humans do (Color-Hist vs humans: $t(18) = 6.056$, $p < 0.001$; Port-Sim vs humans: $t(18) = 2.356$, $p < 0.05$, t-test), suggesting that humans do not rely on the same cues as these simple classifiers (Figure 2.1D).

As an initial attempt to investigate the potential of CNNs to predict human performance, we trained 10 instantiations of a feedforward network with three convolution layers (see Figure A.2B and). The 10 instances had identical architecture and training, but different initial random weights. On the same random images as before, the networks achieved and accuracy that superseded humans, and correlated with mean human responses within the same range as individual humans did, thus outperforming the two 'hand engineered' classifiers. This suggests that CNNs learn features that are inherently superior to the simple color and texture statistics. This in itself is unsurprising as the CNNs learn many more features (99,410), and thus perform the classification in a higher dimensional space. For the purposes of understanding biological vision, the key question is whether the features learnt by the CNNs resemble those used by the human visual system.

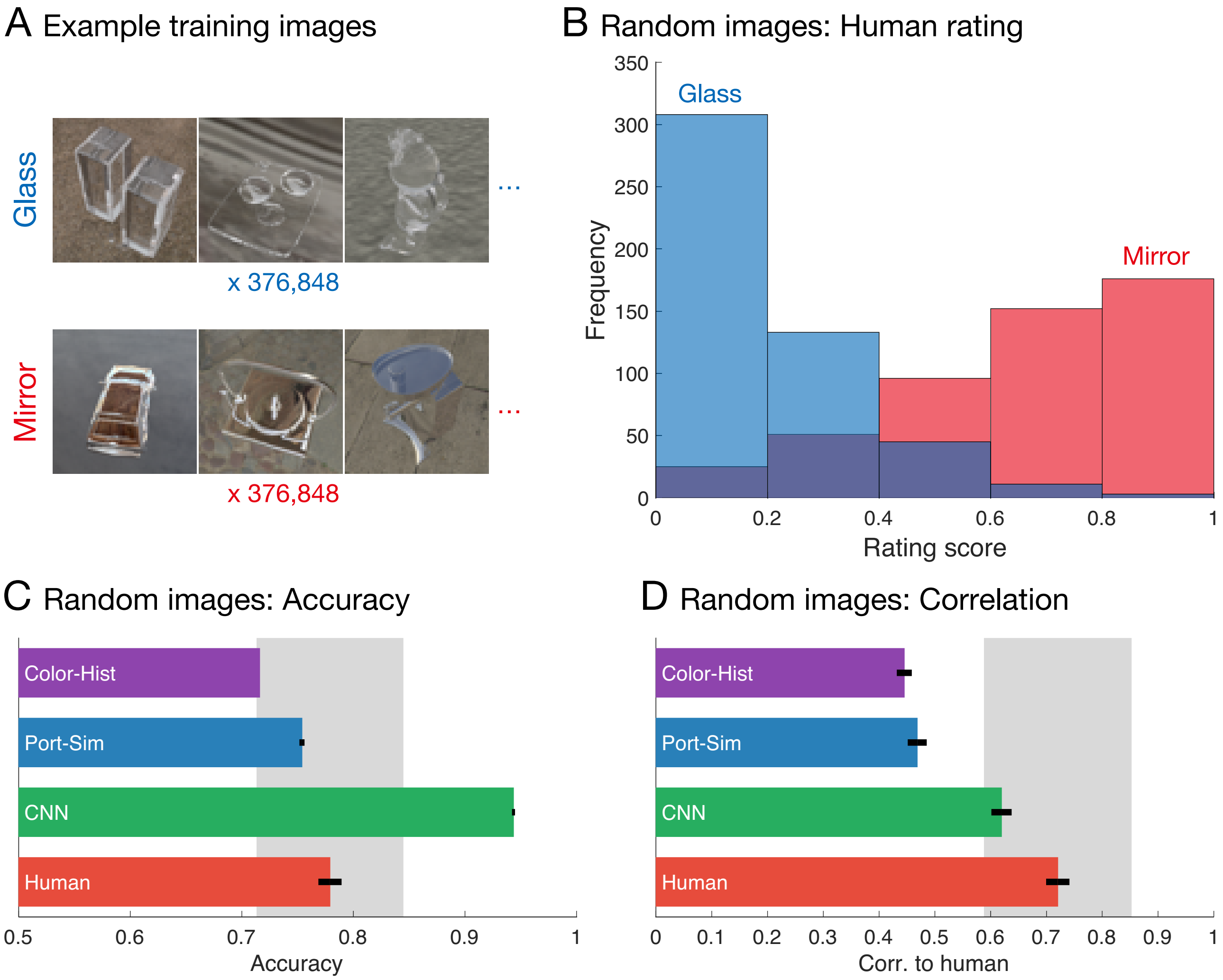


Figure 2.1: Results of randomly selected renderings

(A) Examples of images from the test stimuli (randomly selected renderings) from the data set. Each row shows mirror and glass renderings. (B) Human rating result. The horizontal axis indicates human rating score of 10 observers from glass to mirror (0-1). The vertical axis indicates a frequency of each bin. (C) Accuracy of human and the classifiers' response for the test stimuli. These are average accuracy of 10 repetitions in each classifier and 10 observers. Error bars represent the standard error of the mean across all 10 classifiers or observers. Note that the error bar of Color-Hist is not visible but just tiny in the scale of the horizontal axis. The gray area indicates mean ± 2SD of all human observers. (D) The correlation coefficient between human to the classifiers for the test stimuli. The human-to-human correlation was defined as the average of 10 correlations that between each observer and the left observers' average. The symbols are the same as C.

To gain further insights into the nature of the internal representations of the classifiers, we then performed representational similarity analysis (RSA) (Kriegeskorte et al., 2008) using the images that had been rated by humans. Figure 2.2A shows the Representational Dissimilarity Matrices (RDMs; (Kriegeskorte et al., 2008)) for each of the classifiers, as well as ground truth. The rows/columns of the matrix represent the different images, ordered into two blocks by their true class (mirror vs. glass) and within a block by their mean human ratings (from most mirror-like to most glass-like). Individual entries represent the dissimilarity between the corresponding pair of images in terms of the perceived or predicted mirror vs. glass ratings. Thus, low values indicate the corresponding pair of images are represented as highly similar, while higher values indicate they are more dissimilar.

The patterns in the matrices suggest that for these randomly selected images, humans and classifiers broadly agree. We can summarize the relationships between the RDMs in a Classifier Correlation Matrix (CCM; also known as a 'second-order RDM' (Kriegeskorte et al., 2008)), in which each row/column indicates a different observer or computational classifier, and each entry contains the mean correlation between the RDMs for the corresponding pair of observers/models (Figure 2.2B). For comparison, we also included 10 random RDMs, to characterize how much more similar the classifiers are to humans than would occur by chance. Applying multidimensional scaling (MDS) to the CCM allows us to visualize the relationships in 3D (Figure 2.2C). The correlation between each classifier and humans were 0.30 (Color-Hist), 0.31 (Port-Sim), and 0.58 (CNN), respectively. This reveals that all three classifier types learn inter-image relations that are significantly and substantially closer to humans than occurs by chance, and of all classifier types, the CNNs appear to acquire the most similar representation to humans. These results tend to suggest that feedforward convolution neural networks have significant potential as models of human visual judgments of mirror and glass materials.

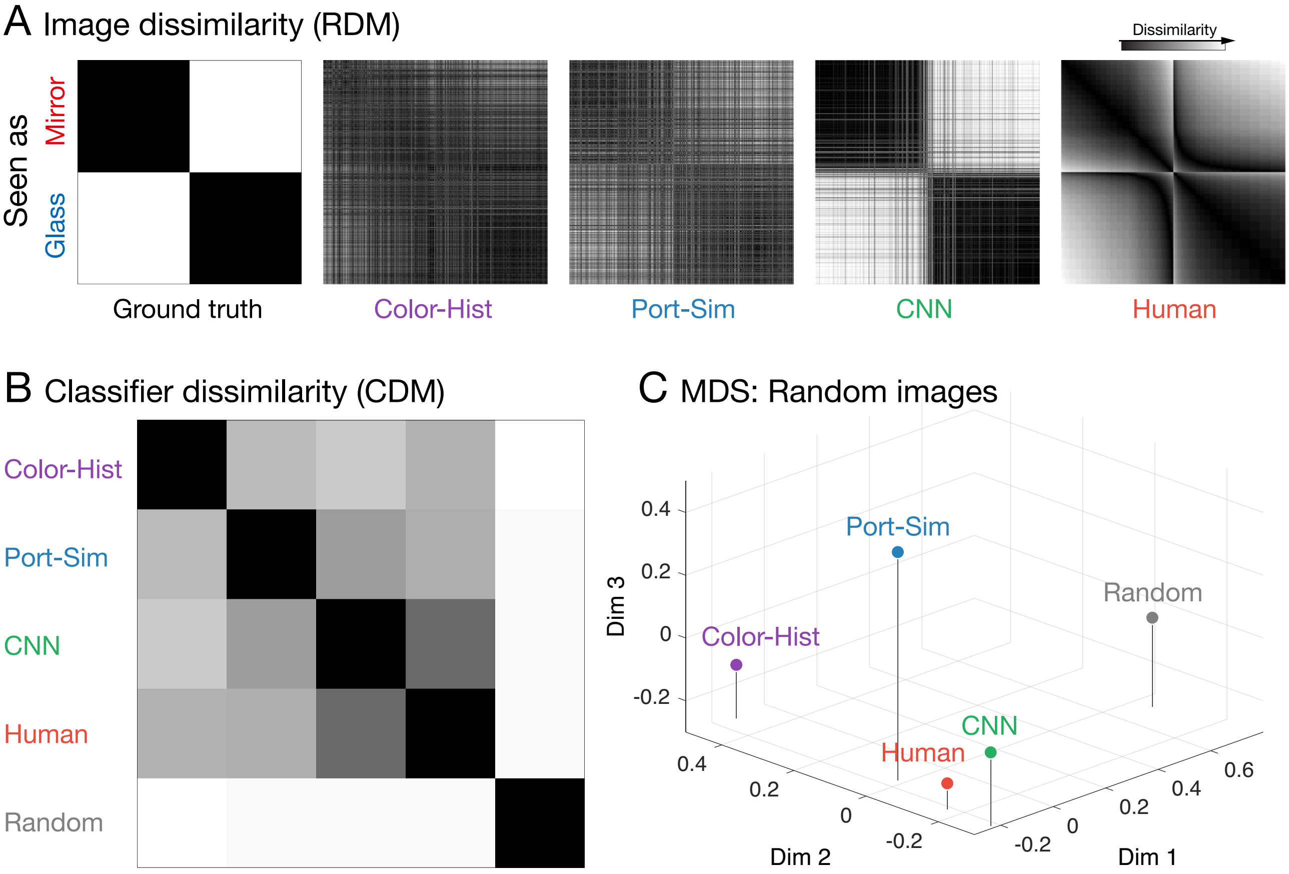


Figure 2.2: RSA of randomly selected renderings

(A) RDMs for each of the classifiers and the ground truth. The rows/columns of the matrix represent 1,000 images, ordered into two blocks by their true class (mirror vs. glass) and within a block by their mean human ratings from most mirror-like to most glass-like. Individual entries represent the dissimilarity between the corresponding pair of images in terms of the perceived or predicted mirror vs. glass ratings. The darker entries indicate the corresponding pair of images are represented as highly similar, whereas the brighter entries indicate they are more dissimilar. (B) CCM among the models. Each row/column indicates a different classifier, human, and random RDM as a control. Each entry contains the mean correlation between the RDMs for the corresponding pair of observers/models. The color code of entries is the same as A. (C) 3D visualization of the relationship between the models by MDS using the CCM. The three axes indicate the first three dimensions obtained by MDS.

2.2.2 Creating a dataset of images diagnostic of human vision

Based on the high correlations between observers and the computational models, it could be tempting to conclude that the models accurately simulate human visual processes. However, there are several reasons for caution. First, the main purpose of comparing models based on different features is to identify which features best predict human material perception. Yet, for randomly selected images, even the most primitive models appear to match human perception quite well. Given what we know about early vision and material perception (Anderson and Kim, 2009; Kim and Anderson, 2010; Marlow et al., 2012), it seems highly unlikely that visual perception of mirror vs glass is based on raw luminance and color distributions, which are entirely insensitive to the spatial structure of the image. Second, and more importantly, it is possible that the high correlations simply result from the fact that both humans and classifiers achieve quite high accuracies. If all models correctly assign most images to one of the two distinct modes ('mirror' or 'glass'), then it follows that they will tend to correlate with one another. Indeed, in Figure 2.2A, most (58%) of the variance in the human judgments is accounted for by the ground truth. A good model must be able to predict not only the successes of human vision, but also the specific pattern of errors, on an image-by-image basis. To test this, we need a set of diagnostic images that decouples accuracy from human judgments.

Creating such a dataset is nontrivial as most images are perceived correctly. It is not sufficient to identify images for which participants are inconsistent in their interpretation, as a deterministic, image-computable model cannot even in principle account for variations between observers when presented with the same image. Our goal is to predict that proportion of the variance in judgments which is *consistent* across observers, and therefore we need an image set that includes images that are consistently misperceived. Specifically, the goal was to identify a 'benchmark' set of images with a flat distribution across the five bins ranging from 'mirror' to 'glass' ratings, for both mirror and glass images (in contrast to the skewed distributions for random images in Figure 2.1B), thereby perfectly decorrelating the true material class from the perceived class. We set as our criterion of consistency that ten out of ten naïve observers should rate each image in the same bin of the 5-point rating scale.

To obtain this diagnostic image set, we performed a sequence of experiments using both crowd sourcing and laboratory judgments to progressively funnel images down to a set that is highly diagnostic of human performance (Figure 2.3A; see Figure A.3 and 2.4 for details). Identifying images that are perceived veridically is relatively straightforward, so we prioritized identifying 'illusory' images that are misperceived, reasoning that the bins corresponding to veridical percepts could be filled in afterwards. Specifically, from the full set of renderings, we selected 30,000 images at random, 1,500 of which were presented to each of 20 observers. Based on the responses, we then selected about 11,000 images—sampling uniformly from the judgments—which proceeded to the next round, in which images were rated by more observers via crowdsourcing. From their responses, we selected the 522 images which had been consistently rated as belonging to the wrong class by (at least) three observers. In the final round, ten observers rated each of these images, resulting in a total of 102 images which had been consistently rated as belonging to non-veridical bins. To fill in the veridical bins of the distribution we then had another ten observers rate 1,000 randomly chosen images from the renderings. From the responses, we selected 68 images that were consistently perceived correctly by all ten observers, yielding 170 images, i.e., 34 images in each of the five bins from perceived mirror to perceived glass, half of which were actually mirror and half glass. Some examples are shown in Figure 2.3B (see Figure

A.4 for details).

To increase the number of diagnostic images, we also trained a generative adversarial network (GAN; (Goodfellow et al., 2014; Radford et al., 2015)) on our renderings. GANs consist of a generator network, G, that is trained to produce images, which a discriminator network, D, has to distinguish from a given dataset. During training D improves at distinguishing the synthesized images from the training data, while G learns to create images that are hard to discriminate from the training data. The result is a model that can synthesize images with many of the visual characteristics in common with the renderings (see Figure 2.3B, also Figure A.4 for details). We find that such images include many cases that are more ambiguous than renderings, appearing somewhere between mirror and glass. In another 'funnel' sequence of experiments, we identified 95 images (out a set of 1,400), which were consistently rated by ten observers as belonging to specific bins. Combining the selected GAN images with the renderings yielded a total 265 diagnostic images, in which true material class was perfectly decorrelated from perception (Figure 2.3B).

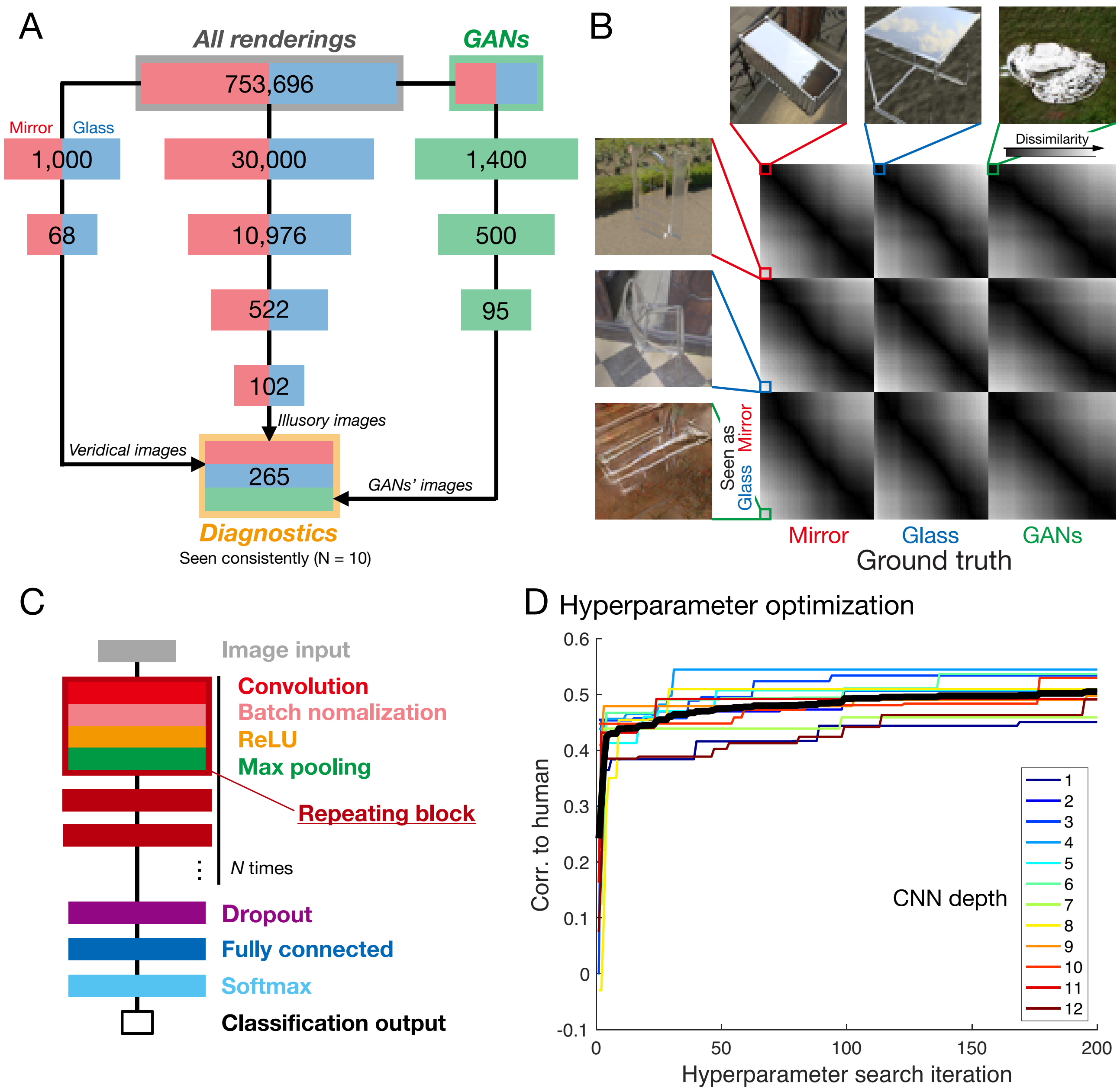


Figure 2.3: Diagnostic image set and systematic exploration of the space of the feedforward networks

(A) A flowchart of creating a diagnostic image set. The diagnostic images are composed of three different image types, such as 'veridical', 'illusory', and 'GANs'. Veridical images that their true and perceived classes are identical were selected from the result of the human rating experiment in Figure 2.1B. In contrast, those of illusory images are opposite. They and the GANs' images were funneled through each sequence of experiments (see 2.4 and Figure A.5). (B) RDM of diagnostic image set. The format is the same as Figure 2.2A except adding GANs as the third true class. The panel shows six example images that are extremely high or low rating score in each class. (C) A general form of the feedforward network architecture in this study (see also Figure A.2). (D) Results of BHS. The horizontal axis indicates the number of iterations of BHS. The vertical axis indicates correlation between human to each model with different network depth (from 1 to 12). A black thicker line represents the average of all 12 depths.

2.2.3 Systematic exploration of the space of feedforward networks

With this diagnostic image set in hand, we then sought neural network models that would correlate strongly with humans for these diagnostic images. It is important to note that the space of potential convolution network models is very large: they can vary widely in terms of their architectures, hyper-parameters and training schedules. We reasoned that within the space of feedforward neural networks, some networks are likely to approximate human visual processing better than others. We therefore ran a large-scale search through the space of feedforward networks with the general form depicted in Figure 2.3C, varying the network depth systematically (see also Figure A.5A). All networks consisted of an input layer followed by a basic 'block' of layers comprised of convolution, batch normalization (Ioffe and Szegedy, 2015), ReLU (Glorot et al., 2011), and max pooling layers, which were repeated several times, followed by dropout (Srivastava et al., 2014), fully connected and softmax layers, and ending with a two-unit classification output ('mirror' vs 'glass'). To investigate the effect of architecture depth, we systematically varied the number of basic blocks in the sequence from one to twelve. (See Methods). Then, for each network depth, we ran 200 iterations of Bayesian hyperparameter search (BHS) to identify the values of 11 hyper-parameters controlling the network architecture and training (e.g., number of filters per layer, initial learning rate, momentum; see Figure A.5B) in 'the optimization-stage'. The objective of the BHS was to optimize correlation to human judgments on the diagnostic image set, which progressively improved across iterations (Figure 2.3D). All networks were trained on a randomly chosen 400,000 renderings (200,000 images in each class) with 2-fold cross validation in order to converge the network quickly.

Having identified promising hyperparameters for each architecture depth, in 'the validation-stage', we then trained 30 instances of each of the resulting neural networks (differing only in the initial random state), again using the same number of renderings with half for training and the other half for testing. Importantly, these networks were never trained on the diagnostic images, and training proceeded until the validation accuracy had not improved for at least three validations, independently for each architecture depth. The mean correlations between the networks and humans on the diagnostic image set is shown in Figure 2.4. Of all depths, the 3-layer network architecture ('OptCNN') was the one that correlated best with human judgments, and was the network class we considered for further analysis. Importantly, however, none of the thousands of networks trained throughout the BHS or the final validation exceeded a correlation with humans of 0.6; whereas human-to-human correlations was between 0.61 and 0.81. In other words, although OptCNN was the closest of all models we considered, it still failed to capture average human performance as well as even the most unrepresentative of the individual humans did.

Figures 2.4A and 2.4B compares the highest of OptCNN to humans and the other classifiers on the diagnostic image set. By design, humans perform at chance on these images (Figure 2.4A). All classifiers outperform humans in terms of accuracy, yet OptCNN is the closest, making the most errors on these images, even though it was trained with the same objective function and a very similar training set to the original CNN, which performs too well to resemble humans. In terms of image-by-image correlations to human judgments (Figure 2.4B), none of the classifiers reaches the noise ceiling, but OptCNN significantly outperformed all other classifiers (Color-Hist vs OptCNN: $t(9) = 4.79 \times 10^{13}$, $p < 0.001$; Port-Sim vs OptCNN: $t(9) = 5.19$, $p < 0.001$; CNN vs OptCNN: $t(9) = 20.23$, $p < 0.001$, t-test).

To compare the nature of the representations in the different classifiers and humans, we then performed RSA. To do this, we measured the similarity between all images in the diagnostic image set according to the final classification output of each classifier. To visualize the relationships between the different classifiers, we computed the mean correlation between the different classifier types, and then performed MDS as same as Figure 2.2D. Figure 2.4D shows the different classifier types arranged in the first three resulting dimensions. This analysis reveals that OptCNN was the most similar to humans, although, there is still a substantial residual difference.

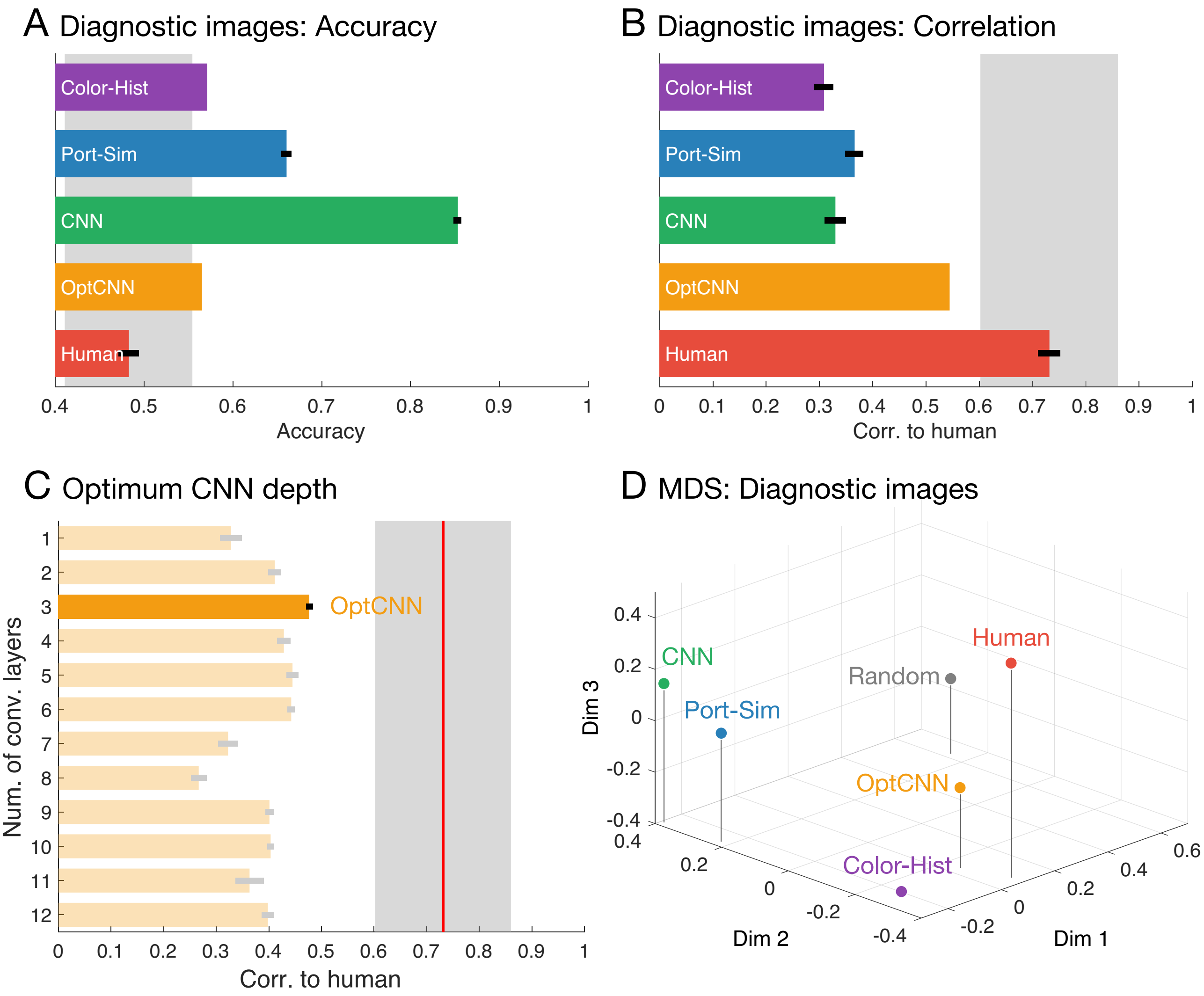


Figure 2.4: Results of the diagnostic image set

(A) Accuracy of human and the classifiers' response for the test stimuli. (B) The correlation coefficient between human to the classifiers for the test stimuli. The symbols of A and B are the same as Figure 2.1C and 2.1D, respectively. Note that OptCNN represents the highest correlation of 30 instances of 3-layer CNN in the exploration. (C) Relationship between the number of depths of the feedforward networks and the correlation to humans. The horizontal axis indicates the correlation to humans. The vertical axis indicates the number of convolution layers (i.e. the number of repeating blocks) of the feedforward neural networks in our systematic exploration. The red vertical line and gray area indicate mean ± 2SD of all observers. (D) 3D visualization of the relationship between the models by MDS. A sequence of analyses using RSA is the same as Figure 2.2 and this panel shows the final stage only. The format is the same as Figure 2.2C.

The greater similarity between OptCNN and the default CNN is also revealed by a more detailed view into the representations at different processing stages of the networks. We applied RSA using the diagnostic image set, to the input stage, the three ReLU stages after each convolution layer, the fully connected layer (FC) and the final output in both the original CNN and OptCNN. We then computed correlations between each of the resulting RDMs ('Layer Correlation Matrix', LCM), and performed MDS to visualize the relationships between the different representations, along with the true labels and human judgments (Figure 2.5A). At the input stages of the network, images that are perceived by humans as mirror and glass materials are thoroughly entangled, such the further processing is required to separate them. As we proceed through network layers, we see that the representations in CNN and OptCNN diverge. While CNN's representations increasingly approach the ground truth, OptCNN's representation increasingly organizes images in the way that humans do, such that in the output stages, images that subjectively appear like mirror or glass are teased apart.

OptCNN also outperforms the original CNN in terms of its robustness to noise, another key characteristic of human vision (Geirhos et al., 2017). If we perturb the input images with noise, the networks' predictions about the material tend to change (Figure 2.5B). Importantly, we find that while the correlation between CNN's predictions and the human judgments of the unperturbed images falls precipitously as noise is added, for OptCNN, not only is the correlation higher across all noise levels, the decline is also gentler. This suggests that by identifying networks that more closely resemble humans in terms of their solution to the objective function, we also identify representations that capture other aspects of human perception, such as robustness to noise.

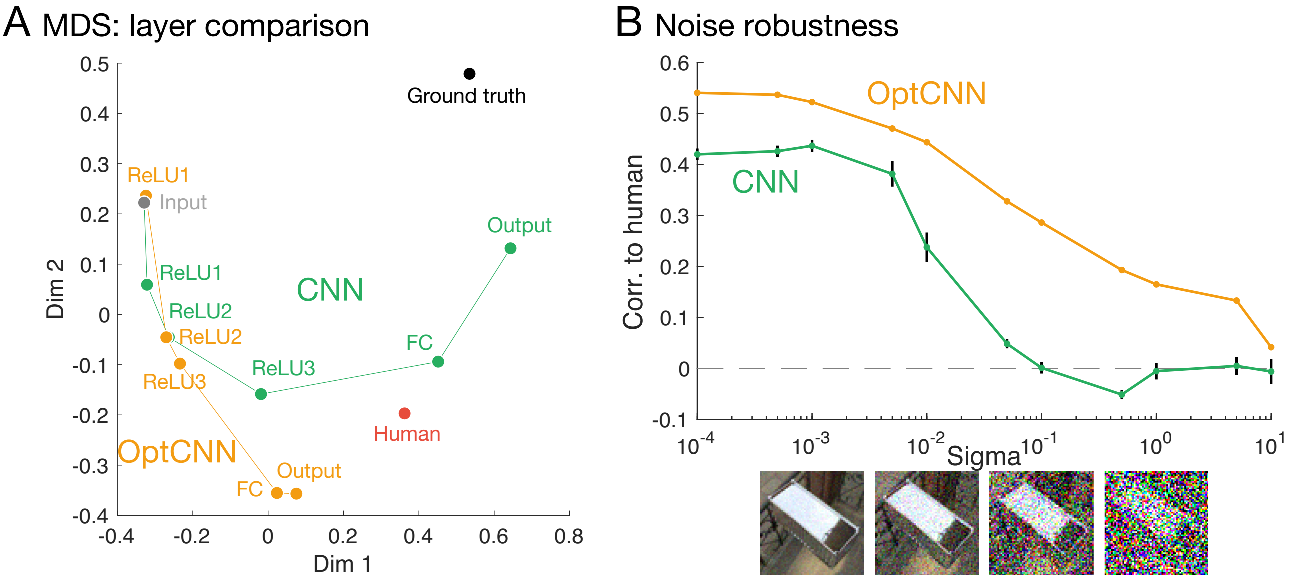


Figure 2.5: Comparing CNN and OptCNN in terms of the internal representation and robustness to noise

(A) 2D visualization of network internal representations between the layers by MDS. The horizontal and vertical axes indicate the first two dimensions obtained by MDS using LCM (the 2nd-stage RDMs between the layers). Two main streams show the relationship between layers in CNN (one of 10 instances) and OptCNN (the highest of 30 instances) with human judgements and the ground truth (see also text). (B) Robustness to noise. The horizontal axis indicates sigma of Gaussian noise (i.e., the amount of image perturbations). Four images show examples with different sigma (10^{-3}, 10^{-2}, 10^{-1}, and 10^{0}). The vertical axis indicates the correlation to humans. Error bars represent the standard error of the mean across all 10 classifiers (CNN). Note that OptCNN represents the highest correlation of 30 instances (the same as OptCNN in A).

2.3 General Discussion

As many materials that we can easily recognize did not exist until the last few centuries—or even decades—our ability to recognize them must be learned rather than evolved. How the visual system acquires the visual computations and internal representations that allow us to succeed at material perception remains poorly understood. Here we investigated the extent to which deep learning can reveal representations that resemble human judgments.

Studies comparing humans to machine learning models often focus on overall performance at a task (Ghodrati et al., 2014; Hong et al., 2016; Kheradpisheh et al., 2016a,b; Kubilius et al., 2016; Majaj et al., 2015), or correlation on arbitrary images (Hong et al., 2016; Kheradpisheh et al., 2016a,b; Kubilius et al., 2016; Majaj et al., 2015) for which both humans and successful machine vision system tend to perform well. Here, by contrast, we defined a new criterion for comparing neural networks with human judgments, by creating a *diagnostic* set of images in which human performance is decorrelated from ground truth. This allows us search for networks that capture the characteristics eccentricities of human vision, reproducing the tell-tale errors that humans tend to make. Although identifying such an image set is time consuming and effortful, it provides a benchmark against which all future models of human vision can be tested. Here, we created one such set for a challenging material perception task: the discrimination of mirror and glass materials.

By comparing human performance on the diagnostic image set against an array of models, we were able to show that neither simple color statistics, nor more sophisticated texture statistics can predict human judgments of mirror vs glass. More importantly, we also showed that an arbitrary CNN, which appeared to resemble humans when tested on arbitrary images, did not resemble humans very closely at all when evaluated on the diagnostic image set. We then performed a large-scale, systematic search through the space of feedforward networks trained to distinguish mirror from glass objects in search of a neural network that more closely resembled human performance. The network architecture that performed most similarly to humans (OptCNN) was a three-layer network, which was not especially good at distinguishing mirror from glass in terms of the true physical labels (at least compared to rival models). This suggests that humans are far from optimal at the task; indeed, many artificial neural networks can outperform them. However, importantly, to our surprise, even the optimised model did not reproduce average human behaviour on the diagnostic image set as well individual humans do. While previous studies have noted that neural networks have reached the 'noise ceiling' of human-to-human performance (e.g., Kheradpisheh et al., 2016a; Kubilius et al., 2016), we find that when tested on a test set that is truly diagnostic of human vision, such conclusions might not in fact be warranted, and that further research is required to find good models of human vision.

Why did OptCNN fail to match human performance? There are at least three important respects in which the models differ from humans and the human brain. First, one of the most striking aspects of human visual cortical processing is the massive amount of feedback (Budd, 1998; Felleman and Van Essen, 1991; Muckli and Petro, 2013). It is widely believed that feedforward processing is responsible for many of our visual abilities. For example, high-level visual tasks, such as animal detection (Thorpe et al., 1996) can be completed successfully too rapidly for feedback to contribute. Nevertheless, the feedback presumably plays an important role, potentially in selective visual attention, visual imagery and the learning process that establishes the representations in the first place. Here, we considered only feedforward architectures. It could be that a key missing ingredient in OptCNN is recurrent processing, and that adding feedback signal flow could make up some of the shortfall in correlation with humans.

A second important difference is the nature of the training objective. Here—as in almost all neural network-based putative models of human vision (Kriegeskorte, 2015; Majaj and Pelli, 2018; Yamins and DiCarlo, 2016)—we used *supervised learning* in which the network is trained on hundreds of thousands of accurately labelled images. Human vision cannot, even in principle, be trained this way, as there is no objective way to establish the ground truth, and the scale of the training set almost certainly exceeds human visual experience with mirror and glass objects. In particular, we very rarely get to see mirror and glass versions of the same objects under the same lighting and viewpoint, and we presumably also exploit that vision unfolds continuously over time, rather than in independent static snapshots as CNNs are typically trained (although see Karpathy et al. (2014); van Assen et al. (2018a)). It is much more likely that visual representations are learned through unsupervised processes, and this may have a critical effect on the internal representations that the visual system learns.

A third important difference between the artificial neural networks and humans lies in the nature of the task that the networks are trained on. Human vision is not tailored solely to the task of distinguishing mirror from glass objects, whereas here, we trained the networks on a binary classification, effectively separating the entire world into two possible states. The representations that optimize performance on this task may well be quite general purpose, as has been found with neural networks optimized for object recognition, which can easily be repurposed for other tasks, such as action recognition (Simonyan and Zisserman, 2014) and image semantic segmentation (Dai et al., 2015). Nevertheless, it is quite possible that in being trained on such a constrained task, the networks learned representations that do not resemble human visual processes.

Future work should use a combination of unsupervised learning, more naturalistic objective functions and network architectures that more closely resemble primate cortex to tease these possibilities apart. Although artificial neural networks have substantial potential as models of human visual processes, they still have important shortcomings. They should be seen not so much as an accurate working model of human brain processes, but rather as an experimental platform for further research, much as animal models of neurological disorders are.

2.4 Methods

2.4.1 Observers

Lab experiments (i.e., not ‘crowdsourcing’)

60 observers were students of Justus-Liebig-University Giessen and Toyohashi University of Technology with normal or corrected-to-normal vision. All experimental protocols were approved by the Ethics board at Justus-Liebig-University Giessen and were conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Observers in the former were paid 8 Euro/hour. Informed consent was obtained from all observers.

Crowdsourcing

247 participants were recruited via the Clickworker platform and were paid 1.2 Euro each. Before the beginning of the experiment participants were shown a consent form that explained the purpose and procedure of the experiment, as well as the uses and benefits of their participation. All participants

that took part in the experiment agreed to these conditions and that their data be recorded and stored anonymously for research and publication in scientific journals.

2.4.2 Stimuli

Renderings

Images were rendered using Mitsuba renderer (Jakob, 2010). We selected 1,583 objects from Evermotion (https://evermotion.org) and 253 light fields for the illumination from the Southampton-York Natural Scenes (SYNS) dataset (Adams et al., 2016) and the other sources (Debevec, 1998; Debevec et al., 2000) (see also A.1). For the 'mirror' objects, the BSDF was a 'conductor' model with 100% specular reflectance. For the 'glass' material, BSDF was 'dielectric', with internal refractive index of 1.5. Objects were uniformly scaled to fit within the unit sphere, and placed at the origin. The camera was randomly located at a position between 30 and 60 degrees elevation angle and any azimuth with a constant distance of 2 units in Mitsuba. The sampling count was 512 per pixel with the Sobol Quasi-Monte Carlo sampler. The reconstruction filter was set as Gaussian. The renderer generated the final image, at 256 × 256 pixel resolution with gamma correction (Reinhard et al., 2002). Then, they were resized to 64 × 64. Note that mirror and glass images were paired using same object and illumination but different camera locations to avoid that the classifiers simply learn a pixel difference between mirror and glass images. We screened all images and excluded a small number of images with rendering artifacts. The final dataset contains 753,696 images.

GAN images

Two generative adversarial networks, GANs (Goodfellow et al., 2014; Radford et al., 2015) were trained to synthesize images that they could not distinguish from a given training set of renderings. Specifically, the rendered images (see 2.4.2) were split into two subsets (mirror and glass) as the training sets (376,848 images in each). The network architecture and the hyper parameters were same as a previous network (Radford et al., 2015), except for the minor modifications in the standard Tensorflow DCGAN implementation that avoid the discriminator of the network converging too fast (Kim, 2017). After 20 epochs training, we then generated 700 images from each GAN by inputting random noise vectors to create a total of 1,400 images, which were then rated by humans.

2.4.3 Apparatus

Stimuli were displayed on a 27-inch liquid crystal display (Eizo CG276 at TUT and CG277 at JLU) using factory default settings with a resolution of 1920 × 1200 pixels. Stimulus presentation was controlled by MATLAB using Psychtoolbox 3.0 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997).

2.4.4 Experiment 1: Random renderings

Ten observers participated (7 women; age range: 19 to 38 years; average 25.1 ± 5.2 years) in the lab. We randomly selected 1,000 images from the dataset (500 mirror and 500 glass images) and presented them in random order to each observer (i.e., each image was rated by all 10 observers). A uniform gray with a fixation cross was displayed as a background. The observers were asked to rate each stimulus with a five-point scaling (glass to mirror) by pressing a corresponding key on the keyboard. They could respond at any time, but the stimulus disappeared after one second.

2.4.5 Experiment 2: Diagnostic images

The purpose of this experiment was to create a 'benchmark' set of diagnostic images, with (1) a uniform distribution of appearances ranging between mirror from glass; (2) perceptual appearance that is decorrelated from the true material class ('ground truth') and (3) consistent judgments across observers. Identifying images that correlate with ground truth (upper left and lower right quadrants of matrix in Figure 2.3B) is straightforward, as humans are generally very good at distinguishing mirror vs glass for our renderings. Thus, most of the procedure revolves around finding images that systematically yield errors (i.e., the upper right and lower left quadrants of Figure 2.3B). To achieve this, we ran two parallel series of experiments using renderings (series A) and images generated by GANs (series B), respectively. Each series starts with a large number of images, with images being progressively excluded in each round, to arrive at a much smaller final set covering the desired distribution (see also Figure A.3 and A.4).

Round A1 (Rendering ratings)

Twenty observers (all men; age range: 21 to 26 years; average 23.1 ± 1.4 years) participated in the lab. We randomly selected 30,000 renderings from the dataset (50% mirror, 50% glass) and distributed 1,500 images to each observer. The procedure was the same as in experiment 1, except that the task was a three-way judgment ('mirror', 'glass', or 'hard to recognize'). The last option was used to exclude rendering artifacts for further rounds (2.9% of the images were excluded here). Figure A.6A shows results of this round and 11,192 images (randomly selected 2,798 images in each bin except 'hard to recognize') moved ahead the round A2.

Round A2 (Rendering ratings)

247 crowdsourced participants observed the stimuli selected by round A1, and were asked to rate them with a five-point scale (glass to mirror). They were shown 100 images - 98 test images and 2 catch trial images, consisting of photographs with clear mirror or glass appearance. Only 5,586 images that were rated by at least three crowd-workers were analyzed further. Figure A.6B shows rating results of this round. We selected 522 images, in which ground truth are actual material but often seen as the other material. Specifically, 261 mirror images which rating score was ranged from 0.0 to 0.4 (often seen

as glass) and 261 glass images which rating score was ranged from 0.6 to 1.0 (often seen as mirror). These selected images were moved ahead the round A3.

Round A3 (Rendering ratings)

Ten observers participated in the lab (9 women; age range: 21 to 30 years; average 24.8 ± 2.8 years). The procedure was the same as in experiment 1 (rating task). The experiment was composed of 1,566 trials (3 trial × 522 images from round A2), and all trials were randomly ordered. Figure A.6C shows rating result of this round. We selected 102 images as well as Round A2. Specifically, 51 mirror images which rating score was ranged from 0.0 to 0.6 (often seen as glass and ambiguous) and 51 glass images which rating score was ranged from 0.4 to 1.0 (often seen as mirror and ambiguous). These selected images were included the diagnostic image set.

Round B1 (GAN-image screening)

Some GAN-generated images resemble textures rather than objects with distinct material properties. The purpose Round B1 was to exclude such images from subsequent rounds. Ten observers participated in the lab (8 women; age range: 20 to 32 years (average 24.8 ± 4.1 years). The stimuli were 1,400 images generated by GANs (see 2.4.2). The procedure was the same as in experiment 1, except that the task was to indicate in a binary decision whether the object shape and material are recognizable or not). Figure A.7A shows result of this round. Based on the responses, 500 images that were judged to be recognizable by at least six observers were moved ahead to Round B2.

Round B2 (GAN-image rating)

Ten observers participated in the lab (all women; age range: from 21 to 34 years; average 25.1 *pm* 3.8 years). The stimuli were 560 images including 500 GAN images from Round B1 and 60 renderings (30 mirror and 30 glass images) from round A2, which had received ratings that were highly consistent with ground truth. The procedure was the same as in experiment 1 (rating task). The experiment was composed of 1,680 trials (3 trial × 560 images from Rounds B1 and A2), and all trials were randomly ordered. Figure A.7B shows result of this round. We selected 95 images (19 images from each bin) as GANs' images.

Final Diagnostic Image set

The two streams of experiments resulted in a final diagnostic image set of 265 images including both mirror and glass renderings, and GANs images with prediction score uniformly distributed from 0.0 to 1.0 (Figure 2.3B and Figure A.3). These are composed of 68 veridical images (from experiment 1), 102 illusory images (from Round A1-A3), and 95 GANs' images (from Round B1 and B2).

2.4.6 'Hand-Engineered' Classifiers

We developed three different classifiers (Figure A.2A), Color-Hist, Port-Sim, and CNN with manually selected parameters. Color-Hist used eight features: mean, variance, skewness and kurtosis of intensity and color saturation from 64 × 64 RGB image. The features of Color-Hist were z-scored across all images. To get the features of Port-Sim, we first used the texture analysis algorithm of Portilla and Simoncelli (Portilla and Simoncelli, 2000) to extract 3,381 higher-order image statistics. These were z-scored and the number of dimensions reduced to 1,052 by principal component analysis (cumulative explained variance of complete image set >99%). For both Color-Hist and Port-Sim, a logistic regression was trained to distinguish mirror from glass based on the ground truth labels. CNN was defined as a three-layer convolutional neural network with 64 × 64 RGB image input and the binary (mirror vs. glass) classification output. The network architecture and training hyperparameters are shown in Figure A.2B. All classifiers trained and tested with 2-fold cross validation, which was repeated 10 times with different randomly selected training and test sets (images that were shown to human were excluded here). The final output—a prediction score ranging from zero (as glass) to one (as mirror) —was averaged across training repetitions.

2.4.7 Identifying optimal CNN models through Bayesian hyperparameter search

We used Bayesian hyperparameter search through a space of feedforward architectures to identify which CNNs correlated best with humans (using MATLAB R2017b with Neural Network Toolbox and Statistics and Machine Learning Toolbox; Figure A.5A). The objective was to maximize the correlation coefficient between CNN and human on the diagnostic image set. The network architectures were basically the same as the CNN in experiment 1, except that we parametrically varied the 'depth', i.e., the number of layers (convolution, batch normalization, rectified linear unit, and max pooling layers) before the first fully connected layer, from 1 to 12. Note that the max pooling layers were only used up to 3 layers (the last 3 layers) because of the size constraints of the filters. For each depth, we ran 200s iterations of the Bayesian hyperparameter search (i.e., 200 CNNs were generated with different hyperparameters, in search of the optimal values for each depth (Figure A.5B)). Each CNN was trained and tested with same training and test set. Having identified the optimal hyperparameter values, we then trained 30 new CNNs with those optimal parameters, but with different random initial weights and training/test images. These are the networks that are reported in Figure 2.3C.

2.4.8 Representational similarity analysis (RSA)

We defined two different representational dissimilarity matrices (RDM), a 1st-stage RDM to identify how dissimilarity relationships between images are represented in each human and classifiers; and a 2nd-stage RDM ('Classifier Correlation Matrix', CCM), characterizing how similar the 1st-stage RDMs are across different humans and classifiers, allowing us to compare their internal representations. We also defined a 'Layer Correlation Matrix' (LCM) as well as CCM to compare different layers. The 1st-

stage RDM was defined as Euclidean distance of prediction scores (final output) from each classifier or average of observers' response from human. The 2nd-stage RDM was defined as a dissimilarity, which was one minus Pearson's correlation between each 1st-stage RDMs.

The number of dimensions of the 2nd-stage RDM was reduced to three or two dimensions using MDS and we visualized a relationship between the targets as three-dimensional space in Figure 2.2C (1,000 images from the random image set) and Figure 2.4D (265 images from the diagnostic image set), and as two-dimensional space in Figure 2.5A (265 images from the diagnostic image set).

Chapter 3

Static visual cues

A part of this chapter has been presented as:

Tamura, H., Prokott, K.E.,& Fleming, R.W. (2018). Modulating luminance and color saturation disambiguates mirror and glass, *European Conference on Visual Perception 2018 (ECVP 2018).*

The visual system uses static cues of some kind to recognize object material categories because we can easily distinguish them at a glance. Here, we examined what static cues contributes to distinguishing mirror (perfect specular surfaces) from glass (transparent and refracted objects) using an image editing approach. First, we observed rendered and generated images of the data set (Chapter 2) to find common appearances, which can verbally express each material feature. We found that the common features on a pixel-by-pixel basis strongly relate to each material's impression but not the actual material. Then, we developed an image editing method using luminance and color saturation profiles along the object contour to its center, to change the material appearance to be mirror-like or glass-like. To test the effect of the editing, we ran a series of psychophysical experiments. The results were that the mirror-like images were perceived more as being a mirror than the original with the background. In contrast, the glass-like were perceived more as being glass without the background. These results suggest that the luminance and saturation profiles contribute to distinguishing mirror from glass as static cues with an interaction between materials and their backgrounds.

3.1 Introduction

3.1.1 Common appearances based on our perception

It is important to clarify what kind of static cues the visual system uses to recognize object material categories from a visual image. This is one of the main streams of material perception research. For example, simple image statistics such as the skewness of the luminance histogram and the skewness of sub-band filter outputs contribute to estimating surface properties (glossiness) (Motoyoshi et al., 2007). Here, we examine what visual static cues contribute to distinguishing mirror from glass using an image editing method.

First, we observed the images from the data set in chapter 2, and looked for some common appearances among the images, which can verbally express certain kinds of material features. Figure 3.1 shows example images from the category of the mirror, glass, and GANs images with the scores rated by the previous series of experiments in chapter 2. We found that the images highly seen as mirror (on the rightmost column in Figure 3.1) have in common a flat surface on the object. This is highly likely to be based on a similarity to common mirrors in our daily life. Moreover, they show some bluish pixels on the surface, suggesting that a blue sky is reflected on the object surface from natural illumination. This type of illumination also gives the object's coloring a gradation (e.g., from structural colors). Thus, we predict that these findings would be highly related to determining an impression of a mirror.

In contrast, in the leftmost column in Figure 3.1, the images highly seen as glass have bright pixels on the object fringe. This area strongly represents the specular highlights on the surface. Moreover, the pixels inside the object contour have the same color as the background, suggesting that they play an important role in determining an impression of glass because we see the background through the object without any doubts.

It is an important point that these descriptions are not related to their original actual material (mirror or glass) and the generating methods (computer graphics rendering or GANs generating). The material impressions were simply derived from the regions of the images. Therefore, some static visual cues contributed to distinguishing mirror from glass on a pixel-by-pixel basis.

3.1.2 Luminance and color saturation profiles

Next, we quantified the characteristic features differentiating mirror and glass using pixel information from the images, and employed luminance and color saturation profiles along trajectories from the object contour to its center. Figure 3.2 shows a quantified difference between mirror and glass based on this index. We randomly selected 1,000 images from the data set (500 images of each material), and took their average luminance and saturation profiles along the trajectories from the object contour to its center. From the average luminance profile (in the left panel of Figure 3.2), the mirror images tend to increase in luminance from the object contour to the center. In contrast, the profile of glass images has an opposite tendency, which decreases along the trajectory from the object contour to the

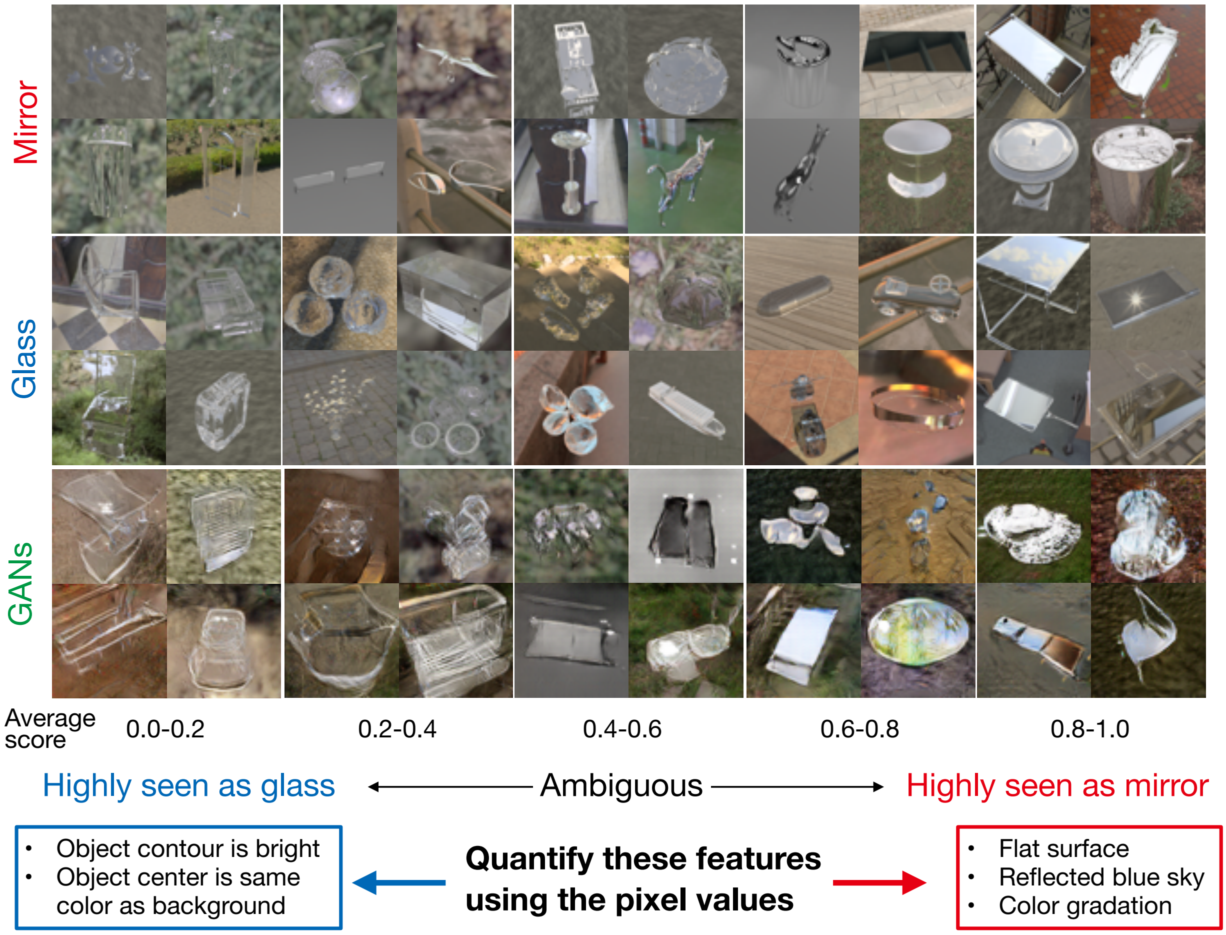


Figure 3.1: Example images

The horizontal direction indicates the average score by the human observers in chapter 2. The vertical direction indicates the different image classes (the mirror rendering, glass rendering, and GANs images) from top to bottom. The four images in each panel were selected from each bin and ranged from 0 to 1. We briefly describe some image features or aspects from the image observation below the example images.

center. Additionally, the average saturation of the mirror images was relatively higher than that of the glass (in the right panel of Figure 3.2). These findings demonstrate that there is a difference between mirror and glass derived from the luminance and saturation profiles. Therefore, we hypothesized that these quantified differences are highly likely to provide each mirror and glass impression. To test this hypothesis, we developed an image editing method to change the static cues. After that, we ran a series of experiments using images in which the static cues were edited to verify their effects.

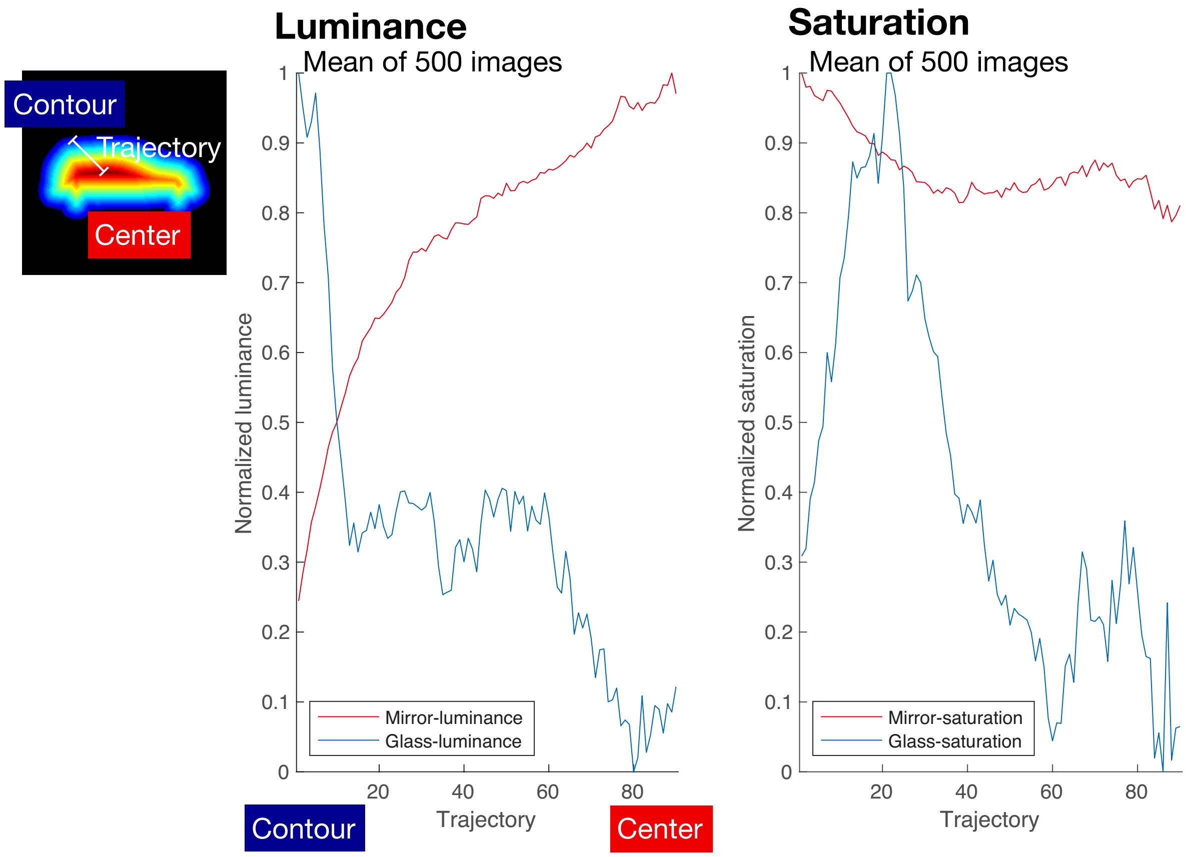


Figure 3.2: Luminance and color saturation differences between mirror and glass

The left panel shows the average luminance profile of mirror and glass along the trajectory from the object contour to its center. The horizontal axis indicates the trajectory. The vertical axis indicates a normalized luminance, which is based on an average across all possible trajectories around the object contour in an image. This shows the average across 500 images in each material. The right panel shows the average color saturation profile. The format is the same as the left one.

3.2 Methods

3.2.1 Editing material appearance

If we modulate the luminance and saturation profiles to change an image appearance to be more mirror-like or glass-like, the modulated information is regarded as static cues for determining mirror vs. glass. Here, we propose an image editing method modulating the luminance and saturation profiles of the images, which contain an object with certain materials. Figure 3.3 illustrates the editing procedure as a flow diagram. The procedure is described as follows:

1. Convert an original RGB image to HSV image.
   - V of the HSV image is defined as the luminance (or the intensity) of the image.
   - S of the HSV image is defined as the color saturation of the image.
2. Obtain an alpha map to separate a region of the object and its background from the original RGB.
3. Obtain pixel profiles of the object contour to its center by repeating binary erosion on the alpha map.
4. Modulate the V and S of the item 1 and obtain V' and S'. Specific modulating functions we used are shown in Figure B.1.
5. Reconstruct the HSV image from the original H, V', and S'.
6. Convert the HSV image of the item 5 to RGB image.

3.3 Human psychophysics

3.3.1 Overview

We designed two stages of psychophysical experiments. In the first stage, we selected rendering images as stimuli from the data set (in experiments 1A and 1B). In the second stage, we tested how impressions of the material appearance changes using the image editing method. We prepared two conditions as "with background" (in experiment 2A) and "without background" (in experiment 2B) to test how the material appearance is affected by the background. In the latter case, the background was excluded and a uniform gray was used instead.

3.3.2 Apparatus

Stimuli were displayed on a 27-inch LCD (CG277, Eizo) using factory default settings with a resolution of 1920 × 1200 pixels. Stimulus presentation was controlled by MATLAB using Psychtoolbox 3.0 (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007).

3.3.3 Experiment 1

All observers had normal or corrected-to-normal acuity. All experimental protocols were approved by the Ethics board at Justus-Liebig-University Giessen and were conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Informed consent was obtained from all observers.

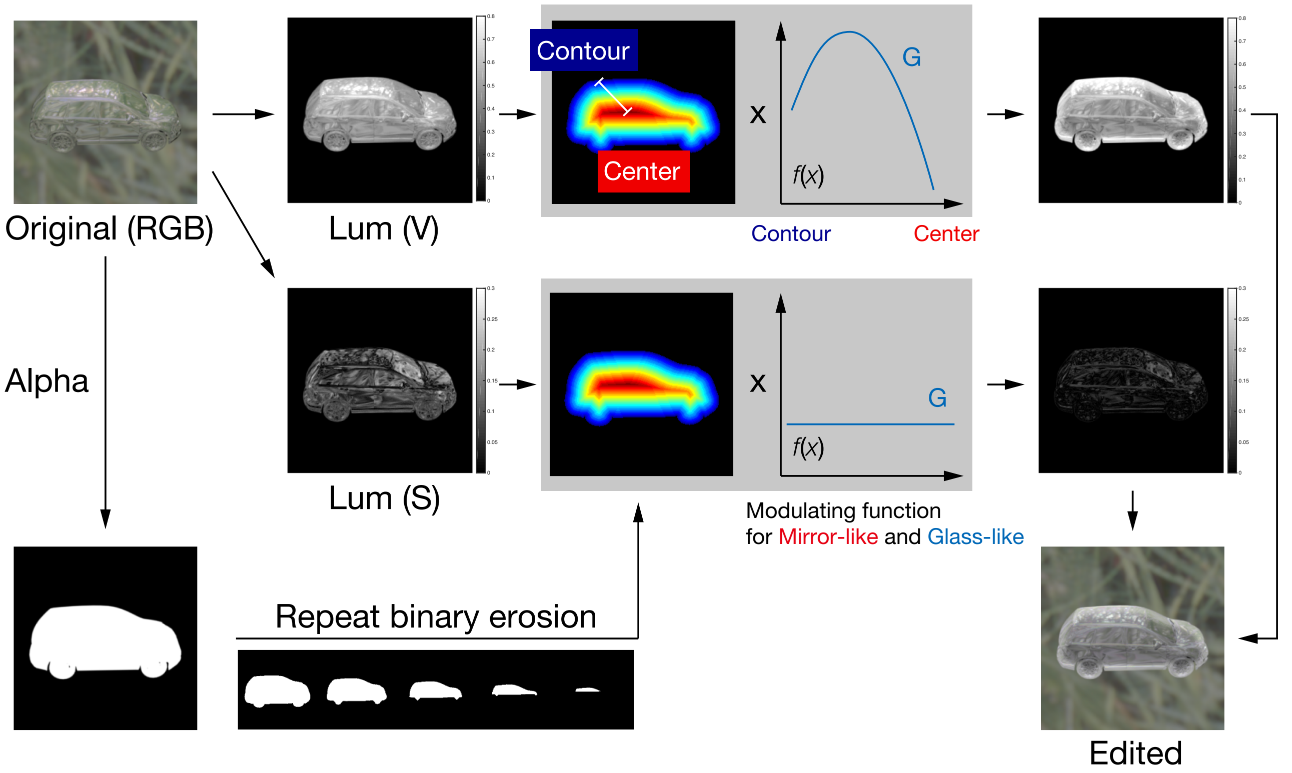


Figure 3.3: Flow diagram of editing material appearance

This is an example to edit the material appearance from an original image to a glass-ish edited one. Note that the images of the modulating function (described as "$f(x)$") were overdrawed to make the description easy to understand. The actual functions and values were shown in Figure B.1.

Observers

Experiment 1A (pre-screening using the images rendered by computer graphics: Ten naïve observers participated in this experiment. Their ages ranged from 18 to 31 years (average 25.6 ± 4.2 years).

Experiment 1B (distinguishing mirror from glass materials using the images rendered by computer graphics): Fifteen naïve observers participated in this experiment. Their ages ranged from 20 to 32 years (average 24.7 ± 3.7 years).

Stimuli

For experiment 1A, the stimuli were selected from the dataset and the number of images was 1,400, that is, 700 images for each material (mirror and glass) with 256 × 256 pixel resolution.

Then, we selected 169 mirror and 203 glass images as candidates for stimuli that met the criterion that all observers answered yes (the number of counts was 10) in experiment 1A (see Figure 3.4). Then, we reduced the number of glass images to 169, the same as the number of mirror images. Finally, 338 images were used as stimuli for experiment 1B.

Procedure and Task

The purpose of experiment 1A was to exclude stimuli that had an ambiguity depending on the object shape and rendering artifact. A grey background and a fixation cross were displayed and remained throughout the experiment. The first trial was started by a key press and a target stimulus was randomly presented. The observers were asked, "is the object shape and material appearance clearly recognizable?" and responded with a key press from a numerical keyboard to indicate yes or no. The next trial was started with a further key press. The experiment consisted of 1,400 trials (one trial × 1,400 images), and all trials were randomly ordered.

For experiment 1B, the procedure was the same as in experiment 1A, except that the task was replaced with a material distinguishing task (mirror or glass). The experiment consisted of 1,690 trials (five trials × 338 images), and all trials were randomly ordered.

3.3.4 Experiment 2

Observers

Experiment 2A (with background): Eight naïve observers participated in this experiment. Their ages ranged from 21 to 37 years (average 25.9 ± 4.9 yeas).

Experiment 2B (without background): Ten naïve observers, who had not participated in experiment 2A, participated in this experiment. Their ages ranged from 21 to 27 years (average 23.7 ± 1.9 years).

Stimuli

The images used in experiment 1B were sorted into five subsets by the average rating scores among all observers, such as 0.0 - 0.2, 0.2 - 0.4, 0.4 - 0.6, 0.6 - 0.8, and 0.8 - 1.0. Then, we randomly selected 25 images in each bin; in total, 125 images were selected as the stimuli based on the results of experiment 1B. Note that we did not consider the actual material of the original images in these selections. Finally, the images were edited to mirror-like and glass-like by the method we proposed above.

Procedure and Task

A pair of two images was presented on the display and the observers were asked which looked more like a mirror (mirror-like task), or more like glass (glass-like task). The pair had an unedited image (original) and a mirror edited (mirror-like) one with the same objects in the former task. In the latter task, instead of a mirror-like image, a glass edited (glass-like) image was used.

Experiment 2A consisted of 1,250 trials (five trials × 125 images × two tasks), each task was separated as two different blocks, and all trials in each block were randomly ordered. Experiment 2B was the same as 2A except the stimuli were different (with/without background).

3.4 Results and Discussion

3.4.1 Image selection for the material editing

As the stimuli, we used images that possessed a clear object shape and material properties, because they increased the effect of changing the material appearance using the image editing method. Therefore, we performed an image selection to obtain a few hundred of images from the data set, on the basis of the results of psychophysical experiments. Figure 3.4A shows the histogram of yes answers from the observers. We found that 372 images were answered with yes (in the rightmost bin) from all observers. This also shows that the number of images in each bin decreased depending on the number of yes answers. A few example images selected from bins 1, 5, and 10 are shown in Figure 3.4.

After that, we selected 338 images (169 images for each material) for the stimuli from the bin in which all observers answered yes. In the next experiment, the observers were asked to judge the object material (mirror or glass). Figure 3.4B shows the rating score, which is clearly bimodal and distributed with 84.6% accuracy. The number of images with ambiguous responses (in the range between 0.4 - 0.6) was smaller than the result of chapter 2 (see also Figure 2.1B). We suggest that this difference depends on the image resolution of this experiment (256 × 256) and the previous one (64 × 64) because the observers could use more precise information from fine pixels, which were made by light reflection, refraction, and transmission, except that at this point, there was the almost same tendency as in that experiment, even though the task was different (a binary judgment or a five-point scale rating).

3.4.2 Test of the edited images

With background

The stimuli were labeled with the average score of the observers in experiment 1B. On this basis, 125 images were selected as candidate images for image editing (25 images from each bin divided by 0.2 steps). Then, we edited these images using the proposed method described above to make the image look more like a mirror (mirror-like editing) or more like glass (glass-like editing). The observers were asked to judge which looked more like a mirror between the original image and the same image edited as mirror-like in a mirror judging task. In the same way but in a different block, they were also asked to judge which looked more look like glass between the original image and the glass-like edited image in a glass judging task (see 3.3.4). Next, we estimated how effective the editing was by a chose-edited ratio.

Figure 3.5A shows the results of mirror-like and glass-like images with a background condition (in

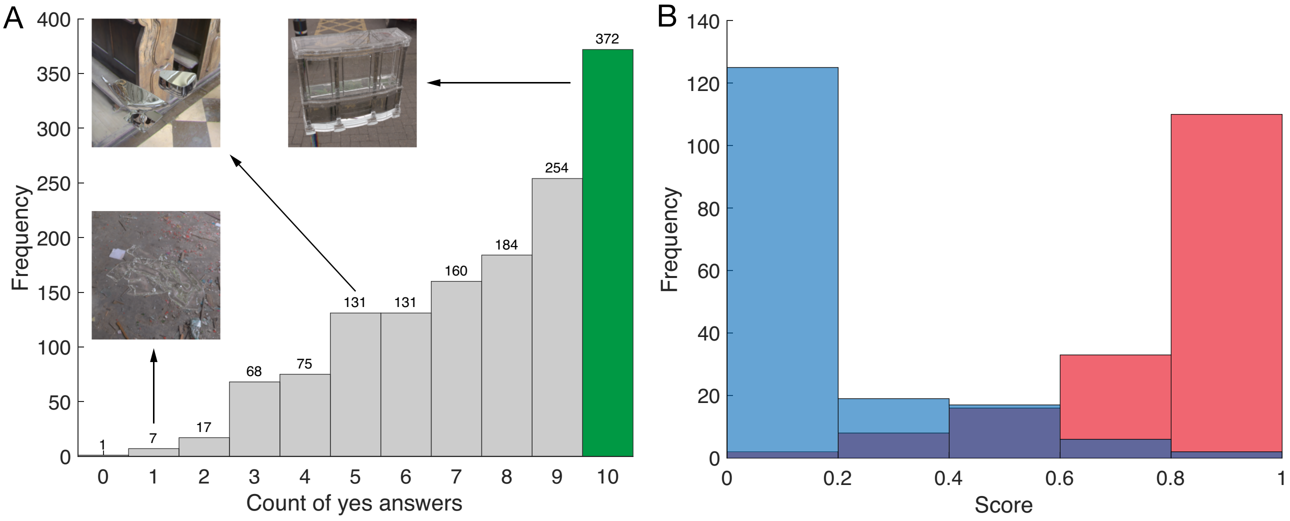


Figure 3.4: Image selection results

(A) Image screening. The horizontal axis indicates the count of yes answers, which means how clearly recognizable the stimulus was. The vertical axis indicates the number of images (frequency). This figure also shows three example images from bins 1, 5, and 10. (B) The results of the material judging task. This panel shows the histogram of the average responses across all observers with five bins.

the left and right panels, respectively). In the mirror judging task, the average of all observers was higher than the chance level (0.5). Moreover, the results of six out of eight observers were significantly higher than the chance level in at least one of these bins (binomial test, $p < 0.05$).

By contrast, in the glass judging task, the average of all observers was almost lower than the chance level. Specifically, even in the highest bin (the most seen as the mirror), the average ratio was almost the same as the chance level. Only one out of the eight observers was significantly higher than the chance level (binomial test, $p < 0.05$). These results suggest that some static cues related to the mirror were enhanced by the material editing and helped to distinguish mirror from glass with the background condition. However, the editing was only effective for the mirror judging task in this case.

Figure 3.5B shows the average ratio of each stimulus. In the mirror judging task, 40 out of 125 images were significantly higher than the chance level (one-tailed t-test, $p < 0.05$). Moreover, the images with a lower score (seen as the glass) tended to have relatively a higher ratio. In the glass judging task, although only two out of 125 images were significantly higher than the chance level (one-tailed t-test, $p < 0.05$), they showed an opposite tendency to the mirror judging task between the original score and the chose-edited ratio. This suggests that the magnitude of the editing effect depends on the original score.

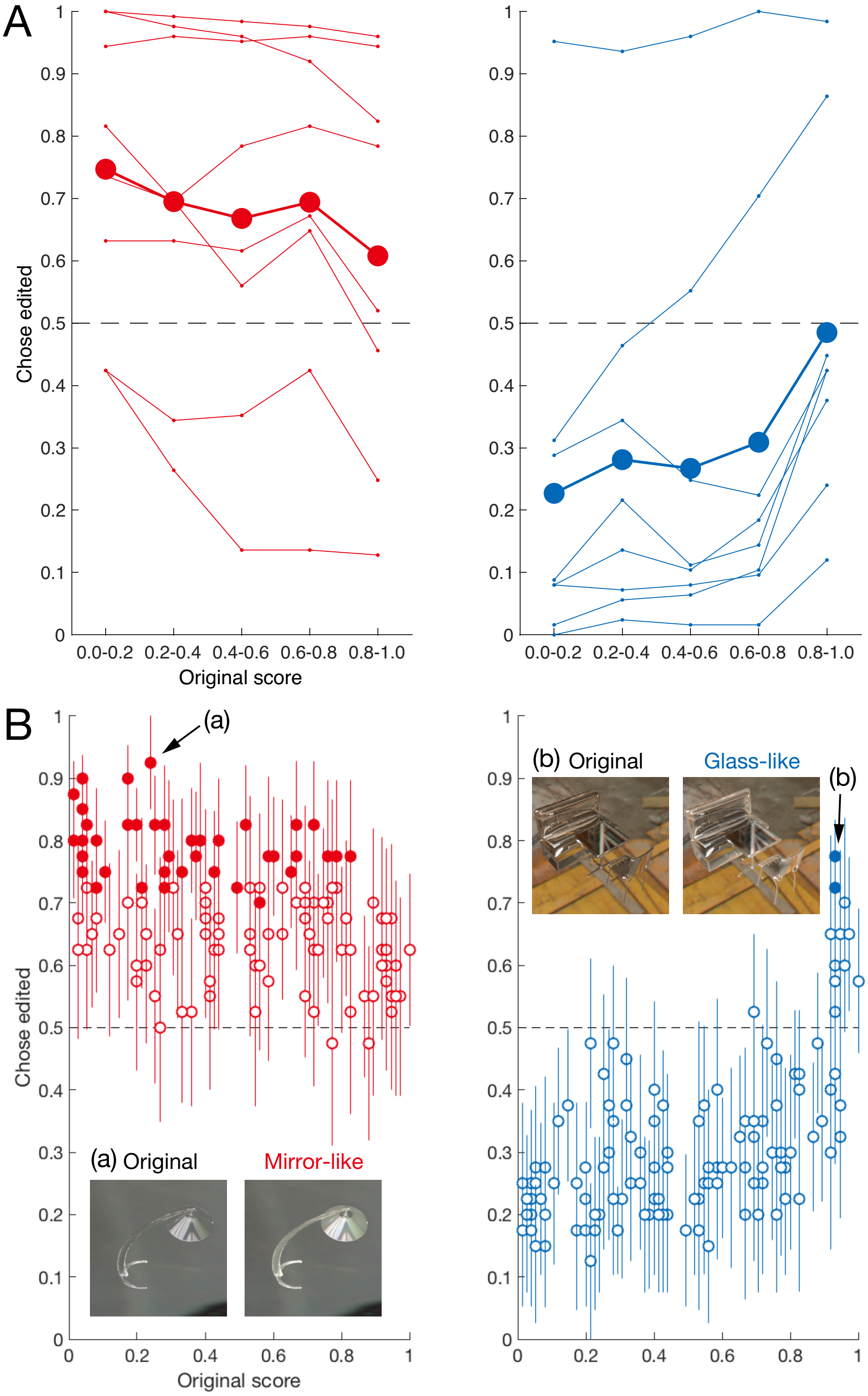


Figure 3.5: Results with the background condition

(A) The average and each observer's chose-edited ratio in each material. The horizontal axis indicates the original score of stimuli, which were rated by experiment 1B. The vertical axis indicates the average chose-edited ratio in each bin. The thicker line shows the average result of all observers and the thin lines show the results of each individual observer. The left and right panels are mirror-edited and glass-edited conditions, respectively. (B) The average chose-edited ratio in each image. The filled and blank circles indicate whether significant difference or not from 0.5 (one-tailed t-test, $p < 0.05$), respectively. The error bars represent the standard error of the mean across all eight observers. The images labeled as (a) and (b) are the highest of all in each task, respectively. The left image indicates the original and the right one indicates the edited one (mirror-like or glass-like). The other format is the same as in A.

Without background

One of the reasons the glass-like images were less effective with background (Figure 3.5) was that a higher contrast between the object edge and its background made the object shape stands out from the background. Then, it provided an unnatural impression for the material judgement. Therefore, if the contrast was lower, the editing effect of the glass-like images might be bigger. To test this, we ran the same experiment excluding the background. We simply changed the background to a uniform gray (without the background condition).

Figure 3.6 shows the results without the background condition. We found that the average response of all observers of the glass-like images was higher than the chance level in the glass judging task (in the right panel of Figure 3.6A). Then, six out of ten observers were significantly higher than the chance level in at least one of these bins (binomial test, $p < 0.05$). Moreover, 49 out of 125 images were significantly higher than the chance level (one-tailed t-test, $p < 0.05$). This suggests that in the glass judging task, the material editing without background is basically effective but it depends on the observers.

However, in the mirror judging task, five out of ten observers were significantly higher but their average was almost at the chance level. Moreover, a one-tailed t-test revealed that there no image was significantly higher than the chance level. This means that there were large individual differences between the observers. Furthermore, the material condition (mirror and glass) and the background condition (with/without background) had an interaction. In fact, the mirror-like images with the background and the glass-like images without the background were only effective under these specific conditions. This suggests that background information works as one of the cues for distinguishing between mirror and glass.

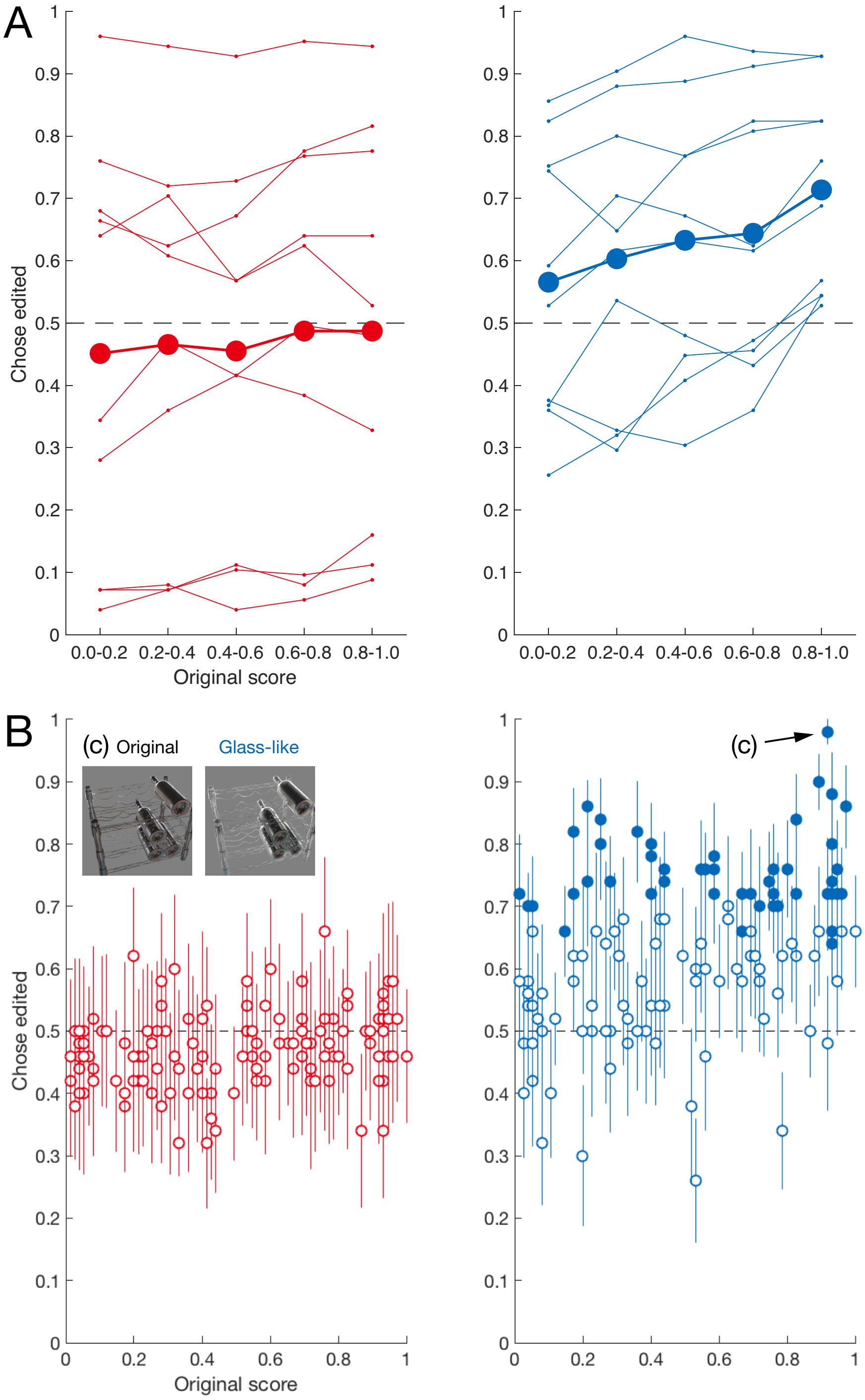


Figure 3.6: Results without the background condition

Two images (the original and glass-like) labeled as (c) are from the highest of all stimuli in the glass judging task. The other format is the same as in Figure 3.5.

3.5 General Discussion

In this study, we hypothesized that the luminance and saturation profiles along the object contour to its center are static visual cues to distinguish mirror from glass materials. On this basis, we proposed an easy and simple material editing method, and tested the magnitude of the editing effect by human psychophysics. The result was that the mirror-like images were perceived more as mirror than the original images with the background condition. In contrast, the glass-like images were perceived more as glass than the original images without the background condition. A glass material object has transparent and refractive components, and a distorted image on the object surface is strongly related to the background behind the objects (e.g., the distortion field: Fleming et al., 2011). Likewise, the relationship between the materials and with or without background condition suggests that this kind of image cue plays an important role in distinguishing glass from other materials.

Previous studies reported that a simple luminance modulation can change an image's material appearance. For example, Fleming and Bülthoff (2005) reported that the image appearance of an opaque object can be edited to a translucent object using a luminance modulation only. Similarly, Motoyoshi et al. (2005) reported using luminance re-mapping to make an opaque object translucent. Moreover, this also makes a metallic surface like the mirror material. Although our method is similar to these in that the editing provides some permeability to the object, this is the first study to propose modulation to make clear transparent objects, such as glass. In addition, our results suggest that color saturation also plays a role for determining materials (see also Zaidi (2011)). From observation of the characteristic images in the data set, the color saturation of an image on an object surface with mirror material tends to be high because it reflects from an image of various natural illuminations, which has a blue sky or the green of a forest. Thus, we suggest that the visual system uses such heuristics (Fleming, 2012) to distinguish mirror from glass.

There was a big variance between the observers in Figures 3.5 and 3.6. We speculate that the perceptual border between mirrors and glass is ambiguous and relatively depends on each observer's top-down bias. Furthermore, common images of mirrors and glass may not be very specific. Even if there are enough visual cues to determine the material, it is possible that some of them are highly effective for seeing an image as a mirror but others are not so. To decrease the bias effect, we could show some example images of mirror and glass as a catch trial to provide observers with a specific material image. We could also employ a rating task instead of the 2AFC task because this way can more precisely represent the magnitude of perception between mirror and glass.

The material editing method we proposed in this study does not depend on an object's shape and the number of regions of the object. This is a simple modulating method and possible to use generally. Moreover, the method only needs an RGB image and its alpha map for editing. If we include any other information related to a cue for determining the materials, we could get more natural material appearances and consistent results. For example, an image modulation including three-dimensional shape information of the target object could work well because it contributes a cue to albedo estimation (Marlow and Anderson, 2015; Marlow et al., 2015). Object depth estimated by a two-dimensional image could provide more natural luminance and saturation biases for the image.

Through the series of experiments in this study, we found that the luminance and saturation profiles along the object contour to its center disambiguate mirror and glass. In mirror-like editing, increasing luminance and saturation can enhance pixels, which has a brighter luminance as specular highlights or is reflected from illumination from the natural environment. These contribute to providing the original images a mirror-like impression as static cues. In addition, in glass-like editing, increasing luminance near the edge of the object and decreasing saturation work as static cues to provide the original images a glass-like impression. This editing enhances the specular highlights on the object surface and emphasizes light transmittance from the background through the object. These findings suggest that common image features of mirror and glass derived from luminance and saturation are used as static cues to distinguish the two materials.

Chapter 4

Dynamic visual cues

A similar version of this chapter has been published as:

Tamura, H., Higashi, H., & Nakauchi, S. (2018). Dynamic Visual Cues for Differentiating Mirror and Glass, *Scientific Reports*, 8, 8403. DOI:10.1038/s41598-018-26720-x

Mirror materials (perfect specular surfaces such as polished metal) and glass materials (transparent and refraction media) are quite commonly encountered in everyday life. The human visual system can discriminate these complex distorted images formed by reflection or transmission of the surrounding environment even though they do not intrinsically possess surface colour. In this study, we determined the cues that aid mirror and glass discrimination. From video analysis, we found that glass objects have more opposite motion components relative to the direction of object rotation. Then, we hypothesised a model developed using motion transparency because motion information is not only present on the front side, but also on the rear side of the object surface in the glass material object. In materials judging experiments, we found that human performance with rotating video stimuli is higher than that with static stimuli (simple images). Subsequently, we compared the developed model derived from motion coherency to human rating performance for transparency and specular reflection. The model sufficiently identified the different materials using dynamic information. These results suggest that the visual system relies on dynamic cues that indicate the difference between mirror and glass.

4.1 Introduction

Visual information reaching the human retina potentially contains temporal components due to the motion of objects and/or observers. It has been revealed that several dynamic cues contribute to perception of x-from-motion: e.g., depth, shape, and structure. Although most studies on material perception have focused predominantly on static cues to reveal how the visual system recognises materials, it is highly probable that dynamic visual cues strongly contribute to perception of certain kinds of material, e.g., glossiness (Doerschner et al., 2011; Dovencioglu et al., 2017; Doerschner et al., 2013; Dovencioglu et al., 2015; Sakano and Ando, 2010; Tani et al., 2013; Marlow and Anderson, 2016; Wendt et al., 2010), transparent flow perception (such as for liquid (Kawabe et al., 2015b) and hot air (Kawabe and Kogovšek, 2017)) from image deformation, liquid viscosity perception (Kawabe et al., 2015a; van Assen and Fleming, 2016), and stiffness (Paulun et al., 2017; Schmidt et al., 2017).

Mirror and glass have completely different optical properties. They do not intrinsically possess surface colour; thus, humans somehow differentiate mirror and glass by perception of the distorted images formed by reflection or transmission of the surrounding environment. However, the visual features the human visual system uses to achieve this task are still unknown, because the distortion of the reflected/transmitted image is largely and complicatedly dependent on the 3D shape and surroundings of the object. The visual system has constancy for surface properties under natural illumination to some extent (Fleming et al., 2003), meaning that humans can accurately distinguish various materials under natural illumination via static visual features (Hiramatsu et al., 2011; Fleming et al., 2013; Nagai et al., 2015; Tanaka and Horiuchi, 2015). Therefore, it is natural to assume that the visual system discriminates mirror/glass based on certain visual cues of the distorted images. For example, Kim and Marlow (2016) previously reported an illusion in which a specular reflected object is seen as a refracted and transparent object when it is turned upside-down. They suggested that the reflected/refracted illumination appears right side up for convex mirror surfaces whereas it appears upside down for convex glass surfaces.

This study focuses on mirror and glass materials, because both are, as stated above, quite common in everyday life and easy to discriminate under natural conditions, but both are also determined only based on visual cues existing in the image distorted by the reflection (for mirror materials) or transmission (for glass materials). The type and degree of distortion varies significantly according to the 3D shape of the object, which is generally unknown and should be recovered by the visual system. Therefore, even though it should be difficult for the visual system to discriminate mirror and glass, this is actually quite easy for humans under everyday conditions. We therefore hypothesised that, in addition to static visual cues, certain dynamic cues caused by the motion of the object and/or observers play an important role in the discrimination of these materials.

Although it was previously unknown that dynamic cues are used to directly distinguish mirror and glass, some related cues have been reported for specular reflected or transparent objects. Doerschner et al. (2011) have reported that there are characteristic differences between moving matte and shiny objects. They proposed three cues (coverage, divergence, and 3D shape reliability) on which the brain relies to distinguish those objects. In this study, we do not consider matte objects, but aim to distinguish mirror and glass, which can both be regarded as shiny. This is highly challenging, because it is more difficult to distinguish these materials using optic flows only than to distinguish matte and shiny objects (see Fig. C.1). Fleming et al. (2011) have reported that the distortion field derived from the refractive index of a thick transparent object (glass) determines human perception when judging refractive indices. Although this provides a cue for perceiving glass, this perception might arise because of similar distorted images derived from the specular surface depending on the object's shape, motion, and surrounding environment. Further, Kawabe et al. (2015b) have reported that image deformation by some specific spatiotemporal frequencies allows perception of transparent liquid layers. This is one of the cues for perceived transparency for non-rigid objects, but may also be related to rigid glass appearance. These previously proposed cues provided by motion do not directly explain the mechanism used to distinguish mirror and glass.

When a transparent object rotates about its vertical axis, the rear side of the object surface moves in the direction opposite to the object's rotational direction (e.g., an Ullman cylinder Ullman, 1979). In other words, dynamic information is not only present on the front side of the object surface, but also on the rear side of the object surface in the glass object. This is because the transparent medium of the glass object transmits the light through the object. This phenomenon is similar to motion transparency (Qian et al., 1994; Snowden, 1999). Therefore, we hypothesised that a model developed using motion information in the same manner as motion transparency (e.g., Qian and Andersen, 1994; Nowlan and Sejnowski, 1995) explains perceptual material discrimination between mirror and glass, based on the results of human psychophysics.

To identify the difference between the motion information of the mirror and glass materials, we first computed optic flows using the method previously reported by (Lucas and Kanade, 1981). Examples are shown in Figure 4.1A (see also, Movie 5.1). A mirror-like material has a specular highlight component on the surface of the object (e.g., polished metal), and glass is a transparent and refractive medium (e.g., ice). The mirror material was defined as a material with perfectly specular reflection on the surface, which is similar to an object having an infinite refractive index. The glass material was defined as a material with a transparent and refractive medium, with a refractive index of 1.5, to be similar to common glass (see Stimuli). The stimuli comprised 60 video frames rendered with three different shapes under five real-world illuminations (Adams et al., 2016; Debevec et al., 2000), and an object rotating about the vertical axis. A colour map indicating the magnitudes of the horizontal motion components in each pixel is shown in Figure 4.1B and Movie 5.2. In the mirror object, brighter luminance areas, such as specular highlights and the contour configuring the object's shape, predominantly moved in the object's rotational direction. In the glass object, the components described above and additional opposite motion components were present around the centre of the object. Here, we quantitatively express this difference using a histogram (Figure 4.1C). The histogram indicates the directions of motion of the optic flows of the mirror and glass materials. Figure 4.1C shows that there are peaks towards the right (0 rad) in the glass materials in each shape. The glass material has more components towards the right than the mirror material. This is because it is a transparent object and opposite direction components exist on the rear side of the object surface such as motion transparency.

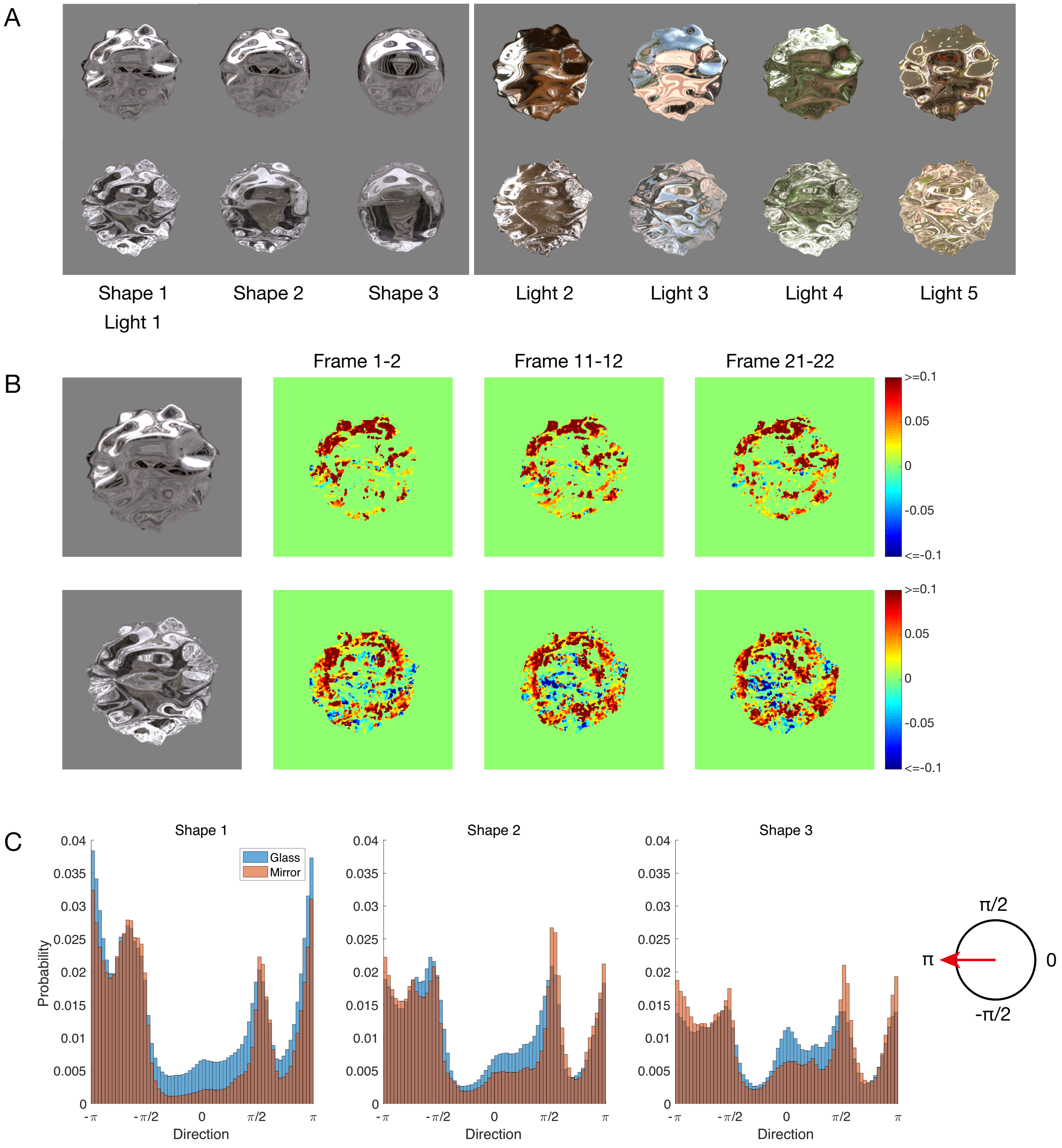


Figure 4.1: Differences between mirror and glass materials in video stimuli

(A) Examples of stimuli (see Movie 5.1 for the dynamic condition). The top row contains the mirror material objects and the bottom row contains the glass material objects. The left block shows three different shapes under illumination

1 (environment light field). The right block shows five different illuminations with object shape 1. (B) Visualisations of motion components in object rotation direction. The top column shows the mirror material objects and the bottom column shows the glass material objects. Colour maps indicate the magnitude of the motion components in each pixel. Red indicates the left direction and blue indicates the right direction. We selected three examples, frames 1-2, 11-12, and 21-22, for shape 1 under illumination 1. (C) Histogram indicating the directions motion of the optic flows of mirror and glass materials. The horizontal axis indicates the direction in radians (right is zero and left is pi). The vertical axis indicates the probability of appearance frequency. The optic flows were included for all frames and the five natural illuminations. Note that these components only move in the horizontal direction because the object was rotated around the vertical axis.

From these observations, the visual system would use the distinctive motion information to distinguish mirror and glass. If so, perceptual material discrimination performance with motion could be better than without motion. In experiment 1, we compare the performances with/without motion. Then, in experiment 2, we collect human rating data for perception of transparent and specular reflection to propose a model explaining the human perception in this task. Finally, in experiment 3 and experiment 4, we validate the model by excluding motion transparency and static cue.

4.2 Methods

4.2.1 Observers

Experiment 1 (perceptual material discrimination between mirror and glass materials): Ten naïve observers participated in this experiment. Their ages ranged from 21 to 25 years (average 22.4 ± 1.4 years).

Experiment 2 (rating for perception of transparent and specular surface): Ten naïve observers, who had not participated in experiment 1, participated in this experiment. Their ages ranged from 22 to 25 years (average 23.2 ± 1.1 years).

Experiment 3 (rating using edited videos): Ten naïve observers participated in this experiment. Their ages ranged from 23 to 26 years (average 24.2 ± 1.1 years).

Experiment 4 (perceptual material discrimination using binary noise stimuli): Ten naïve observers participated in this experiment. Three observers were excluded because their performance for the static condition was equal to or higher than that for the dynamic condition and quite different from the others. Thus, the final sample consisted of seven observers aged from 23 to 26 years (average 24.1 ± 1.1 years).

All observers had normal or corrected-to-normal acuity. All experimental protocols were approved by the institutional review board of Toyohashi University of Technology on the use of humans in experiments. Informed consent was obtained from all observers and all methods were performed in accordance with the approved guidelines and regulations of the review board.

4.2.2 Apparatus

Stimuli were displayed on a calibrated 32-inch LCD (Display++, Cambridge Research Systems) with 1920 × 1080 pixel resolution and a 60-Hz refresh rate. Stimulus presentation was controlled by MATLAB using Psychtoolbox 3.0 (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). While they observed the stimuli, each observer was seated on a chair facing the display in a dark booth, with their head secured on a chin rest to maintain a constant distance (57 cm).

4.2.3 Stimuli

Modelling

The stimuli were modelled using Blender 2.77. First, the object was created as a UV sphere with 10 segments and 10 rings. Then, the subdivide function of the mesh tool was used with the number of cuts set to one, fractal set to 15, and along the normal set to one as parameters for subdividing shape 1. For shape 2, these parameters were 1, 10, and 1, respectively; and for shape 3, 1, 5, and 1, respectively. Then, a subdivision surface was added to these objects, with two views and two renderers using the modifier tool. Finally, the object was set as having a smooth surface using the shading smooth function. The specific modelling procedures were as described above and the other procedures retained the default parameters.

Rendering

The reflectance and transmittance of smooth-surface materials are theoretically defined by the refractive index of the material (Hecht, 1998). For example, the refractive index of common glass is 1.5 and its reflectance and transmittance are 0.04 and 0.96, respectively. Therefore, most of the incoming light passes through the object and we perceive the object as being made of glass. As the refractive index increases, so does its reflectance (see Fig. C.2). Moreover, when the refractive index approaches infinity, all components of light are reflected. Some metals, such as silver and aluminium, have a complex refractive index and their reflectance exceeds 0.9. Therefore, the light is almost perfectly reflected by the material surface and humans perceive it as a polished metal such as a mirror. All objects were rendered using the Mitsuba renderer (Jakob, 2010). For the natural illuminations, we used two light fields, called 'Uffizi Gallery' and 'Dining room of the Ennis-Brown House', from the High-Resolution Light Probe Image Gallery (Debevec et al., 2000), and three light fields, called 'Whiteley shopping centre', 'Bolderwood forest', and 'Piazza café', from the Southampton-York Natural Scenes (SYNS) dataset (Adams et al., 2016). For the binary noise illumination, we set a random binary noise image as an environment illumination. The object was set on the default position in the scene. The camera was located at five units from the object. For the mirror material properties, the bidirectional scattering distribution function (BSDF) parameter was a conductor and the surface was set to 100% specular reflection. For the glass material properties, the BSDF parameter was a dielectric, the internal refraction index was 1.5 (as with common glass), and the external refraction index was 1.0 (as in air).

With the camera fixed, the object was rotated about the vertical axis by 0.5° per frame. We rendered 60 frames per stimulus. The sampling count was 512 per pixel with a low-discrepancy sampler. The reconstruction filter was set as Gaussian and each output frame was a 512 × 512 image. Finally, all images were resized to 256 × 256, and a 1000-ms video containing 60 frames was created. Stimuli for the 'original' condition were created using the above procedures. Then, stimuli for the 'upside-down' condition were finally obtained via an operation that rotated the image by 180°.

Image morphing for rating experiment

When light is input from a medium with refractive index n_i to another medium with refractive index n_t, the reflectance R and transmittance T of the output light are expressed as follows:

$$R = \left\{ \frac{n_t - n_i}{n_t + n_i} \right\}^2 \tag{4.1}$$

$$T = \frac{4 n_t n_i}{(n_t + n_i)^2} \tag{4.2}$$

For example, when the light is input from air, which has a refractive index of one ($n_i = 1$), to a medium with refractive index 1.5 ($n_t = 1.5$), such as common glass, R and T are 0.04 and 0.96, respectively. A target image that has an arbitrary refractive index was made using (a) a perfect reflectance image, (b) a perfect transmittance image, (c) the target reflectance, and (d) the target transmittance. The perfect reflectance image was simply an image of the mirror material. The perfect transmittance image was created using the following steps: 1) the image of the mirror was multiplied by 0.04 and the resultant image was outputted; 2) the image of the glass material was subtracted from the image in 1) and the output was divided by 0.96. Finally, the sum of the multiplication products of (a) and (b) and (c) and (d) was used as the target image. This image morphing was performed for the hi-dynamic range components and the image was processed using gamma curve correction (gamma = 2.2). See also Figure C.4.

4.2.4 Procedure and Task

Experiment 1

At the beginning of the experiment, a grey background (33.7 cd/m^2) and fixation point were displayed, which remained throughout the experiment. After a 1-min adaptation period, the first trial was started by pressing a key. A target stimulus was randomly presented for 1000 ms at a 60-Hz refresh rate. An image was presented as the static stimulus for the static condition. The image was one frame selected once from four frames (frame numbers 1, 16, 31, and 46) extracted from the video (60 frames). The performance for the static condition was the average result for those four frames. A video was presented as the dynamic stimulus for the dynamic condition. To cancel any effect of rotation direction, we prepared trials in which the object rotated leftward and rightward about the vertical axis; these rotations were each included twice in the stimulus conditions. The performance of the dynamic condition was the average result for those four conditions (two trials × two directions). The observer responded by pressing a key on a numerical keyboard to indicate mirror or glass material. The experiment was composed of 480 trials (two materials × three shapes × five illuminations × two rotations × eight present conditions), and all trials were randomly ordered. As the control, we prepared the shuffle condition, which had four frames as described above, but it randomly presented four frames (frame numbers 1, 16, 31, and 46) in 15 frames as a cut-off animation. For this condition, the observers could obtain image information from various viewpoints, similar to the dynamic condition, but virtually no dynamic information (such as rotational motion) was provided (see supplementary experiment).

Experiment 2

The procedure was the same as in experiment 1, except the yes/no task was replaced with the seven-point rating task. All stimuli were presented under the dynamic conditions. The experiment was composed of 270 trials (two trials × nine materials × three shapes × five illuminations), and all trials were randomly ordered.

Experiment 3

The procedure was the same as in experiment 2. The experiment was composed of 60 trials (four trials × five conditions × three shapes), and all trials were randomly ordered.

Experiment 4

The procedure was the same as in experiment 1, except the upside-down condition was excluded, and the binary noise condition was included. The experiment was composed of 96 trials (two materials × three shapes × two illuminations × eight present conditions), and all trials were randomly ordered.

4.2.5 Video Analysis

We estimated the optic flows from each pair of frames of the video stimuli using the Lucas-Kanade method (Lucas and Kanade, 1981) in the MATLAB Computer Vision Toolbox. For preprocessing, the stimuli were sharpened using image sharpening processing. The optic flows in the vicinity along the object contour were excluded, because the effect of the optic flows along an object contour inhibits true temporal deformation information on the object surface.

4.2.6 Quantification for Dynamic Cue (Model Output)

The dynamic cue was defined as the ratio between the positive and negative value motion components along the object rotation direction. In this case, all stimuli were horizontally rotated, and the object rotation direction was the x direction. The ratio k_t between the t th frame and the $t+1$ th frame was defined as described below. The horizontal components of the optic flows $v_x(i,j,t)$ between two frames (t and $t+1$) of the material were used to calculate k. The step function f was defined with parameter a, which is related to the threshold of the step function. We set $a = 0.001$. The stimulus had 60 frames and the optic flows were produced as 59 frames ($T = 59$). Finally, the dynamic cue of one stimulus, in other words, the model output k, was defined as the average of the ratios of all frames.

$$k = \frac{1}{T}\sum_{t=1}^{T} k_t \tag{4.3}$$

$$k_t = \frac{\sum_{i,j} f\left(-v_x(i,j,t);a\right)}{\sum_{i,j} f(v_x(i,j,t);a) + \sum_{i,f} f\left(-v_x(i,j,t);a\right)} \tag{4.4}$$

$$f(x;a) = \begin{cases} 1(\text{ if } x > a) \\ 0(\text{ if } x \leq a) \end{cases} \tag{4.5}$$

4.3 Results

4.3.1 Perceptual Discrimination between Mirror and Glass

To test whether the visual system uses distinctive motion information to distinguish mirror and glass, we compared the performance of two presenting conditions: static and dynamic. Under the static condition, the observer was presented with a single image frame, randomly extracted from the video. The dynamic condition was simply presented in video form. These conditions were defined as ‘original’, and an additional condition was defined as ‘upside-down’ stimuli, in which images were rotated by 180°; this operation was performed in order to measure the amount of additional information provided by static cues (Kim and Marlow, 2016). Both original and upside-down stimuli were intermingled in one block. One stimulus (60 frames) was displayed on a liquid crystal display (LCD) monitor for 1000 ms (i.e. the frame rate was 60 frame/s), and observers were asked to state their opinion on which material (mirror or glass) was being observed in a yes/no task paradigm. The observer performance was defined as the average of the percentage of correct answers for the mirror and glass materials (see Procedure & Task).

Figure 4.2A shows the performance for each condition. We performed two-way repeated-measures analysis of variance (ANOVA) for the presenting condition and the rotating condition. The main effect of the presenting condition was significant ($F(1,\ 9) = 18.311$, $p < 0.005$) and indicated that the performance under the dynamic condition was higher than that under the static conditions. The main effect of the rotating condition was significant ($F(1,\ 9) = 10.018$, $p < 0.01$), meaning that turning the image upside-down decreased the performance of perceptual material discrimination. There was no significant interaction ($F(1,\ 9) = 0.048$, $p = 0.832$).

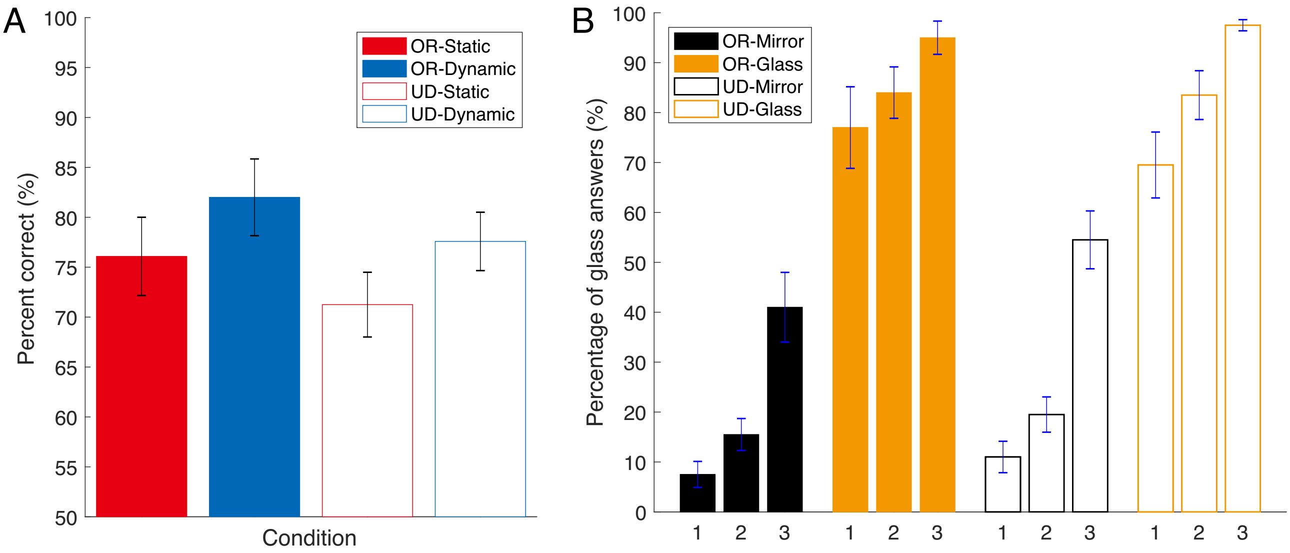


Figure 4.2: Perceptual material discrimination between mirror and glass

(A) Results for perceptual material discrimination under two presenting conditions, along with the rotating conditions. The horizontal axis indicates each condition combined with the rotating and presenting conditions. The vertical axis indicates the percentage of correct answers. 'OR' signifies original, and 'UD' signifies upside-down stimuli. Averages and standard errors among observers were obtained. The error bars represent the standard error of the mean across all ten observers. (B) Percentage of glass answers for different shapes. The horizontal axis indicates the shape number and the vertical axis indicates the percentage of glass answers. Different numbers along the horizontal axis indicate different shapes. The symbols are the same as in A.

The dynamic condition exhibited higher performance, in support of our hypothesis. Under this condition, the visual system acquired consecutive images as rich three-dimensional structural information from the video, as in Ullman (1979), for example. The three-dimensional information gives a cue to albedo estimation on the surface (Marlow and Anderson, 2015; Marlow et al., 2015). This result suggests that there is a cue for not only albedo estimation, but also perceptual material discrimination between mirror and glass materials, even those with complex optical features on their surface.

The performance for the upside-down stimuli was significantly lower than that for the original stimuli. One reason for this is that turning an image upside-down decreases a cue derived from 'the light-from-above prior', i.e. the assumption that light simply comes from above one's head (Mamassian and Goutcher, 2001; Ramachandran, 1988; Sun and Perona, 1998; Adams et al., 2004). These results suggest that both the dynamic cue derived from the video stimuli containing consecutive images and the static cue derived from the image itself contribute to perceptual material discrimination.

Then, to determine the motion information, we confirmed a tendency associated with the percentage of glass answers in each shape and the rotating conditions under the dynamic condition (Figure

4.2B). Specifically, the percentage of glass answers increases depending on the bumpiness of the shape, regardless of the material. This suggests that the more spherical an object is, the more easily it is perceived as being made of glass material, as shape 1 in Figure 4.1 has the bumpiest surface, whereas shape 3 has the least bumpy surface.

We also included a 'shuffle' condition to examine whether the increasing performance was not simply caused by the number of viewpoints. Although the shuffle condition provided image information from various viewpoints, similar to the dynamic condition, its performance was almost same as the static condition. Simultaneously, we tested the effect of changing the luminance polarity to measure the amount of information provided by static cues derived from the natural environment, similar to the upside-down condition. This operation had the same tendency as that of turning upside-down (see supplementary experiment).

4.3.2 Model Development and Performance

Our findings (Figure 4.1B, C) suggest that it is possible for the simple quantified index to express perceptual material discrimination between the mirror and glass materials. We defined a quantified index k as the motion ratio between the direction of object rotation and its opposite direction based on the relationship between motion coherency and behavioural performance (Nowlan and Sejnowski, 1995) (see Quantification for the dynamic cue). Figure 4.3 illustrates the model. We assumed that the visual system detects two kinds of motion: (1) motion in the object rotation direction and (2) motion in its opposite direction, and we estimated the ratio of the opposite motion in all motions.

It is possible to develop a model that can explain not only two specific materials (mirror and glass), but also a general material with a smooth surface depending on the refractive index. To achieve this goal, we collected the rating data for specular reflection and transparency as perceived by the human observers. We assumed a general material with a smooth surface and material property changing with the refractive index. Then, we created images of an object composed of a material with an arbitrary refractive index via image morphing, using the mirror and glass materials (see Stimuli). Finally, we prepared nine different materials as stimuli with nine different refractive indexes (1.33, 1.5, 1.7, 2.42, 5, 10, 20, 50, and infinite). The stimuli were then presented under the dynamic condition (i.e. as video) and the human observers rated the materials on a seven-point scale from perceived specular reflection to perceived transparency. The other conditions were the same as in the experiment for perceptual material discrimination.

Figure 4.4A shows the rating score obtained for each condition. We performed two-way repeated-measures ANOVA for the refractive index and shapes. The main effect of the refractive indexes was significant ($F(1.665, 14.989) = 58.57$, $p < 0.001$). The material with the same refractive index (infinite) as the mirror material was rated significantly lower than others; in other words, it was perceived as having the most specular reflection among all the different refractive index materials (multiple comparison test). The main effect of the shapes was significant ($F(1.263, 11.366) = 50.562$, $p < 0.001$). The ratings for the shapes were significantly different (multiple comparison test; shape 1

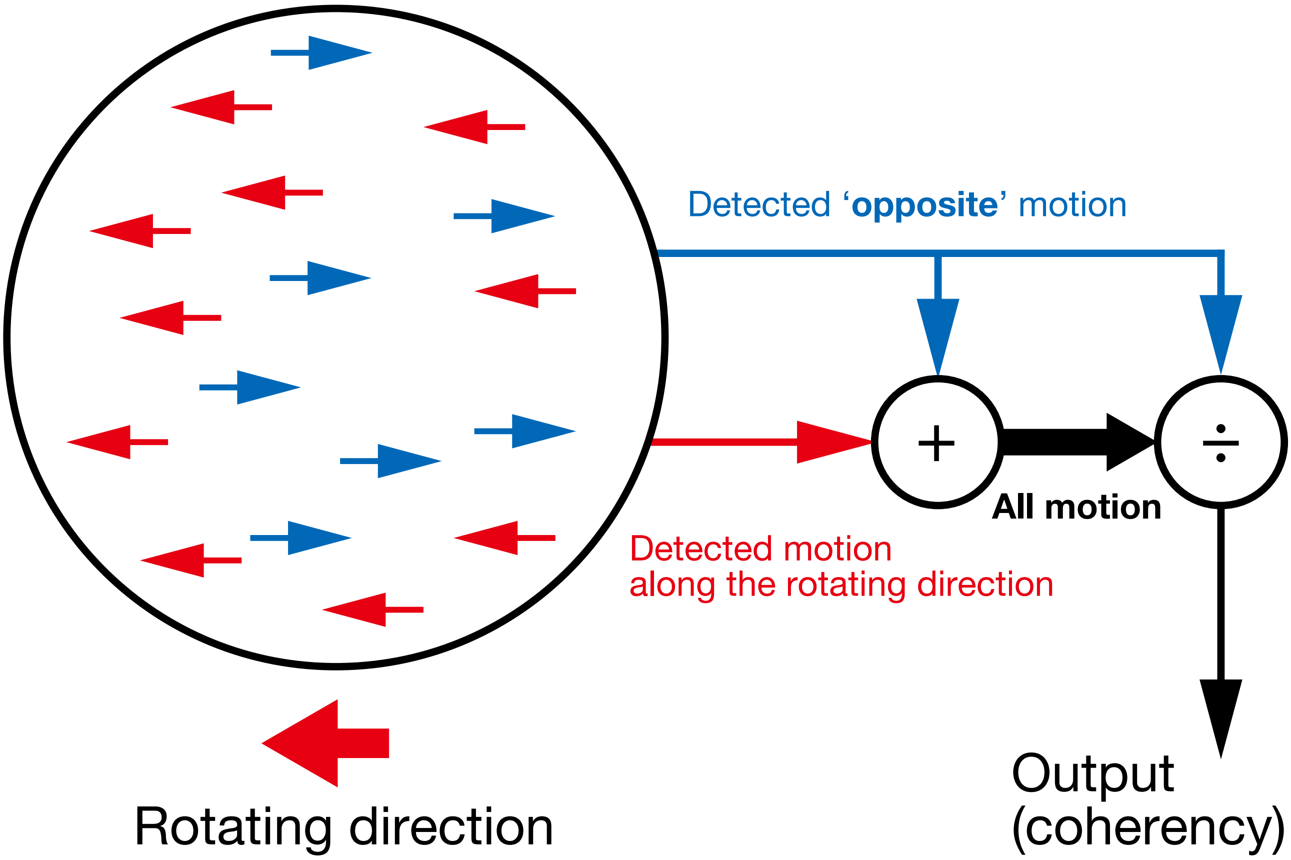


Figure 4.3: Model description

vs. shape 2 ($p < 0.01$); shape 1 vs. shape 3 ($p < 0.001$); shape 2 vs. shape 3 $p < 0.001$) and shape 3 obtained the highest rating; in other words, it was perceived as being the most transparent object. There was no significant interaction ($F(16, 144) = 1.100$, $p = 0.361$). Note that we described adjusted degrees-of-freedom using Greenhouse-Geisser correction when the assumption for sphericity was not in effect (Mauchly's test of sphericity).

The ratings changed depending on the materials, which physically changed in accordance with the refractive index. This means that the visual system can appropriately rate stimuli simulated by image morphing. In addition, there was a tendency to perceive less bumpy objects as being more transparent. Thus, the perception of transparency changes for three-dimensional shapes, much like the perception of glossiness 36. The difference in rating score between shape 1 (most bumpy) and shape 3 (least bumpy) was 1.48 on average. This suggests that the object shape easily changes human perception of a material even if it has the same refractive index.

Figure 4.4B shows the relationship between the refractive indexes and the output score k of the model. As in Figure 4.4A, it expresses a nonlinear function that indicates the relationship between the refractive indexes and the rating given by the human observers. Further, the model output well explains the rating via the refractive index shown in Figure 4.4C. Note that these results are significantly correlated ($r = 0.83$, $p < 0.001$). We also performed a two-way ANOVA without replication of the model output. The main effect of the refractive indexes was significant ($F(8, 16) = 175.15$, $p < 0.001$). The main effect of the shapes was also significant ($F(2, 16) = 1854.03$, $p < 0.001$). These statistical features of the model output are consistent with the statistical features of the ratings given by the human observers. Therefore, this model offers an explanation, i.e. that the visual system uses dynamic information for material perception of a rotating object. Note that the correlation between these results is not perfect, because some factors that are not affected by either the refractive indexes or the shapes remain.

Moreover, the model can also explain the results of perceptual material discrimination. Figure 4.4D shows the relationship between the model output, which is used instead of the shape conditions, and the percentage of glass answers. This indicates the extent to which the model estimates the observer performance for perceptual material discrimination. If the dynamic cue contributes to this material perception, the model output and the percentage of glass answers should be connected in a correlative relationship. Note that our premise is that the relationship between the rating dimensions from the perceived specular reflection to transparency and the percentage of glass answers obtained in a yes/no task paradigm is a monotonically increasing function. In the original case, although the percentage of glass answers increased with the model output, there is a separation between the mirror and glass materials (in Figure 4.4D, the filled symbols). This suggests that the dynamic information and other factors contribute to the material perception. The question now becomes the following: which static cues contribute to the distinction between mirror and glass materials? Considering that these cues exist among the original stimuli but not among the upside-down stimuli, one possibility is the luminance distribution along the vertical direction of the stimuli, because the effect derives from 'the light-from-above prior' (Mamassian and Goutcher, 2001; Ramachandran, 1988; Sun and Perona, 1998; Adams et al., 2004). These results suggest that the visual system naturally hypothesises that the light is located above the observer's head. We thus speculate that the other factor is a static cue from the image.

In contrast, when we rotated the image by 180°, the model output and percentage of glass answers were significantly and more highly correlated ($r = 0.88$, $p < 0.05$) (in Figure 4.4D, open symbols). This correlation coefficient was greater than that obtained with the original stimuli. Although both the original and upside-down stimuli had the static and dynamic cues, the visual system could not sufficiently use the static cue in the upside-down stimuli because the images were rotated, which inhibits the static cue. Therefore, the visual system was basically more dependent on the dynamic cues.

In addition, turning upside-down seems to more strongly affect to our perception when the k value was medium (in this case, at around 0.3). To quantify the effect of turning upside-down, we focused on the change of material appearance (mirror-to-glass or glass-to-mirror) by turning upside-down to manipulate the static cue and calculated the ratio of the appearance change for various k values. Figure 4.4E shows the appearance change ratio significantly varied depending on the k value ($F(5, 45) = 10.281$, $p < 0.001$; one-way repeated-measures ANOVA). A multiple comparison test showed that the change rate of the mirror of shape 3 (black circle) was significantly higher than that of the mirrors of shapes 1 and 2, and the glass of shape 3 ($p < 0.05$). These results suggest that the effect of the dynamic cue is limited, i.e. the visual system relies more on the static cue when the dynamic cue is ambiguous for distinguishing mirror and glass.

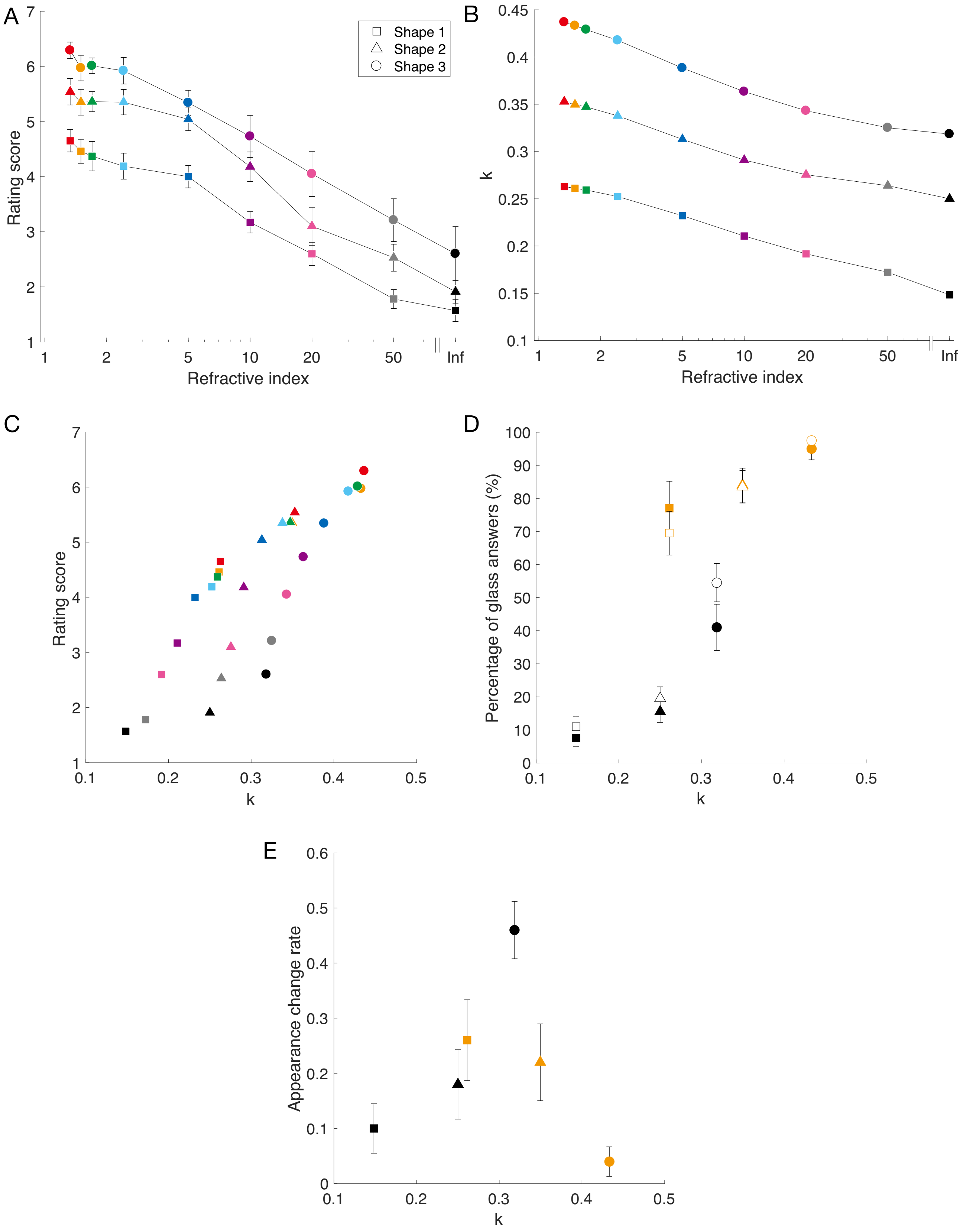


Figure 4.4: Developed model and its performance

(A) Rating score of the perceived specular reflection and transparency. The horizontal axis indicates the refractive index of the stimuli on a log scale. The vertical axis indicates the rating score on a seven-point rating system, where perceived specular reflection is one and transparency is seven. Averages and standard errors among observers were obtained. The error bars represent the standard error of the mean across all ten observers. (B) Model output. The horizontal axis is the same as in A. The vertical axis indicates the model output k. (C) Correlation between model output and ratings given by human observers. (D) Correlation between the model output k and the percentage of glass answers for original and upside-down stimuli under dynamic conditions. The horizontal axis indicates the model output k.The vertical axis indicates the percentage of glass answers; this axis is the same as in Figure 4.2B. The filled and open symbols indicate the original and upside-down stimuli, respectively. Averages and standard errors among observers were obtained. The error bars represent the standard error of the mean across all ten observers.(E) The appearance change rate. The horizontal axis is the same as in C and D. The vertical axis shows the rate indicating how much the material appearance changes. In B-E, the symbols are the same as in A.

4.3.3 Model validation

Our model predicts that glass objects may be misperceived as mirror surfaces if the glass generates little or no motion transparency. To test this prediction, we attempted to reduce the amount of motion transparency that glass objects generate by rendering the glass as a purely transmittance component that only refracts light and does not generate any specular reflection.

We generated three stimuli: 1) glass only, possessing the transmittance component of the light ('transmittance only', Movie 5.3A) and created through image morphing (see Stimuli); images created by superimposing two different shape mirrors images in the 2) same ('superimposed', Movie 5.3B) and 3) opposite directions ('oppositely superimposed', Movie 5.3C). The last two also changed the amount of motion transparency. Figure 4.5A shows the rating scores for five conditions including actual mirror and glass. We performed one-way repeated-measures ANOVA for the stimulus conditions. The main effect for the stimulus conditions was significant ($F(1.962, 17.656) = 68.842$, $p < 0.001$, adjusted degree-of-freedom using Greenhouse-Geisser correction). We predicted that the 'transmittance only' condition would be perceived as being more transparent than the glass, because the reflectance component of the former is eliminated, and its k value increases according to the model. However, there was no significant difference (multiple comparison test; transmittance only vs. glass ($p > 0.05$)), suggesting that the transmittance component mainly provides us with transparency perception. The superimposed and oppositely superimposed conditions were perceived as being more transparent than the mirror (multiple comparison test; mirror vs. superimposed ($p < 0.001$); mirror vs. oppositely superimposed ($p < 0.001$)). We speculate that the superimposition of two different shapes provides humans with an impression of a transparent layer on the surface. The oppositely superimposed condition additionally increased the opposite motion relative to the rotation direction (the model output k, see Figure 4.5B). This was not the same rating as the glass, because the pixel information derived from the mirror reduced the perceived transparency. If we could naturally match two different shapes on a pixel-by-pixel basis, a clearer glass appearance would be obtained. Then, the relationship between the k value and the rating score would more generally support the model.

Even if the rotating stimuli in which static cues were maximally eliminated, motion transparency would still provide sufficient cues to distinguish the materials. To test this, we rendered new stimuli (Figure 4.5C, see also Movie 5.4A and Movie 5.4B) in a random binary noise illumination using more complex shaped objects based on our proposed model. We tested them using a material judgement test. Figure 4.5D shows that the dynamic condition was significantly higher than the static condition ($t(12) = 3.399$, $p < 0.01$, t-test), supporting our findings. This result suggests that the dynamic cue is informative for the distinction of materials, even without a simple static cue. Moreover, we found a tendency that the luminance distribution in the vertical axis of the binary noise condition was almost flat and the absolute difference between mirror and glass was smaller than that for the natural illumination (Figure 4.5E). In the real world, we naturally exhibit a luminance bias, i.e. the upper area of the object surface is brighter, and the bottom area of the object surface is darker because the lights from sources such as the sun and illumination lamps tend to come from above. However, in the binary noise condition, this bias does not exist, and the visual system cannot rely on this cue. We speculate that measuring this kind of difference provides us with information on how the visual system is sensitive to the static cue.

4.4 Discussion

Previous studies on surface properties focussed on highly reflective surfaces and included dynamic information (Doerschner et al., 2011; Dovencioglu et al., 2017; Doerschner et al., 2013; Dovencioglu et al., 2015; Sakano and Ando, 2010; Tani et al., 2013; Marlow and Anderson, 2016; Wendt et al., 2010). It has also been suggested that three kinds of cues from optic flows can distinguish the surface property (matte or shiny) of an object (Doerschner et al., 2011). Both the mirror and glass materials in our study were shiny objects and had similar features to identify the shiny surface. In addition, our results reveal that the optic flows can identify whether the material of a rotating object is mirror, glass, or its medial materials depending on the refractive index. We suggest that this dynamic cue from the video can somehow be detected by the visual system, for example, through perception of motion transparency, and be used for perception of smooth surface materials and distinction between mirror and glass.

In this study, the visual system could distinguish between materials under the static condition to an extent, because it can use a distorted image reflected from surrounding illumination on the surface of the mirror object (Fleming et al., 2004), and a refraction image derived from the shape of thick transparent objects such as glass (Fleming et al., 2011). A recent study reported that a specular object and a transparent object perceptually change each other by vertically reversing, and suggested that the direction of illumination affects this phenomenon (Kim and Marlow, 2016). In 'the light-from-above prior', for the mirror object, the upper area of the object surface becomes brighter because the light reflects off the object. Likewise, for a glass object, the bottom area of the object

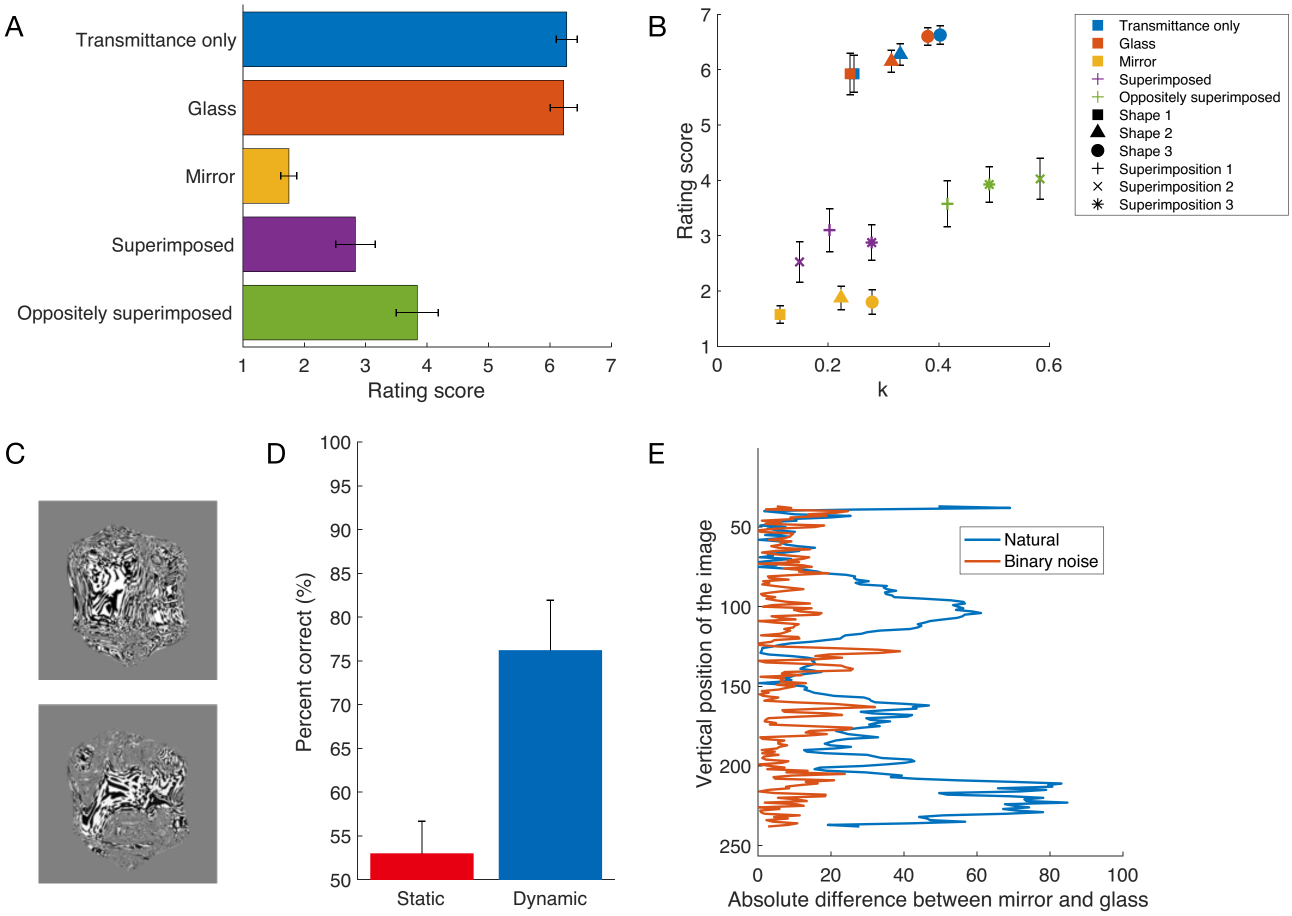


Figure 4.5: Validation of the proposed model with new stimuli

(A) Rating scores of perceived specular reflection and transparency for new stimuli. The horizontal axis indicates the rating score corresponding to the vertical axis of Figure 4.4A. The vertical axis indicates the stimulus conditions. The error bars represent the standard error of the mean across all ten observers. (B) Relationship between model output and rating score. (C) Stimuli rendered in random binary noise illumination (top: mirror, bottom: glass). See also, Movie 5.4A and Movie 5.4B. (D) Result of perceptual material discrimination using the binary noise condition. The horizontal axis indicates each condition. The vertical axis indicates the percentage of correct answers. Averages and standard errors among observers were obtained. The error bars represent the standard error of the mean across all seven observers. (E) Different luminance distributions between original and the binary noise conditions. The horizontal axis indicates the absolute differences of the image pixel intensities (luminance) between mirror and glass. The pixel intensities were averaged along each row in the image, excluding the background. The vertical axis indicates the vertical position of the image.

surface tends to become brighter because the light is transmitted through the object. Our results suggest that the visual system uses such a simple static cue for perceptual material discrimination between mirror and glass materials. Further, image statistics could be given as the other static cue in this case, because the visual system can estimate a surface property (glossiness) on the object using those statistics (Motoyoshi et al., 2007) (but subsequent works have questioned the idea that those statistics play a causal role in material perception (see Anderson and Kim, 2009; Kim and Anderson, 2010). The stimuli for our study were comprised of mirror and glass materials, where both had a specular highlight component and their surfaces were glossy. Therefore, we speculated that a cue from the image statistics was not the only contribution to the material distinction, and that other luminance information, such as the luminance distribution along the vertical direction of the stimuli, also contributed to this distinction.

An image on the surface of mirror object is more clearly visible than that on glass because the surrounding illumination is directly reflected by the surface. By contrast, an image on the surface of a glass object has complex patterns due to the distortion field (Fleming et al., 2011). We premised that these midlevel features distinguish the reflected image from the surrounding illumination or the refracted image through the background. The static cue surely has an influence on overwriting the dynamic cue. We suggest that the visual system automatically shifts its focus to use only one cue or multiple cues in various contexts. We found that the magnitude of the object's bumpiness provides an effect to aid perception of specular and transparent materials (Fig. 4.2B, Fig. 4.4A). A smooth surface tends to be perceived as glass, because it is easy to track the motion on the object's rear surface. In contrast, a bumpy surface tends to be perceived as specular, because there are enough feature points to track the motion on the object's front surface. Some material appearances change largely depending on bumpiness (e.g. specularity (Ho et al., 2008; Anderson et al., 2014) and translucency (Chowdhury et al., 2017)). Thus, we speculate that the ability to track the object motion aids distinction between mirror and glass.

Some observers reported perceiving liquid in the rotating glass material object, especially in shape 3 (lowest bumpiness). Almost all liquids are transparent or translucent, and recent studies have reported the importance of liquid perception for the visual system (Kawabe et al., 2015b,a; Kawabe and Kogovšek, 2017; van Assen and Fleming, 2016; Paulun et al., 2015). In our study, although the glass material object was definitely rigid and solid, the visual system perceived a non-rigid object like water, because the visual system distinguishes non-rigid object shapes from the motion cue (Jain and Zaidi, 2011). This study was mainly aimed at rigid objects only; however, it is necessary to discover how the visual system distinguishes between non-rigid specular and transparent objects such as mercury and water.

From neuroscience and neurophysiological perspectives, our glossiness perception (Nishio et al., 2014; Goda et al., 2014)4 and material perception (Hiramatsu et al., 2011; Goda et al., 2014) are represented in the inferior temporal (IT) cortex through the ventral pathway. The structures determined from motion (Bradley et al., 1998; Grunewald et al., 2002) and motion transparency are represented in the middle temporal area (MT) (Qian and Andersen, 1994; Snowden et al., 1991) and the fundus of the superior temporal sulcus (FST) (Rosenberg et al., 2008) through the dorsal pathway. It is possible that both processes are in parallel or provide mutual feedback to each other, raising the recognition level. We suggest that examination of the relationship between them will provide further insight into the perception of smooth surface materials by dynamic cues.

In-depth understanding of the material perception performed by the visual system can be obtained by considering not only the static cue from the image, but also the dynamic cue, because the findings of this study provide one piece of evidence to explain this mechanism. Our finding can also explain perception of any motions other than rotation. In future work, we would like to study not only smooth surface objects such as mirror and glass materials, but also objects with matte surfaces such as wood, fabric, and other materials, or a mixture of materials.

Chapter 5

The glare illusion

A similar version of this chapter has been published as:

Tamura, H., Nakauchi, S., & Koida, K. (2016). Robust brightness enhancement across a luminance range of the glare illusion, *Journal of Vision*, 16(1), 10, 1-13. DOI:10.1167/16.1.10.

The glare illusion refers to brightness enhancement and the perception of a self-luminous appearance that occurs when a central region is surrounded by a luminance gradient. The center region appears to be a light source, with its light dispersing into the surrounding region. If the luminous edge is critical for generating the illusion, modulating the perceived luminance of the image, and switching its appearance from luminous to nonluminous, would have a strong impact on lightness and brightness estimation. Here we quantified the illusion in two ways, by assessing brightness enhancement and examining whether the center region appeared luminous. Thus, we could determine whether the two effects occurred jointly or independently. We examined a wide luminance range of center regions, from 0 to 200% relative to background. Brightness enhancement in the illusion was observed for a wide range of luminances (20-200% relative to background), while a luminous-white appearance was observed when the center region luminance was 145% of the background. These results exclude the possibility that brightness enhancement occurs because the stimuli appear self-luminous. We suggest that restoring the original image intensity precedes the perceptual process of lightness estimation.

5.1 Introduction

Lightness estimation of objects within scenes is an important function of our visual system. The perception of lightness is related to the proportion of reflected light from a surface (Kingdom, 1997). This lightness percept can be categorized, such as by the descriptive terms white, gray, or black. Since the Middle ages, numerous reports, such as those of Alhazen, and later by Helmholz and Hering (Kingdom, 1997), have described lightness perception and its basis. Current vision researchers have built upon such works (Gilchrist, 2007). To estimate the lightness of an object's surface, several cues are available, such as its luminance, spatially adjacent stimuli, three dimensional configuration, and texture statistics (Adelson, 1993; Anderson and Winawer, 2005; Gilchrist, 2007; Knill and Kersten, 1991). Based on these cues, the visual system could estimate the mapping between the retinal output and reflectance, which is required for lightness perception (Adelson, 2000; Gilchrist, 2007; Murray, 2013). Numerous studies have addressed lightness perception and illumination estimation (Adelson, 2000; Adelson and Pentland, 1996; Arend and Goldstein, 1987; Gilchrist, 2007; Kingdom, 2011; Land and McCann, 1971). One hypothesis is that lightness perception describes how the visual system decomposes images into separate layers based on illumination geometry, shading due to shape, and surface albedo (Gilchrist, 2007). These three components correspond to the physical attributes of scenes, and are indispensable for rendering objects in computer graphics. If two of them, such as illumination and shape, are clearly defined, the other component, in this case, reflectance, can be determined theoretically. However, if the target object is self-luminous, this calculation is not possible. Thus, it is important to know how humans discriminate self-luminous from non-luminous surfaces. A stimulus that is higher in luminance than its surroundings has a luminous appearance (Bonato and Gilchrist, 1994; Evans, 1959; Gilchrist et al., 1999; Radonjić et al., 2011; Uchikawa et al., 2001). The contrast between an object and its surroundings is important for lightness and brightness perception. The physiological basis of this function is lateral inhibition in the retina (Schneck, 2010; Shapley et al., 1984; Walraven et al., 1990) and higher-order cortical mechanisms (Boyaci et al., 2007; Haynes et al., 2004; Kinoshita and Komatsu, 2001; MacEvoy and Paradiso, 2001; Roe and Tso, 1995; Shevell et al., 1992; Whittle, 1994). If there exist circumstances whereby a luminous appearance is not explicable by the contrast model, these would potentially assist our understanding of lightness/brightness perception mechanisms. The glare illusion is an optical illusion in which brightness enhancement and self-luminosity occur for a bright central region relative to the equiluminant surface (Agostini and Galmonte, 2002; Zavagno, 1999). It has a surrounding gradient that mimics the spread of intense light due to atmospheric or ocular dispersion (Kakimoto et al., 2005; Spencer et al., 1995). Figure 5.1A shows examples of the glare illusion. In the top right of the figure, the central circles appear relatively bright and self-luminous due to their annular blur, even though the central region has the same luminance as the circle with no annular blur, directly below. The glare illusion is very robust across various displays, and even occurs in printed images. As early as the Renaissance period, painters such as Tintoretto and Rembrandt employed blur to produce the appearance of self-luminosity (Zavagno and Massironi, 1997). Perceived self-luminosity induced by annular blur has been frequently used as a stimulus in psychological experiments (Correani et al., 2006; Hanada, 2012; Keil, 2007; Yoshida et al., 2008; Zavagno et al., 2004; Zavagno and Caputo, 2001, 2005), including fMRI studies (Leonards et al., 2005). Further, such blur is widely used in computer graphics to represent an intense light source or specular reflection (Nakamae et al., 1990; Rokita, 1993; Shinya et al., 1989; Spencer et al., 1995). This illusion raises questions about how our visual system integrates contextual information when evaluating lightness or brightness, and how self-luminosity is judged.

Measurement of the effect of visual illusions has been considered important in neuroscience (Eagleman, 2001). It is possible to measure illusions consistently in animals as well as humans (Kelley and Kelley, 2014). Thus, quantitative measurement of the illusion is important in order to evaluate similarities in perception across species and to understand the neurophysiological basis of our visual system. Zavagno and Caputo (2001) measured self-luminosity thresholds of the glare illusion by modulating the luminance of the central and inducing regions and found that the illusion occurred even when the central region was darker than the white background. Yoshida et al. (2008) measured the amount of brightness enhancement in the glare illusion. Subjects perceived the central region in the illusion as 20-35% brighter than the reference white. These two studies implicitly assumed that two ways exist to quantify the glare illusion, namely via self-luminosity thresholds and by assessing the amount of brightness enhancement. Further, it was assumed that these two attributes can be measured separately. In this study, we asked how the illusion elicits two kinds of perceptual effect—brightness enhancement and a self-luminous appearance—and whether they are elicited either jointly or independently. To test this, we modulated the stimulus intensity while keeping the background stimulus constant and asked the subjects to report perceived brightness and to provide a categorical judgment of appearance. Does the glare illusion induce a brighter, self-luminous appearance for both dark and bright stimuli, and what luminance level effectively induces the illusion? If the illusion is observed at specific range, do the two aspects of the illusion appear jointly?

5.2 Methods

5.2.1 Subjects

Subjects were eight naïve volunteers aged from 22 to 25 years (average 22.8 years, seven males and one female). They had normal or corrected-to-normal acuity and normal trichromatic color vision. All experimental protocols were approved by the institutional review board of Toyohashi University of Technology with respect to the use of humans in experiments. Informed consent was obtained from all participants.

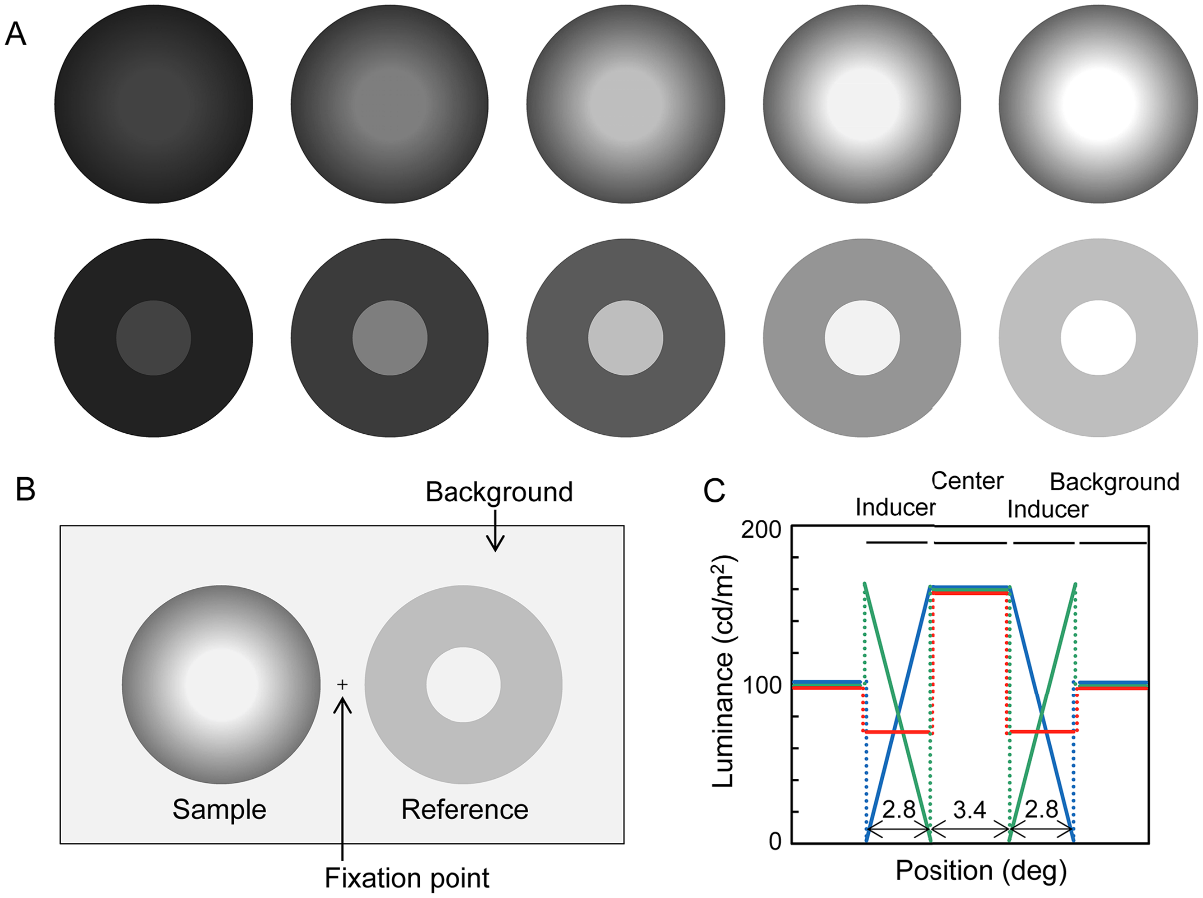


Figure 5.1: Stimuli used in this study

(A) Examples of glare stimuli (top) and reference stimuli (bottom). The luminance of a central circle and surrounding inducer were modulated from dark to light (left to right). Circles in the same column have the same intensity (luminance). (B) Screen appearance: The sample and reference stimuli were placed to the left and right of the central fixation point, respectively. (C) Luminance profiles of Glow (blue), Halo (green), and Uniform (red) stimuli. The annulus is labeled as “inducer,” the inner circle as “patch,” and the surrounding uniform area covering the remaining display as “background.” In this example, the luminance of the center patch was 160 cd/m^2. The background was always 100 cd/m^2 irrespective of sample intensity.

5.2.2 Apparatus

Stimuli were displayed on a 27-inch (screen size 596 × 335 mm) liquid crystal display (LCD; CG276, EIZO) with a resolution of 2560 × 1440 pixels, calibrated by ColorCAL II (CRS). Stimulus presentation was controlled by MATLAB using Psychtoolbox 3.0 (Brainard, 1997) on a desktop computer (Dell Precision T7500). Subjects observed the stimuli binocularly, while seated on a chair facing the display with the head secured on a chin rest to keep a constant distance from the display (1340 mm). The only luminous object present in the room was the LCD monitor.

5.2.3 Stimuli

Figure 5.1B shows the screen layout in both experiments. The stimulus on the left is an example of an image that evokes the glare illusion. That is, the central circle is surrounded by a luminance gradient extending toward the periphery of the annulus. The stimulus on the right is a reference where the center patch is surrounded by an annulus of uniform luminance. The diameters of the outer and inner stimulus boundary were 9 deg and 3.4 deg of visual angle, respectively. Two stimuli were simultaneously displayed 5.5 deg to the left and right of the fixation point. The background was uniform, with a luminance of 100 cd/m^2. Figure 1C illustrates the luminance profiles of the three types of stimuli used in this study. Center patches of the three stimuli were the same, but the surrounding inducers were different. Glow had a linear profile, which modulated from zero on the outer boundary to the highest luminance adjacent to the central region. Halo had a similar profile as Glow but the direction of the gradient was reversed. Halo was used to examine the effect of contrast polarity and to reduce visual adaptation to Glow. Uniform had a spatially uniform profile, with a luminance of 44% of the central region. The value of 44% was arbitrarily determined. We also used a uniform black inducer in an additional experiment for comparison (see D.1). After taking into account the number of pixels, the mean inducer luminance of Glow and Halo compared to the center patch was 41% and 59%, respectively. All stimuli were achromatic gray (CIE: x = 0.290, y = 0.331).

5.2.4 Procedure

Experiment 1: brightness comparison

Two stimuli were presented while participants were fixating; they were asked to report which of the two stimuli had the brighter central region. The two stimuli were presented horizontally to the left and right of the fixation point; one was the sample and the other was the reference. The sample was selected randomly from three profiles (Glow, Halo, and Uniform), and the reference was always Uniform. The horizontal position of the sample was randomly switched. There were 18 luminance levels: 0, 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 140, 160, 180, and 200 cd/m^2. Both the sample and reference in each trial had the same luminance. We also used other reference stimuli with 20% higher (bright reference) or 20% lower (dark reference) luminance. In this case, both center and annulus luminance varied together while the contrast between center and annulus was kept constant. At the beginning of the experiment, the background and a fixation point were displayed and remained throughout the experiment. After three minutes of adaptation, the first trial was started by a mouse click. Stimuli were presented for 0.3 seconds, after which only the background was shown. The subject then responded with a right or left mouse click to indicate which stimulus was brighter. The next trial began following a further click. One session was composed of 648 trials (2 repeats × 2 positions × 18 luminance levels × 3 profiles × 3 references), and all conditions were randomly ordered. Each subject participated in four sessions; thus, each condition was repeated eight times. In the analysis, we combined the data for the two horizontal positions, thus providing 16 repeats per condition.

Experiment 2: categorical labeling

After three minutes of adaptation to the background, a single stimulus was displayed either to the right or to the left of the center. Stimulus position was switched on every trial. Subjects were asked to report the appearance of the central region of the stimulus using four categories: black, gray, white, or luminous-white. The response was recorded by keyboard. There was no fixation point, and the subjects were allowed to freely view the stimuli. We also conducted the additional experiment to confirm that the responses were stable even when eye position was controlled (see D.2). The luminance profiles of the stimuli consisted of Glow, Halo, and Uniform. There were 21 luminance levels, ranging from 0 to 200 cd/m^2. The stimulus appeared after the key press and was displayed until a response was obtained (the mean response time was 1.05 s). The next trial started following a further key press. One session was composed of 504 trials (4 trials × 2 positions × 21 luminance levels × 3 profiles) and all conditions were randomly intermingled. Each subject participated only in one session.

5.3 Results

5.3.1 Results (Experiment 1)

Brightness enhancement was observed across almost the entire range of target luminance (Fig. 5.2A). Significant enhancements of Glow compared to chance (50%) were observed from 20 to 200 cd/m^2 ($p < 0.05$, binomial test), which included both higher and lower luminance values than the background luminance (100 cd/m^2). The probability never reached 100% at any luminance level. This was due to individual differences in the luminance level inducing brightness enhancement (Supplemental Figure 1). The majority of observers (n = 6) showed monotonically increasing functions, while other observers showed a maximum response probability at specific luminance levels (50-100 cd/m^2). Two observers did not show 100% probability even at the most effective luminance level; nevertheless, the response was significantly higher than chance. In the low luminance range, the probabilities were distributed around 50% or slightly below. This significant negative bias was observed in at least two observers. Halo also showed significant brightness enhancement at several luminance levels (60 and 80-200 cd/m^2, $p < 0.05$, binomial test); however, its probabilities were lower than those of Glow. No significant effects were observed for the Uniform control. Similar trends were observed for the different reference conditions (Fig. 5.2B, C). The largest response was observed for the Glow condition, followed by Halo and Uniform. For the bright reference condition (Fig. 5.2C), responses for Glow were still greater than chance, indicating that the magnitude of brightness enhancement for Glow was still higher than the 20% increase in luminance of the reference. Responses for Halo were slightly lower than chance, indicating that the magnitude of brightness enhancement for Halo did not attain the 20% increase in luminance of the reference. By combining the data from the three reference conditions, we obtained psychometric functions for each luminance level (Fig. 5.2D). The response probabilities obtained for the three reference levels were then fit to a maximum-likelihood logistic function. With this function, we could estimate the brightness enhancement quantitatively. The point of subjective equality (PSE) between the Glow and Uniform reference was calculated as the luminance ratio between the mid-point of the function and zero. The observed PSEs were 30%. Similar calculations were performed for each sample luminance (Fig. 5.2E). The average PSE was 42.8% (Fig. 5.2F) when we included data showing good fit (slope significantly different from zero, $p < 0.05$ by Matlab GLMfit). The average PSEs of Halo and Uniform were 8.5% and 2.1%, respectively. The number of samples in the present study was relatively small for calculating psychometric functions (i.e., there were only three reference levels). Therefore, to validate whether the PSEs were estimated appropriately, we performed an additional experiment that used more reference levels (specifically, 7) for one luminance of the target stimulus (120 cd/m^2, see Figure D.2). The observed PSE was 31% of the luminance ratio, i.e., almost identical to that in the main experiment (Fig. 5.2E). Thus, we conclude that the observed PSEs in the main experiment were reliable.

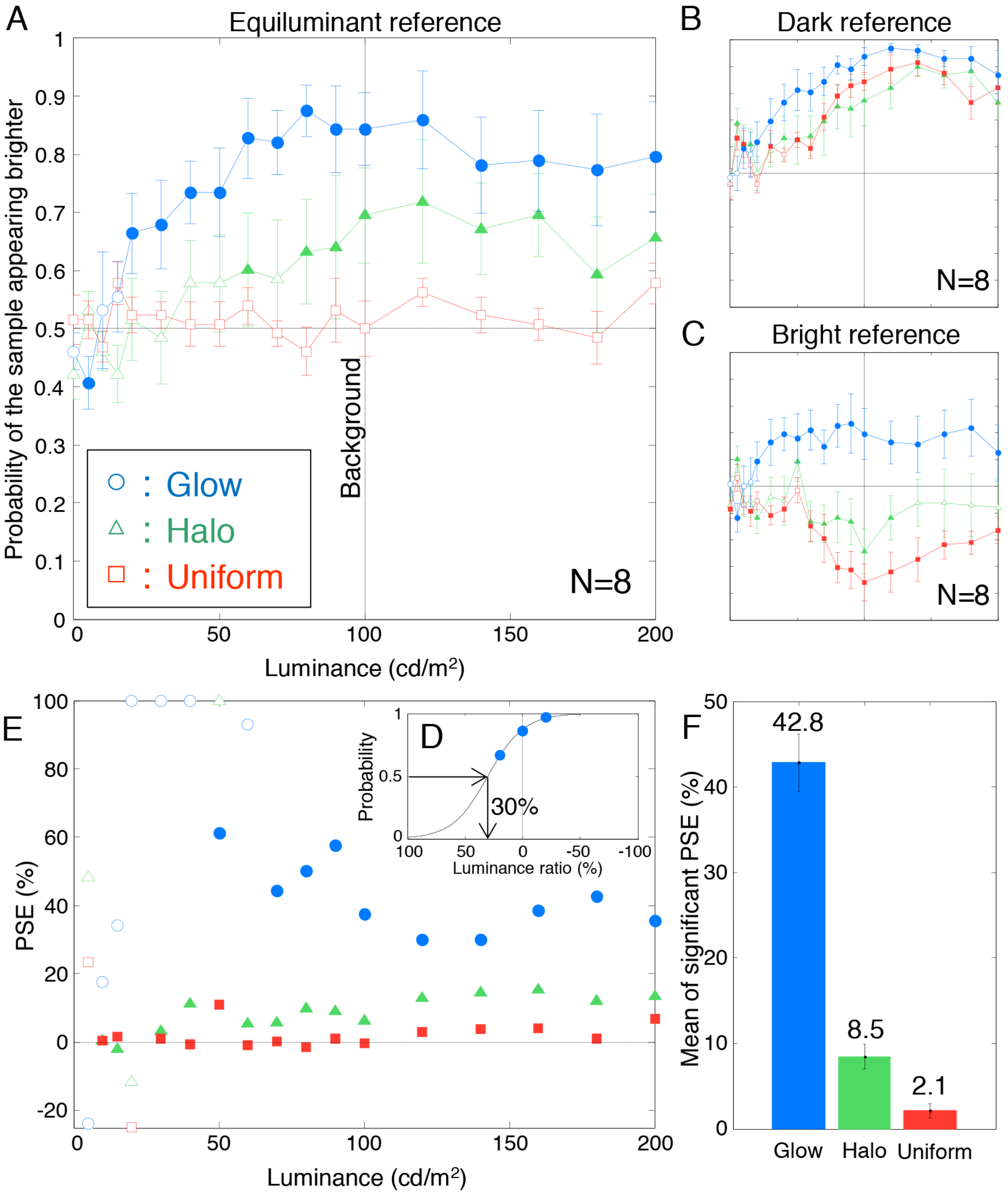


Figure 5.2: Brightness enhancement in the glare illusion

(A) The horizontal axis indicates the luminance of the center patch, and the vertical axis the probability of the sample appearing brighter than the equiluminant reference uniform stimulus. Filled symbols indicate a significant difference from chance (50%). Response probabilities for each subject were calculated, and averages and standard errors among subjects were obtained. Error bars represent the standard error of the mean across all eight subjects. (B) Response probabilities for a dark (-20%) reference condition. (C) Response probabilities for a bright (+20%) reference condition. (D) Example psychometric function at a luminance level of 120 cd/m^2. Response probabilities for three reference conditions with a fitted logistic function are shown. (E) Brightness enhancement quantified by the point of subjective equality (PSE). Symbols are the same as in A. Data were clipped to 100% or -25% if the function returned values greater than 100% or less than -25%, respectively. Filled symbols indicate goodness-of-fit of the psychometric functions. (F) Means and standard errors of significant PSEs for each condition.

5.3.2 Results (Experiment 2)

We found that the Glow stimulus (Fig. 5.3A) was perceived as self-luminous between 100 and 200 cd/m^2, while very few self-luminous percepts were reported for Halo and Uniform stimuli, even at 200 cd/m^2. We calculated the response probabilities of the four alternatives at each luminance level and fit cumulative Gaussian functions via a maximum-likelihood estimation procedure. Using these functions, we determined the categorical thresholds corresponding to the 50% probability level. Note that in Fig. 5.3A, the four responses black, gray, white, and luminous-white are vertically stacked. The threshold between white and luminous-white responses occurred at 145 cd/m^2, a clearly lower luminance than for the other two stimuli. The thresholds for the Halo and Uniform conditions were 248 cd/m^2 and 243 cd/m^2, respectively. Although the illusion had a strong impact on luminance at which the percept transitioned from white to luminous-white, the transition between gray and white percepts was very similar for all conditions (117-135 cd/m^2). Thus, the thresholds between gray and white simply depended on the luminance of the central region; the surrounding annulus had limited influence. The transition between percepts of black and gray for Glow occurred at a slightly higher luminance than for Halo and Uniform stimuli. The majority of subjects (n = 5) showed this tendency. At low luminances, the central region of the Glow stimulus was less visible than in the other conditions due to low edge contrast between the center and annulus.

5.4 Discussion

5.4.1 Comparison

We examined the effect of stimulus luminance on the glare illusion from two aspects: brightness enhancement and the induction of a self-luminous appearance. The results confirmed that the glare illusion enhanced brightness across a wide range of luminances (20-200% relative to the background); the resulting luminance enhancement was 40% (Experiment 1). Self-luminosity emerged when the central stimulus region was 145% of the luminance of the background (Experiment 2). Comparing the results of the two experiments, we found that brightness enhancement arose when the stimulus

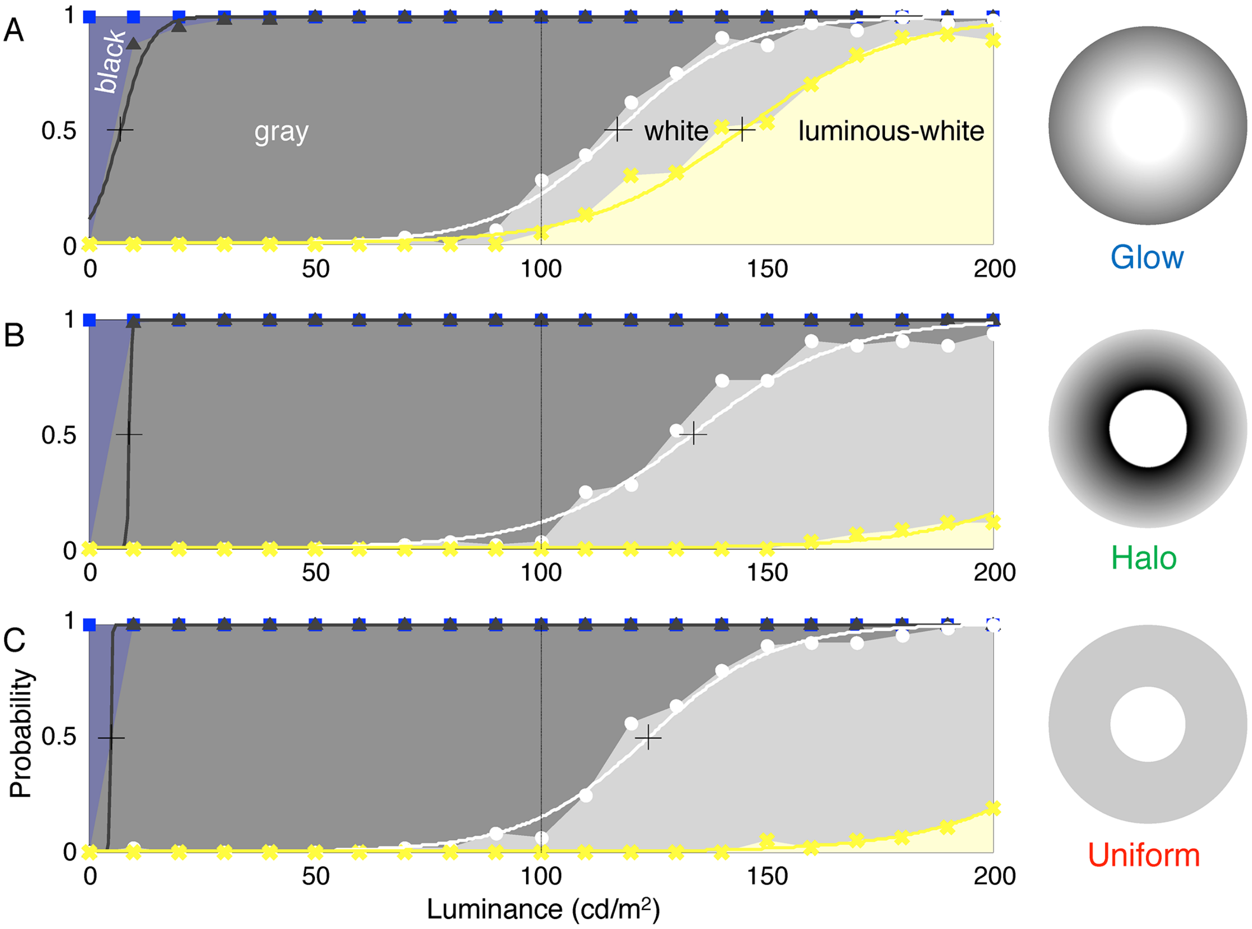


Figure 5.3: Categorical response probability

The horizontal axis indicates the luminance of central circle, and the vertical axis the response probability for each category. A, B, and C show the response for the Glow, Halo, and Uniform conditions, respectively. The categorical responses black, gray, white, and luminous-white increased in probability from low to high luminance. For example, gray, white, and luminous-white responses were observed with approximately equal probability at 120 cd/m^2 for the Glow condition. Functions are cumulative Gaussians, fit via a maximum likelihood method. Crosses indicate 50% probability points of each function.

was not perceived as self-luminous. Indeed, the brightness enhancement of the glare illusion was observed even when the stimulus was categorically perceived as gray. This excludes the possibility that brightness enhancement occurs because the stimulus appears self-luminous. Zavagno and Caputo reported subjective experiences of self-luminous grays in the glare illusion (Zavagno and Caputo, 2001). However, they did not measure these subjective judgments directly. A shift in the categorical threshold due to the illusion was observed only for the white/luminous-white threshold and only a minor shift was observed for the gray/white threshold. In the luminance range of the gray/white threshold (120-130 cd/m^2), glare stimuli enhanced brightness significantly. The white/luminous-white threshold varied by 68% (145-243 cd/m^2) from Uniform to Glow; however, the quantitative effect of brightness enhancement was only 40%. These results suggest that the luminous appearance did not quantitatively match the metric evaluation of brightness. These discrepancies between brightness estimation and categorical reports might originate from independent underlying mechanisms (Witzel and Gegenfurtner, 2013). There was a small but significant decrease in brightness estimates for low-luminance Glow stimuli (Experiment 1), and an increased probability of black responses for these stimuli (Experiment 2). These two occurrences might share the same origin, namely invisibility of the stimulus due to low contrast at the edge of the central region, which might cause filling-in (Paradiso and Hahn, 1996): observers might fail to experience the central region. This would reduce brightness and thus increase the probability of a black categorical judgment. There were differences in viewing conditions between the two experiments. First, eye positions were not controlled. In Experiment 1, the stimuli were presented in the peripheral visual field while the observers were fixating on the center of the display. In Experiment 2, observers freely moved their eyes. Second, sample presentation periods were not matched. Presentation duration was fixed at 0.3 s in Experiment 1 and open-ended in Experiment 2. The typical response time was 1 s, which was longer than the presentation duration in Experiment 1. The result might vary depending on gaze fixation and presentation time; however, brightness enhancement and the categorical labeling would appear the same across viewing conditions as demonstrated in Figure 1. To examine this, we performed an additional experiment that tested the effect of eye position on the appearance of stimuli (see D.2 details). As a result, observed thresholds of gray/white and white/luminous-white were the same between foveal and peripheral observation. In addition, the thresholds were the same as those in the original experiment in which eye position was not controlled. Thus, the discrepancy between brightness estimation and categorical judgments would likely persist, even if gaze were controlled in the categorical judgment task.

5.4.2 Importance of the illusion

The current findings suggest that the glare illusion is not an exceptional phenomenon that is limited to the perception of self-luminosity under specific conditions, but rather a general brightness enhancement that is simply generated by a luminance gradient surrounding an object. The glare illusion induces robust brightness enhancement, which does not depend on the luminance of the central region. This is an important and useful property for both scientific and engineering fields. Intense lights are not only generated from self-luminous objects, but are also present as specular highlights on glossy surfaces. Thus, it is natural to assume that when an intense light is present on part of an object's surface, the surface is glossy. The spatial configuration of an image affects both its perceived lightness and glossiness (Motoyoshi et al., 2007; Sharan et al., 2008). Gloss perception also depends on object segmentation and three-dimensional configuration (Anderson and Kim, 2009; Kim and Anderson, 2010; Marlow et al., 2015; Olkkonen and Brainard, 2010). Indeed, glossiness is perceived when it is generated from both bright specular highlights and dark specular "lowlights" (Kim et al., 2012). That both bright and "dark" lights induce glossiness is consistent with the current findings that illustrate the robustness of the glare illusion. Those two perceptions might share similar mechanisms. Recently, the effect of glossiness on lightness and color perception was investigated (Granzier et al., 2014; Olkkonen and Brainard, 2010); these interactions, and the glare illusion, might help us further understand lightness/brightness perception. Our current finding of constant brightness enhancement in the glare illusion is also potentially useful for computer graphics and image processing, to signify brighter and more luminous objects. The robustness of the illusion across illumination conditions is further useful if one wishes to show printed images.

5.4.3 Halo

We included Halo as one of the test conditions. If Halo were not included, repeated presentations of Glow would have caused an afterimage (Anstis et al., 2007) and subsequently affected the appearance of the following trial's stimulus. If the afterimage added to the sample stimulus brightness, Glow and Uniform would become similar to Uniform and Halo, reducing the illusion effect size. Indeed, we observed such effects in our preliminary experiments. We also noted interactions between the afterimage and the breathing light illusion, which was dynamically induced by the glare pattern (Anstis et al., 2007; Gori et al., 2010; Gori and Stubbs, 2006).

5.4.4 Comparison with studies of Gilchrist and Zavagno

Bonato and Gilchrist (1994) reported that a target begins to appear luminous when its luminance is 1.7 times that of a reference white. The stimulus configuration used in their study was similar to the uniform condition in our experiment, in which subjects responded that the stimulus appeared white above 123.9 cd/m^2, and luminous-white above the extrapolated value of 242.9 cd/m^2 (Fig. 3C). If we use 123.9 cd/m^2 as reference white, 242.9 cd/m^2 corresponds to a 1.96-fold increase. In contrast, the transition between the percepts of white and luminous-white in the Glow condition only required the stimulus luminance to be 1.23 times that of the reference white. This reduction in the threshold for the perception of self-luminosity represents the effect of the glare illusion. Zavagno (1999) indicated that luminosity perception occurred even when the target was darker than subjective white. Thus, there is a possibility of a "luminous-gray" response. Therefore, we should discuss whether our four available choices in Experiment 2 were appropriate. Alternatively, the subjects could be asked to respond in two steps, i.e., color naming and providing a luminosity judgment. If this response schedule was used, the luminosity threshold might be lower than suggested in our data, because our current task did not allow a "luminous-gray" response. However, the magnitude of the decrease would be limited, and it is unlikely that the threshold would drop to a very low level, such as 20 cd/m^2, which corresponds to the lower limit of brightness enhancement in Experiment 1. This is because if the "true" luminosity threshold were very low, white (and non-luminous) responses would not have been observed in our original task; however, the observed data showed substantial white (and non-luminous) responses. Thus, even if the luminosity threshold decreased further in a different task, our conclusion still holds: brightness enhancement is observed below the luminosity threshold.

5.4.5 Effect of edge contrast

Both Glow and Halo stimuli had similar image statistics except for the direction of the luminance gradient. The mean luminances of the annuli were 41% (Glow), 59% (Halo), and 44% (Uniform) to the center patch. If we assume that the contrast between the center and the mean luminance of the annulus is critical in determining brightness enhancement, the largest contrast should be obtained for Glow, followed by Uniform and Halo. This order is not consistent with the observed brightness enhancement in Experiment 1. Another consideration is the adjacent contrast between center and annulus. In this case, contrast is maximal in Halo (∼100%), followed by Uniform (44%), and least in Glow (∼0%). Again, this order is not consistent with the observed data. These luminance contrasts do not explain brightness enhancement in the glare illusion, and thus other image statistics must be considered. Brightness enhancement was determined via comparison to the particular reference stimuli of the annulus's luminance level of 44% to the center. Thus, the use of other luminance might induce different results. Low luminance level of the uniform annulus may have resulted in increased lightness of the center, leading to less brightness enhancement in the glare stimuli, as compared to the reference dark annulus. To examine the effect of reference selection, we performed an additional experiment that was similar to Experiment 1, except for the reference, whose annulus was black. Again, the Glow condition led to significant brightness enhancement as compared with the reference; indeed, the magnitude of the brightness enhancement was almost comparable (35%, see D.1 details). Thus, we conclude that brightness enhancement is robust across different luminances of target and reference stimuli.

5.4.6 Image restoring and glare illusion

An empirical explanation that may be applicable to the glare illusion is that it is similar to a hazy, multilayered scene (Anderson and Winawer, 2005, 2008). Blurring an intensely bright object corresponds to a convolution of the object with a point spread function, which results in a glare image. The glare illusion could be a perceptual process that restores the true image. For example, when an observer sees an object in a foggy scene, he or she would realize the true brightness of the object on coming closer to it. Once one learns the association between blurred and veridical images, it is possible to infer the true brightness of the image, even when only the blurred image is available. Similar optical effects would occur in a variety of situations such as the sun through a cloudy sky, a translucent lampshade, and defocusing the lens of our eye. The original image, which typically has a sharp outline, is always brighter than a blurred version of that image. Thus, perceived brightness would be enhanced in the glare illusion. This perceptual restoring process might be related to lightness constancy of the visual system. Since the blurring and restoring would not be limited to bright and self-luminous objects, but would also apply to surfaces with various reflectances, including those perceived as gray, it results in robust brightness enhancement in the illusion.

5.5 Conclusion

Brightness enhancement in the glare illusion is surprisingly robust across stimulus intensities, ranging from dark to light, including subjectively gray, white, and luminous appearances. The magnitude of brightness enhancement is almost constant and corresponds to an effective increase in the luminance of at least 30%. Thus, the surrounding gradient, which was originally thought to mimic the spread of intense light, strongly and robustly increases brightness estimation even for stimuli that are not self-luminous.

Chapter 6

The rotating glass illusion

A similar version of this chapter has been published as:

Tamura, H., & Nakauchi, S. (2018). The rotating glass illusion: material appearance is bound to perceived shape and motion, *i-Perception*, 9(6), 1-5, DOI:10.1177/2041669518816716.

We report a novel illusion in which a rotating transparent and refractive triangular prism (glass object) is perceived as being made of a specular reflective material (mirror), and simultaneously, its direction of rotation (clockwise or anticlockwise) is also misperceived. Our findings suggest that physical motion strongly influences viewers' judgements of material in some situations.

6.1 Introduction

Although motion aids in perception of a material and its surface properties on a rigid object (e.g., Doerschner et al., 2011; Tamura et al., 2018; Ueda et al., 2015), some particular types of motion of a refractive and transparent rigid object induce mistakes in viewers' perceptions of material and motion. We report a novel illusion in which a rotating refractive triangular prism is perceived as being of a specular reflective material and its direction of rotation is simultaneously misperceived. Figure 6.1(a) shows examples of the illusion (see Movie 6.1). A triangular prism with randomly distributed bumps, rendered using computer graphics, and rotates clockwise (when viewed from above) about the vertical axis. The left panel in Movie 6.1 shows this rotating object made of specular reflective material, and viewers can correctly discern its material and direction of rotation. In the right panel in Movie 6.1, however, viewers see a rotating object made of a transparent and refractive material, such as glass; from certain specific viewpoints, they perceive this as a specular reflective material, such as a mirror. Moreover, at that point, the object's direction of rotation (clockwise or anticlockwise) is perceived to be reversed.

6.2 Methods

We investigated the error rate in perception of material appearance and direction of rotation. The stimuli were videos of a bumpy triangular prism rotating through 30° from 1 of 12 starting positions (see Figure 6.1(b)). Two versions of each stimulus were prepared in which the prism was made of different materials: "mirror," a perfectly specular reflective surface, or "glass," a refractive medium (with a refractive index of 1.5; its reflectance and transmittance were 0.04 and 0.96, respectively). Stimuli were rendered using a physically based renderer Mitsuba (Jakob, 2010) under realistic illumination "Uffizi Gallery" (Debevec et al., 2000). One stimulus was for a second video with 60 frame/s refresh rate and the speed of the object's rotation was 0.5° per frame. Ten observers were exposed to the stimuli and asked to judge the material of the object (mirror or glass) in the material task. In different blocks, they were also asked to judge the object's direction of rotation (clockwise or anticlockwise) in the direction task. The order of the two blocks was counterbalanced.

6.3 Results

Figures 6.1(c) and 6.1(d) show the percentage of correct answers of all observers in the material and direction tasks, respectively. Although performance was stable for mirror stimuli, that for glass stimuli tended to be worse at specific starting angles (30°, 60°, 150°, 180°, 270°, and 300°) in the material task (Figure 6.1(c)), and there was a significant difference in performance depending on the combination of material and starting angle, with a two-way repeated measures analysis of variance indicating a significant interaction, $F(3.849, 34.645) = 9.371$, $p < .001$. Similarly, performance differed at the same specific angles in the direction task (Figure 6.1(d)), $F(3.278, 29.500) = 23.978$, $p < .001$.

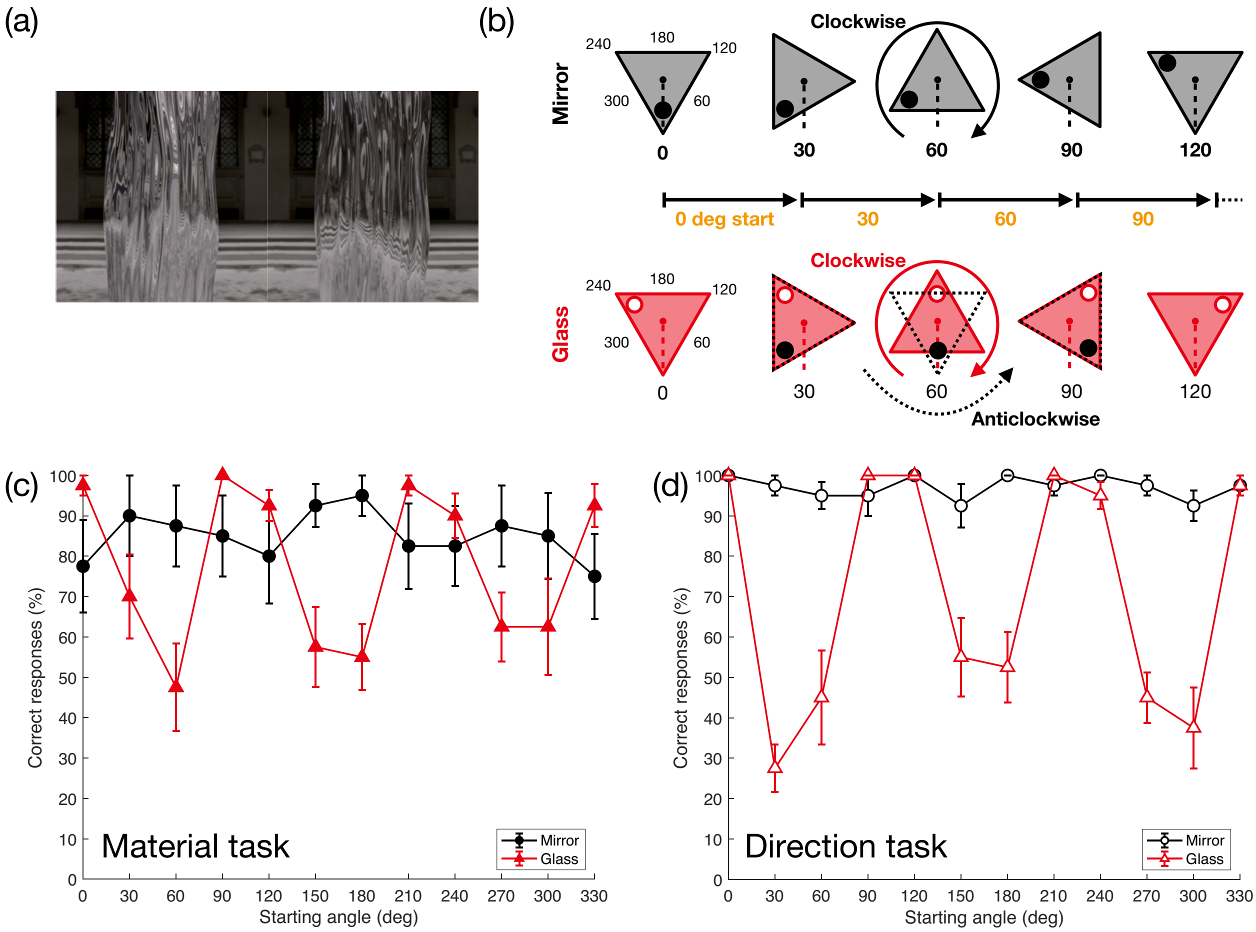


Figure 6.1: The illusion producing misperception of material and direction of rotation

(a) Example stimuli for the material and direction tasks (see also Movie 6.1). The left panel shows the mirror object and the right one, the glass object. (b) A diagram explaining where viewers misperceived the object. (c) Results of the material task. The horizontal axis indicates starting angles for the object's rotation. The vertical axis indicates the percentage of correct answers. Averages across all ten observers are shown; error bars represent the standard error of the mean. (d) Results of the direction task, presented as in (c).

These results suggest that the observers misperceived the appearance of the material and direction of rotation in both tasks, and that the starting positions in which these misperceptions occurred were consistent. Note that we present adjusted degrees of freedom using Greenhouse-Geisser correction when the criterion for the assumption of sphericity (using Mauchly's test) was not met.

6.4 Discussion

Although the object physically rotates in a fixed direction, the viewer's visual system misperceives the object's material and direction of rotation in certain instances, because the visual system relies on the components of motion to distinguish reflective and refractive materials (Tamura et al., 2018). For example, even if the object physically rotates clockwise, the opposite motion components anticlockwise could be dominant, depending on the complex light reflection and refraction resulting from the interactions between shape, surface properties, and illumination. This illusion suggests that these physical motions of a triangular prism with rich optical properties induce confusion in viewers' perceptions of material and motion.

The visual system narrows down a target structure by accumulating relative motion information at a given time for the structure based on its motion (e.g., Ullman, 1979). This allows for the ambiguity to be resolved when viewing only the object's front surface or both the front and the rear surfaces. At the specific starting angles at which the illusion occurs, a convex edge of the triangular prism made of glass was facing toward the rear (see Figure 6.1(b)). The visual system could be misperceiving this edge as that facing the front, as in the hollow-face illusion (Gregory, 1997; Hill and Johnston, 2007), thus reversing perception of the object's direction of rotation. This means that the visual system more easily recognizes the object when it has a specular reflective surface and tends to ignore refractive media.

From the viewpoint of a change in material appearance, this illusion is similar to the type of illusion in which a refractive object is perceived as a specular reflective object when it is turned upside-down (Kim and Marlow, 2016). However, in that case, the authors reported on a static image illusion; the illusion we report here is a video illusion and simultaneously changes the viewer's perceptions of material and motion. This illusion could be a new tool to further explore the relationship between material appearance and motion.

Chapter 7

Conclusion

In this thesis, we investigated how the visual system distinguishes mirror from glass and what visual cues contribute to this challenging task. Here, we summarize the findings and insights from the five studies.

7.1 Comparing humans and models

We compared the performance of humans and models (classifiers) for distinguishing mirror from glass with thousands of neural networks (Chapter 2)

Purpose

- Compare the performance of the visual system, the hand-engineered classifiers, and the feedforward neural networks for mirror/glass judgements.
- Develop a model that would behave like humans according to both the success and systematic errors.

Contribution

- On the random images, the CNNs overperformed the other models because they learn many more features but not simply resemble those used by the visual system.
- We proposed the use of diagnostic images that perfectly decorrelate the true material class from the perceived class.
- Despite the extensive and guided search, none of the OptCNNs we investigated correlated better than 0.6 with human judgements, implying the existence of three important respects, such as the feedback, training objective, and task generality.

- RSA and noise addition revealed that the OptCNN was gradually organized and acquired robustness to noise like humans through the training, respectively.

We mainly discussed why the models differ from the human brain and behaviors. Three important respects, such as feedback processing in the network architecture, the nature of the training objective function, and the nature of the task have potential to be the networks that resemble human visual processes. In future work, we will use a combination of unsupervised learning, more naturalistic objective functions, and network architectures that more closely resemble the human visual system.

7.2 Static and dynamic visual cues

We tested what visual cues contribute to distinguish mirror from glass (Chapters 3 and 4).

Purpose

- Clarify visual static (Chapter 3) and dynamic (Chapter 4) cues that contribute to distinguishing mirror from glass as image and video stimuli, respectively.

Contribution

- Modifying the luminance and color saturation profiles along the trajectory from the object contour to its center changes material appearances between mirror and glass.
- We proposed the image editing method, which can change the object material in an image using these simple cues.
- The motion ratio between the direction of object rotation and its opposite direction contributed to determining the extent of material appearance between transparent and specular reflective objects.
- Our model explained well that the visual system uses dynamic information for material perception of a rotating object.

We showed that both luminance and color saturation of static images contribute to determining material appearances. On this basis, we changed an object material to be more mirror-like or glass-like even though they have complex optical properties. In addition, we connected the refractive index of an object with the magnitude of perception between mirror and glass. This will be a better bridge between a physical factor and our perception. Thus, constructing a model with complex materials and motion is one of the future tasks to further understand the visual system.

7.3 Illuisions

We tested properties of luminosity perception, which occurred by specular highlights derived from mirror and glass materials (Chapter 5). Moreover, we discovered a new illusion in which the material and rotation direction of a rotating glass object is misperceived, and tested the effect of the illusion using human psychophysics (Chapter 6).

Purpose

- Test how the glare illusion elicits a brightness enhancement and a self-luminous appearance.
- Clarify fundamental properties of a new illusion, which is defined as the rotating glass illusion.

Contribution

- Brightness enhancement in the glare illusion was robust across stimulus intensities ranging from dark to light with subjective gray, white, and luminous appearances.
- We found that the interaction between shape, surface properties, and illumination strongly affects our material and motion judgements.

These findings will be a new tool to provide further insights for understanding mirror and glass perception.

7.4 Future work

Although distinguishing mirror from glass is a challenging problem in material recognition/classification, our neural networks performed well in terms of accuracy (CNN with the random images) and correlation (OptCNN with the diagnostic images). Despite this, there is still a gap between our globally optimized model and the visual system. To fill it, we can design a more general objective function for the training and plausible networks with both feedforward and feedback architectures. Furthermore, our approach can expand the possibilities for other materials or optical properties even though we only focused on two specific materials in this thesis.

In addition, the proposed models based on static and dynamic cues correlated well to human behavior. These cues are simple and easy to measure, and allow us to design a new system of object recognition or discrimination based on the human visual system. It would have high accuracy, more robustness to noise, and particularly similar behavior to humans. Thus, this thesis clarified various aspects of distinguishing mirror from glass and provided further challenges.

References

Adams, W. J., Elder, J. H., Graf, E. W., Leyland, J., Lugtigheid, A. J., and Muryy, A. (2016). The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. *Scientific Reports*, 6:35805.

Adams, W. J., Graf, E. W., and Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, 7(10):1057–1058.

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

Adelson, E. H. (2000). Lightness Perception and Lightness Illusions. In Gazzaniga, M., editor, *The New Cognitive Neurosciences*, volume 3, pages 339–351. MIT Press, Cambridge, MA, 2nd edition.

Adelson, E. H. (2001). On Seeing Stuff: The Perception of Materials by Humans and Machines. In Rogowitz, B. E. and Pappas, T. N., editors, *Proceedings of the. SPIE*, pages 1–12.

Adelson, E. H. and Pentland, A. P. (1996). The perception of shading and reflectance. In Knill, D. and Richards, W., editors, *Perception as Baysian Inference*, volume 1, pages 409–423. Cambridge University Press, New York.

Agostini, T. and Galmonte, A. (2002). A new effect of luminance gradient on achromatic simultaneous contrast. *Psychonomic Bulletin and Review*, 9(2):264–269.

Anderson, B., Mooney, S. W. J., and Anderson, B. L. (2014). Specular Image Structure Modulates the Perception of Three-Dimensional Shape Specular Image Structure Modulates the Perception of Three-Dimensional Shape. *Current Biology*, 24(22):2737–2742.

Anderson, B. L. (2011). Visual perception of materials and surfaces. *Current Biology*, 21(24):R978–R983.

Anderson, B. L. and Kim, J. (2009). Image statistics do not explain the perception of gloss and lightness. *Journal of Vision*, 9(11):1–17.

Anderson, B. L. and Winawer, J. (2005). Image segmentation and lightness perception. *Nature*, 434(7029):79–83.

Anderson, B. L. and Winawer, J. (2008). Layered image representations and the computation of surface lightness. *Journal of Vision*, 8(7):1–22.

Anstis, S., Gori, S., and Wehrhahn, C. (2007). Afterimages and the Breathing Light Illusion. *Perception*, 36(5):791–794.

Arend, L. E. and Goldstein, R. (1987). Simultaneous constancy, lightness, and brightness. *Journal of the Optical Society of America A*, 4(12):2281–2285.

Bonato, F. and Gilchrist, A. L. (1994). The perception of luminosity on different backgrounds and in different illuminations. *Perception*, 23(9):991–1006.

Boyaci, H., Fang, F., Murray, S. O., and Kersten, D. (2007). Supplemental Data Responses to Lightness Variations in Early Human Visual Cortex. *Current Biology*, 17(11):5–7.

Bradley, D. C., Chang, G. C., and Andersen, R. A. (1998). Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature*, 392(6677):714–717.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4):433–6.

Budd, J. M. (1998). Extrastriate feedback to primary visual cortex in primates: a quantitative analysis of connectivity. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1400):1037–1044.

Chowdhury, N. S., Marlow, P. J., and Kim, J. (2017). Translucency and the perception of shape. *Journal of Vision*, 17(3):1–14.

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

Correani, A., Scott-Samuel, N. E., and Leonards, U. (2006). Luminosity—A perceptual “ feature ” of light-emitting objects? *Vision Research*, 46(22):3915–3925.

Dai, J., He, K., and Sun, J. (2015). Instance-aware Semantic Segmentation via Multi-task Network Cascades. *arXiv*, 1512.04412.

Debevec, P. (1998). Rendering synthetic objects into real scenes. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques - SIGGRAPH '98*, pages 189–198, New York, New York, USA. ACM Press.

Debevec, P., Hawkins, T., Tchou, C., Duiker, H.-P., Sarokin, W., and Sagar, M. (2000). Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques - SIGGRAPH '00*, pages 145–156, New York, New York, USA. ACM Press.

Doerschner, K., Fleming, R. W., Yilmaz, O., Schrater, P. R., Hartung, B., and Kersten, D. (2011). Visual Motion and the Perception of Surface Material. *Current Biology*, 21(23):2010–2016.

Doerschner, K., Yilmaz, O., Kucukoglu, G., and Fleming, R. W. (2013). Effects of surface reflectance and 3D shape on perceived rotation axis. *Journal of Vision*, 13(11):1–23.

Dovencioglu, D. N., Ben-shahar, O., Barla, P., and Doerschner, K. (2017). Specular motion and 3D shape estimation. *Journal of Vision*, 17(6:3):1–15.

Dovencioglu, D. N., Wijntjes, M. W. A., Ben-Shahar, O., and Doerschner, K. (2015). Effects of surface reflectance on local second order shape estimation in dynamic scenes. *Vision Research*, 115:218–230.

Eagleman, D. M. (2001). Visual illusions and neurobiology. *Nature Reviews Neuroscience*, 2(12):920–926.

Evans, R. M. (1959). Fluorescence and Gray Content of Surface Colors. *Journal of the Optical Society of America*, 49(11):1049–1059.

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

Fleming, R. W. (2012). Human perception: Visual heuristics in the perception of glossiness. *Current Biology*, 22(20):R865–R866.

Fleming, R. W. (2014). Visual perception of materials and their properties. *Vision Research*, 94(24):62–75.

Fleming, R. W. (2017). Material Perception. *Annual Review of Vision Science*, 3(1):365–388.

Fleming, R. W. and Bülthoff, H. H. (2005). Low-Level Image Cues in the Perception of Translucent Materials. *ACM Transactions on Applied Perception*, 2(3):346–382.

Fleming, R. W., Dror, R. O., and Adelson, E. H. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, 3(5):347–368.

Fleming, R. W., Jäkel, F., and Maloney, L. T. (2011). Visual Perception of Thick Transparent Materials. *Psychological Science*, 22(6):812–820.

Fleming, R. W., Torralba, A., and Adelson, E. H. (2004). Specular reflections and the perception of shape. *Journal of Vision*, 4(9):798–820.

Fleming, R. W., Wiebel, C., and Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, 13(8):1–20.

Geirhos, R., Janssen, D. H. J., Schütt, H. H., Rauber, J., Bethge, M., and Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv*, 1706.06969.

Ghodrati, M., Farzmahdi, A., Rajaei, K., Ebrahimpour, R., and Khaligh-Razavi, S.-M. (2014). Feed-forward object-vision models only tolerate small image variations compared to human. *Frontiers in Computational Neuroscience*, 8:74.

Gilchrist, A. (2007). *Seeing Black and White.* Oxford University Press, New York.

Gilchrist, A., Kossyfidis, C., Bonato, F., Cataliotti, J., Annan, V., and Economou, E. (1999). An anchoring theory of lightness perception. *Psychological Review*, 106(4):795–834.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323.

Goda, N., Tachibana, A., Okazawa, G., and Komatsu, H. (2014). Representation of the Material Properties of Objects in the Visual Cortex of Nonhuman Primates. *Journal of Neuroscience*, 34(7):2660–2673.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680.

Gori, S., Agostini, T., and Giora, E. (2010). Measuring the Breathing Light Illusion by means of induced simultaneous contrast. *Perception*, 39(1):5–12.

Gori, S. and Stubbs, D. A. (2006). Last but not least: A new set of illusions - The dynamic luminance-gradient illusion and the breathing light illusion. *Perception*, 35(11):1573–1577.

Granzier, J. J. M., Vergne, R., and Gegenfurtner, K. R. (2014). The effects of surface gloss and roughness on color constancy for real 3-D objects. *Journal of Vision*, 14(2):1–20.

Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352(1358):1121–1127.

Grunewald, A., Bradley, D. C., and Andersen, R. a. (2002). Neural correlates of structure-from-motion perception in macaque V1 and MT. *Journal of Neuroscience*, 22(14):6195–207.

Guclu, U. and van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014.

Hanada, M. (2012). Luminance profiles of luminance gradients affect the feeling of dazzling. *Perception*, 41(7):791–802.

Haynes, J.-D., Lotto, R. B., and Rees, G. (2004). Responses of human visual cortex to uniform surfaces. *Proceedings of the National Academy of Sciences of the United States of America*, 101(12):4286–4291.

Hecht, E. (1998). *Hecht optics.* Addison Wesley.

Hill, H. and Johnston, A. (2007). The Hollow-Face Illusion: Object-Specific Knowledge, General Assumptions or Properties of the Stimulus? *Perception*, 36(2):199–223.

Hiramatsu, C., Goda, N., and Komatsu, H. (2011). Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. *NeuroImage*, 57(2):482–494.

Ho, Y. X., Landy, M. S., and Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture: Research article. *Psychological Science*, 19(2):196–204.

Hong, H., Yamins, D. L. K., Majaj, N. J., and DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, 19(4):613–622.

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv*, 1502.03167.

Jain, A. and Zaidi, Q. (2011). Discerning nonrigid 3D shapes from motion cues. *Proceedings of the National Academy of Sciences of the United States of America*, 108(4):1663–1668.

Jakob, W. (2010). Mitsuba renderer. http:/www.mitsuba-renderer.org.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., and Mur, M. (2017). Deep Convolutional Neural Networks Outperform Feature-Based But Not Categorical Models in Explaining Object Similarity Judgments. *Frontiers in Psychology*, 8:1726.

Kakimoto, M., Matsuoka, K., Nishita, T., Naemura, T., and Harashima, H. (2005). Glare Generation Based on Wave Optics. *Computer Graphics Forum*, 24(2):185–193.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Kawabe, T. and Kogovšek, R. (2017). Image deformation as a cue to material category judgment. *Scientific Reports*, 7:44274.

Kawabe, T., Maruya, K., Fleming, R. W., and Nishida, S. (2015a). Seeing liquids from visual motion. *Vision Research*, 109:125–138.

Kawabe, T., Maruya, K., and Nishida, S. (2015b). Perceptual transparency from image deformation. *Proceedings of the National Academy of Sciences of the United States of America*, 112(33):E4620–E4627.

Keil, M. S. (2007). Gradient representations and the perception of luminosity. *Vision Research*, 47(27):3360–3372.

Kelley, L. A. and Kelley, J. L. (2014). Animal visual illusion and confusion: The importance of a perceptual perspective. *Behavioral Ecology*, 25(3):450–463.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11):e1003915.

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., and Masquelier, T. (2016a). Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. *Scientific Reports*, 6:32672.

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., and Masquelier, T. (2016b). Humans and Deep Networks Largely Agree on Which Kinds of Variation Make Object Recognition Harder. *Frontiers in Computational Neuroscience*, 10:92.

Kietzmann, T. C., McClure, P., and Kriegeskorte, N. (2018). Deep Neural Networks In Computational Neuroscience. *bioRxiv*, 133504.

Kim, J. and Anderson, B. L. (2010). Image statistics and the perception of surface gloss and lightness. *Journal of Vision*, 10(9):1–17.

Kim, J. and Marlow, P. J. (2016). Turning the World Upside Down to Understand Perceived Transparency. *i-Perception*, 7(5):1–5.

Kim, J., Marlow, P. J., and Anderson, B. L. (2012). The dark side of gloss. *Nature Neuroscience*, 15(11):1590–1595.

Kim, T. (2017). DCGAN in Tensorflow. https://github.com/carpedm20/DCGAN-tensorflow.

Kingdom, F. (1997). Simultaneous contrast: the legacies of Hering and Helmholtz. *Perception*, 26(6):673–677.

Kingdom, F. A. A. (2011). Lightness, brightness and transparency: A quarter century of new ideas, captivating demonstrations and unrelenting controversy. *Vision Research*, 51(7):652–673.

Kinoshita, M. and Komatsu, H. (2001). Neural representation of the luminance and brightness of a uniform surface in the macaque primary visual cortex. *Journal of Neurophysiology*, 86(5):2559–70.

Kleiner, M., Brainard, D. H., Pelli, D. G., Broussard, C., Wolf, T., and Niehorster, D. (2007). What's new in Psychtoolbox-3? *Perception*, 36(14):1–16.

Knill, D. C. and Kersten, D. (1991). Apparent surface curvature affects lightness perception. *Nature*, 351(6323):228–30.

Komatsu, H. and Goda, N. (2018). Neural Mechanisms of Material Perception: Quest on Shitsukan. *Neuroscience*, 392:329–347.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1(1):417–446.

Kriegeskorte, N. and Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, 21(9):1148–1160.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Kubilius, J., Bracci, S., and Op de Beeck, H. P. (2016). Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLOS Computational Biology*, 12(4):e1004896.

Land, E. H. and McCann, J. J. (1971). Lightness and Retinex Theory. *Journal of the Optical Society of America*, 61(1):1–11.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

LeCun Jackel, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D., Cun, B. L., Denker, J., and Henderson, D. (1990). Handwritten Digit Recognition with a Back-Propagation Network. *Advances in Neural Information Processing Systems 2*, pages 396–404.

Leonards, U., Troscianko, T., Lazeyras, F., and Ibanez, V. (2005). Cortical distinction between the neural encoding of objects that appear to glow and those that do not. *Cognitive Brain Research*, 24(1):173–176.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *IJCAI'81 Proceedings of the 7th international joint conference on Artificial intelligence*, pages 674–679.

MacEvoy, S. P. and Paradiso, M. a. (2001). Lightness constancy in primary visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 98(15):8827–8831.

Majaj, N. J., Hong, H., Solomon, E. A., and DiCarlo, J. J. (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, 35(39):13402–13418.

Majaj, N. J. and Pelli, D. G. (2018). Deep learning—Using machine learning to study biological vision. *Journal of Vision*, 18(13):1–13.

Mamassian, P. and Goutcher, R. (2001). Prior knowldege on the illumination position. *Cognition*, 81(1):B1–B9.

Marlow, P. J. and Anderson, B. L. (2015). Material properties derived from three-dimensional shape representations. *Vision Research*, 115:199–208.

Marlow, P. J. and Anderson, B. L. (2016). Motion and texture shape cues modulate perceived material properties. *Journal of Vision*, 16(1):1–14.

Marlow, P. J., Kim, J., and Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, 22(20):1909–1913.

Marlow, P. J., Todorović, D., and Anderson, B. L. (2015). Coupled computations of three-dimensional shape and material. *Current Biology*, 25(6):R221–R222.

Motoyoshi, I. (2010). Highlight-shading relationship as a cue for the perception of translucent and transparent materials. *Journal of Vision*, 10(9):1–11.

Motoyoshi, I., Nishida, S., and Adelson, E. H. (2005). Luminance re-mapping for the control of apparent material. In *Proceedings of the 2nd symposium on Appied perception in graphics and visualization - APGV '05*, page 165, New York, New York, USA. ACM Press.

Motoyoshi, I., Nishida, S., Sharan, L., and Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature*, 447(7141):206–209.

Muckli, L. and Petro, L. S. (2013). Network interactions: non-geniculate input to V1. *Current Opinion in Neurobiology*, 23(2):195–201.

Murray, R. F. (2013). The Statistics of Shape, Reflectance, and Lighting in Real-World Scenes. In Dickson, S. J. and Pizlo, Z., editors, *Shape Perception in Human and Computer Vision*, pages 225–235. Springer, London.

Nagai, T., Matsushima, T., Koida, K., Tani, Y., Kitazaki, M., and Nakauchi, S. (2015). Temporal properties of material categorization and material rating: Visual vs non-visual material features. *Vision Research*, 115:259–270.

Nakamae, E., Kaneda, K., Okamoto, T., and Nishita, T. (1990). A lighting model aiming at drive simulators. In *Proceedings of the 17th annual conference on Computer graphics and interactive techniques - SIGGRAPH '90*, pages 395–404.

Nishio, A., Shimokawa, T., Goda, N., and Komatsu, H. (2014). Perceptual Gloss Parameters Are Encoded by Population Responses in the Monkey Inferior Temporal Cortex. *Journal of Neuroscience*, 34(33):11143–11151.

Nowlan, S. and Sejnowski, T. (1995). A selection model for motion processing in area MT of primates. *Journal of Neuroscience*, 15(2):1195–1214.

Olkkonen, M. and Brainard, D. H. (2010). Perceived glossiness and lightness under real-world illumination. *Journal of Vision*, 10(9):1–19.

Paradiso, M. A. and Hahn, S. (1996). Filling-in percepts produced by luminance modulation. *Vision Research*, 36(17):2657–2663.

Paulun, V. C., Kawabe, T., Nishida, S., and Fleming, R. W. (2015). Seeing liquids from static snapshots. *Vision Research*, 115:163–174.

Paulun, V. C., Schmidt, F., van Assen, J. J. R., and Fleming, R. W. (2017). Shape, motion, and optical cues to stiffness of elastic objects. *Journal of Vision*, 17(1):1–22.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, 10(4):437–442.

Portilla, J. and Simoncelli, E. P. (2000). A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision*, 40(1):49–71.

Qian, N. and Andersen, R. (1994). Transparent motion perception as detection of unbalanced motion signals. II. Physiology. *Journal of Neuroscience*, 14(12):7367–7380.

Qian, N., Andersen, R., and Adelson, E. (1994). Transparent motion perception as detection of unbalanced motion signals. I. Psychophysics. *Journal of Neuroscience*, 14(12):7357–7366.

Radford, A., Metz, L., and Chintala, S. (2015). DCGAN: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv*, 1511.06434.

Radonjić, A., Allred, S. R., Gilchrist, A. L., and Brainard, D. H. (2011). The dynamic range of human lightness perception. *Current Biology*, 21(22):1931–1936.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J. J. (2018). Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *Journal of Neuroscience*, 38(33):7255–7269.

Ramachandran, V. (1988). Perception of shape from shading. *Nature*, 331(6152):136–6.

Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J. (2002). Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 21(3):267–276.

Roe, A. W. and Tso, D. Y. (1995). Visual Topography in Primate V2 - Multiple Representation across Functional Stripes. *Journal of Neuroscience*, 15(5):3689–3715.

Rokita, P. (1993). A model for rendering high intensity lights. *Computers and Graphics*, 17(4):431–437.

Rosenberg, A., Wallisch, P., and Bradley, D. C. (2008). Responses to direction and transparent motion stimuli in area FST of the macaque. *Visual Neuroscience*, 25:187–195.

Sakano, Y. and Ando, H. (2010). Effects of head motion and stereo viewing on perceived glossiness. *Journal of Vision*, 10(9):1–14.

Schlüter, N. and Faul, F. (2014). Are optical distortions used as a cue for material properties of thick transparent objects? *Journal of Vision*, 14(14):1–14.

Schlüter, N. and Faul, F. (2016). Matching the Material of Transparent Objects: The Role of Background Distortions. *i-Perception*, 7(5):1–24.

Schmidt, F., Paulun, V. C., van Assen, J. J. R., and Fleming, R. W. (2017). Inferring the stiffness of unfamiliar objects from optical, shape and motion cues. *Journal of Vision*, 17(3):1–17.

Schneck, C. M. (2010). *Visual perception.* Academic Press, New York.

Shapley, R., Shapley, R., Enroth-cugell, C., and Enroth-cugell, C. (1984). Visual adaptation and retinal gain control. In *Progress in Retinal Research*, volume 3, pages 263–346. Elsevier.

Sharan, L., Li, Y., Motoyoshi, I., Nishida, S., and Adelson, E. H. (2008). Image statistics for surface reflectance perception. *Journal of the Optical Society of America A*, 25(4):846–65.

Sharan, L., Rosenholtz, R., and Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of Vision*, 14(9):1–24.

Shevell, S. K., Holliday, I., and Whittle, P. (1992). Two separate neural mechanisms of brightness induction. *Vision Research*, 32(12):2331–2340.

Shinya, M., Saito, T., and Takahashi, T. (1989). Rendering techniques for transparent objects. In *Graphics Interface*, volume 89, pages 173–182.

Simonyan, K. and Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. *arXiv*, 1406.2199.

Snowden, R. (1999). Motion transparency: making models of motion perception transparent. *Trends in Cognitive Sciences*, 3(10):369–377.

Snowden, R., Treue, S., Erickson, R., and Andersen, R. (1991). The response of area MT and V1 neurons to transparent motion. *Journal of Neuroscience*, 11(9):2768–2785.

Spencer, G., Shirley, P., Zimmerman, K., and Greenberg, D. P. (1995). Physically-based glare effects for digital images. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques - SIGGRAPH '95*, pages 325–334. ACM.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Sun, J. and Perona, P. (1998). Where is the sun? *Nature Neuroscience*, 1(3):183–184.

Tamura, H., Higashi, H., and Nakauchi, S. (2018). Dynamic Visual Cues for Differentiating Mirror and Glass. *Scientific Reports*, 8:8403.

Tamura, H. and Nakauchi, S. (2018). The Rotating Glass Illusion: Material Appearance Is Bound to Perceived Shape and Motion. *i-Perception*, 9(6):1–5.

Tanaka, M. and Horiuchi, T. (2015). Investigating perceptual qualities of static surface appearance using real materials and displayed images. *Vision Research*, 115:246–258.

Tani, Y., Araki, K., Nagai, T., Koida, K., Nakauchi, S., and Kitazaki, M. (2013). Enhancement of Glossiness Perception by Retinal-Image Motion: Additional Effect of Head-Yoked Motion Parallax. *PLoS ONE*, 8(1):e54549.

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582):520–522.

Uchikawa, K., Koida, K., Meguro, T., Yamauchi, Y., and Kuriki, I. (2001). Brightness, not luminance, determines transition from the surface-color to the aperture-color mode for colored lights. *Journal of the Optical Society of America A*, 18(4):737–746.

Ueda, S., Tani, Y., Nagai, T., Koida, K., Nakauchi, S., and Kitazaki, M. (2015). Perception of a thick transparent object is affected by object and background motions but not dependent on the motion speed. *Journal of Vision*, 15(12):823.

Ullman, S. (1979). The Interpretation of Structure from Motion. *Proceedings of the Royal Society B: Biological Sciences*, 203(1153):405–426.

van Assen, J. J., Nishida, S., and Fleming, R. W. (2018a). Estimating perceived viscosity of liquids with neural networks. In *41st European Conference on Visual Perception (ECVP2018)*.

van Assen, J. J. R., Barla, P., and Fleming, R. W. (2018b). Visual Features in the Perception of Liquids. *Current Biology*, 28(3):452–458.

van Assen, J. J. R. and Fleming, R. W. (2016). Influence of optical material properties on the perception of liquids. *Journal of Vision*, 16(15):1–20.

Walraven, J., Enroth-Cugell, C., Hood, D. C., MacLeod, D. I., and Schnapf, J. L. (1990). The control of visual sensitivity: Receptoral and postreceptoral processes. In *Visual Perception*, pages 53–101. Elsevier.

Wendt, G., Faul, F., Ekroll, V., and Mausfeld, R. (2010). Disparity, motion, and color information improve gloss constancy performance. *Journal of Vision*, 10(9):1–17.

Whittle, P. (1994). Contrast Brightness and Ordinary Seeing. In Gilchrist, A., editor, *Lightness, Brightness, and Transparency*, pages 111–158. Lawrence Erlbaum Associates, Hillsdale, NJ, England.

Witzel, C. and Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of Vision*, 13(7):1–33.

Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–24.

Yoshida, A., Ihrke, M., Mantiuk, R., and Seidel, H.-P. (2008). Brightness of the glare illusion. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization - APGV '08*, pages 83–90, New York, New York, USA. ACM Press.

Zaidi, Q. (2011). Visual inferences of material changes: color as clue and distraction. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(6):686–700.

Zavagno, D. (1999). Some new luminance-gradient effects. *Perception*, 28:835–838.

Zavagno, D., Annan, V., and Caputo, G. (2004). The problem of being white: Testing the highest luminance rule. *Vision*, 16(3):149–159.

Zavagno, D. and Caputo, G. (2001). The glare effect and the perception of luminosity. *Perception*, 30(2):209–222.

Zavagno, D. and Caputo, G. (2005). Glowing greys and surface-white: The photo-geometric factors of luminosity perception. *Perception*, 34(3):261–274.

Zavagno, D. and Massironi, M. (1997). La Rappresentazione Della Luce Nelle Opere D'arte Grafica. *Giornale Italiano Di Psicologia*, 24(1):135–187.

Appendix A

Comparing humans and models

A.1 Supplement information

Illuminations

- https://syns.soton.ac.uk/
- http://www.pauldebevec.com/
- http://hdrmaps.com/freebies
- http://dativ.at/lightprobes/
- http://www.openfootage.net/?cat=15
- https://hdrihaven.com/hdris.php?thumb=all&sort=date&search=all&page=2&npp=12
- https://www.doschdesign.com/

A.2 Supplement figures

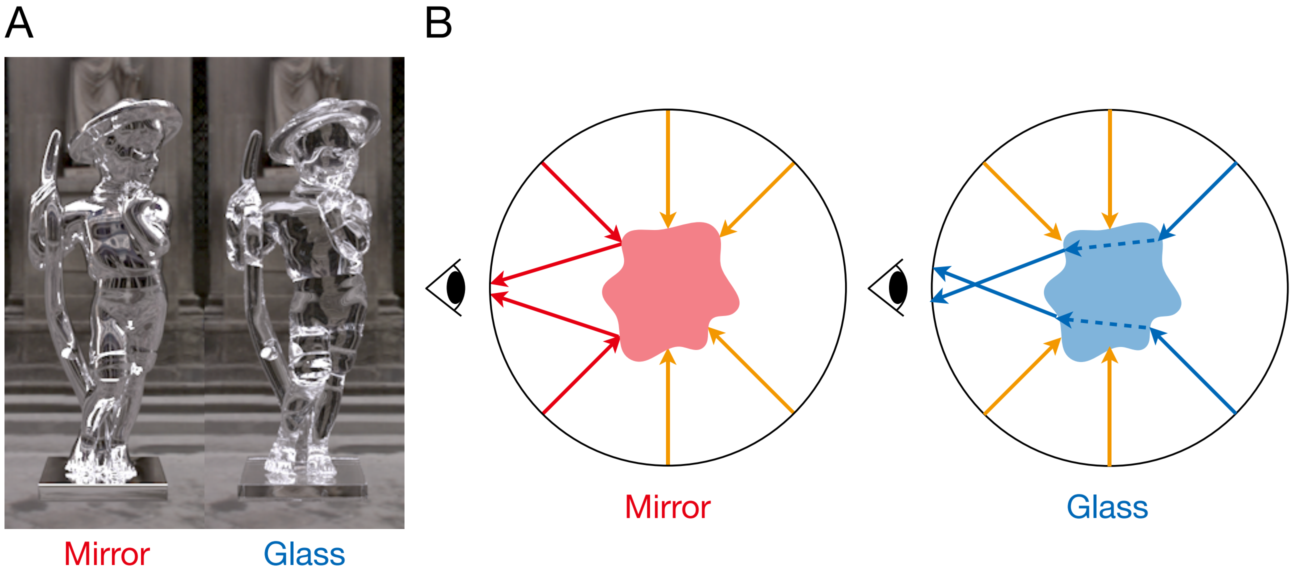
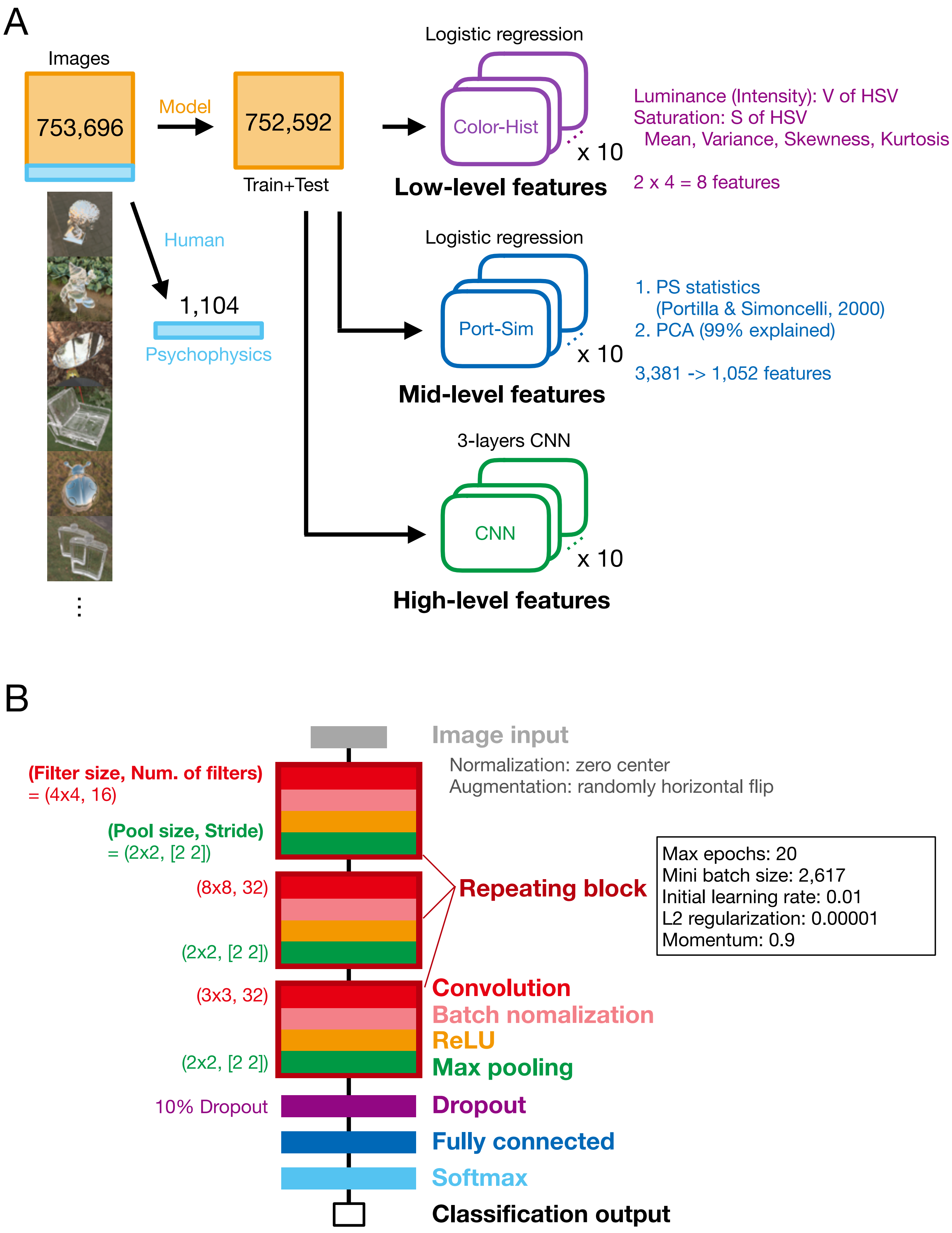


Figure A.1: Distinguishing mirror from glass

(A) Example objects made of mirror and glass materials. The left and right image show mirror and glass objects, respectively. Those of 3D shapes, illuminations, and camera positions are identical but the object's optical properties are different. (B) Illustration of different light paths through mirror and glass objects.



Max epochs: 20
Mini batch size: 2,617
Initial learning rate: 0.01
L2 regularization: 0.00001
Momentum: 0.9

Figure A.2: Classifiers and CNN architecture

(A) Flowchart of developing three classifiers. (B) Network architecture of CNN. The text box shows the hyper-parameters for training the CNN. The other hyper-parameters were the same as default settings in MATLAB.

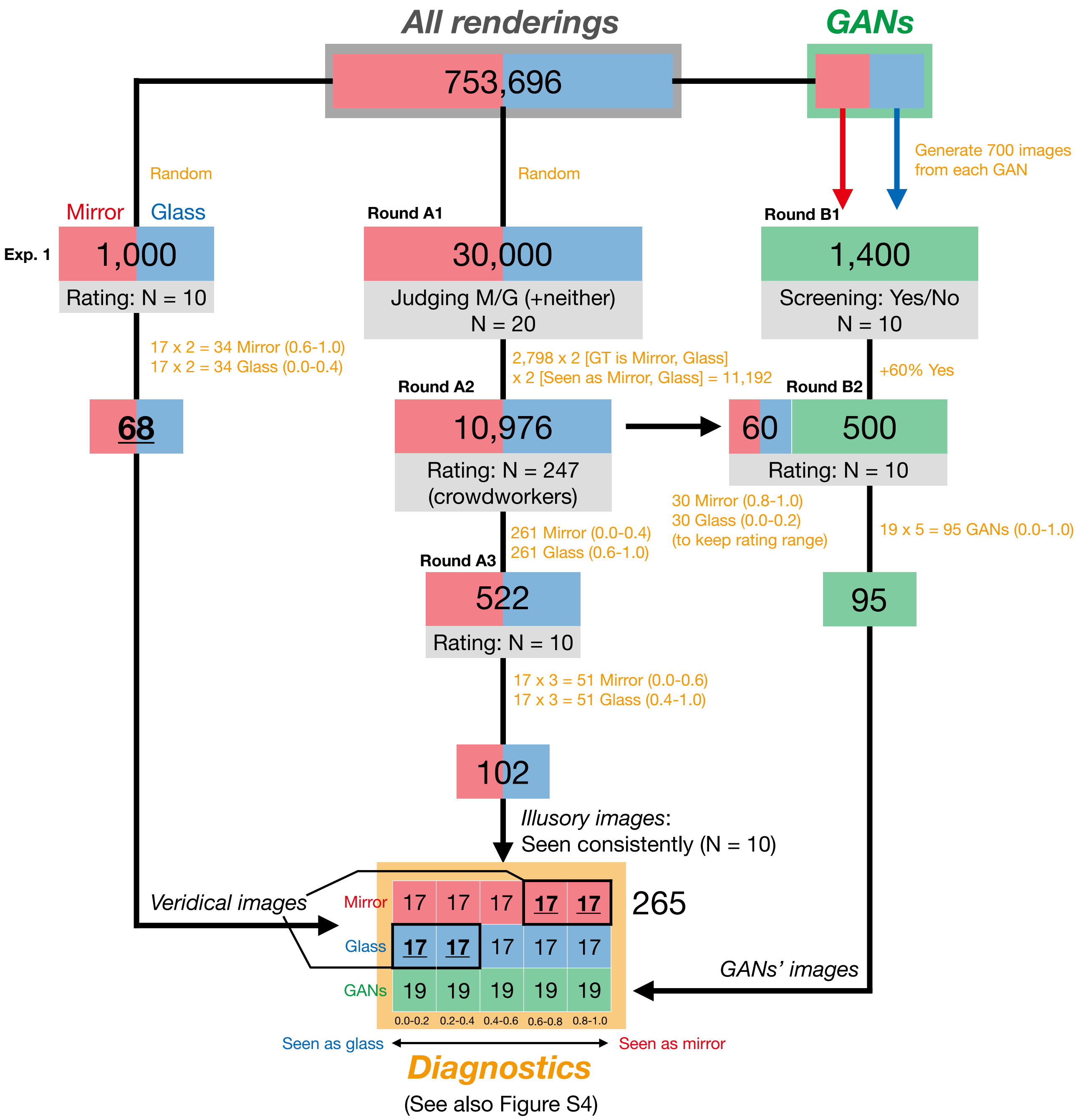


Figure A.3: Flowchart to create the diagnostic image set

See also 2.4.

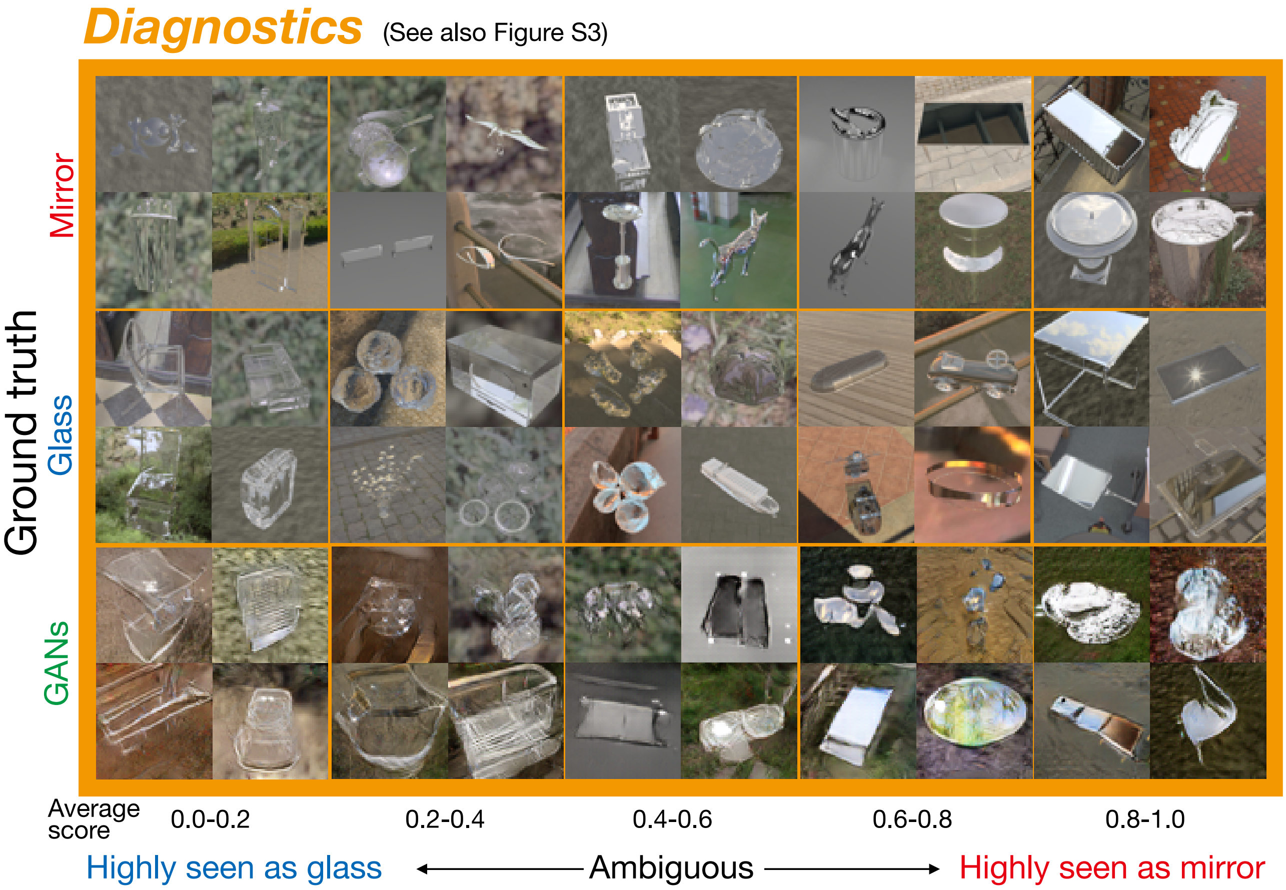
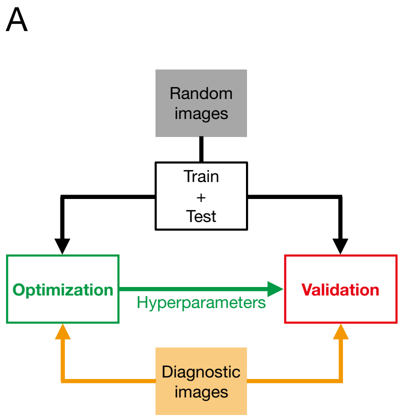


Figure A.4: Example images from the diagnostic image set

Each row indicates different ground truth (mirror, glass, and GANs). Each column indicates average rating score of 10 observers.



B Optimized hyperparameters

	Name	Min	Max	Type	Transform
1	Initial learning rate	10^{-7}	10^{-1}	Real	Log
2	Factor for L_2 regularization	10^{-7}	10^{-1}	Real	Log
3	Mini batch size	100	2000	Real	-
4	Momentum	0.9	0.96	Real	-
5	Dropout rate	0	0.3	Real	-
6	Filter size 1	2	3	Integer	-
7	Filter size 2	2	4	Integer	-
8	Filter size 3	2	4	Integer	-
9	Num. of filters 1	4	8	Integer	-
10	Num. of filters 2	4	30	Integer	-
11	Num. of filters 3	4	30	Integer	-

Figure A.5: Systematic exploration of the space of feedforward networks

(A) Illustration of the optimization-stage and validation-stage in exploration. (B) Eleven hyper-parameters controlling the network architecture. Each hyper-parameter was searched in the range between 'min' and 'max'. The fist two hyper-parameters were transformed to log space during the searching. A pair of filter size 1 and num. of filters 1 was set to the convolution layers from the 1st to the n-2 th repeating block. A pair of filter size 2 and num. of filters 2, and a pair of filter size 3 and num. of filters 3 were set to the convolution layer in the penultimate and last repeating blocks. The learning rate was dropped by 0.1 times in each 10 epochs.

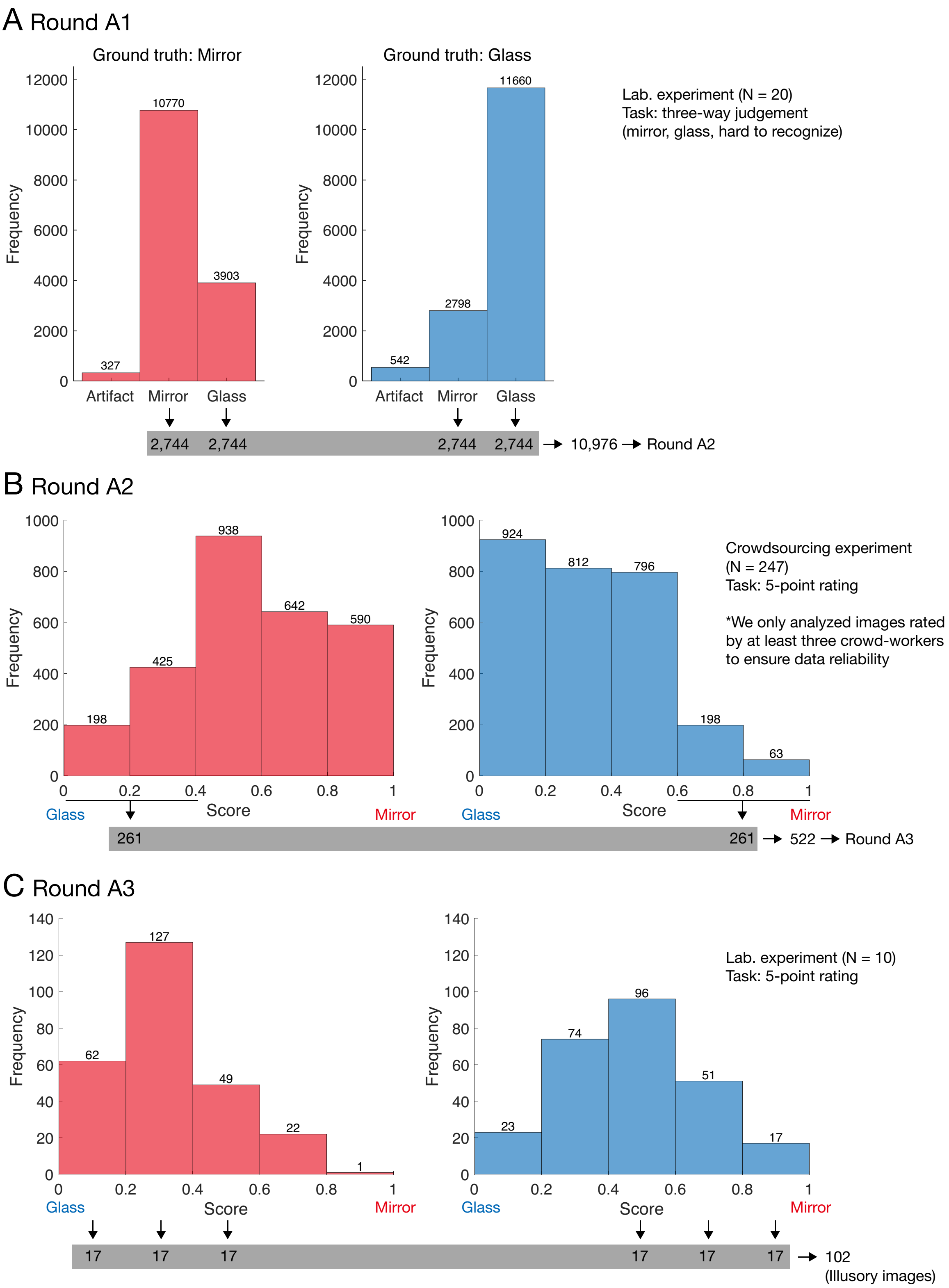


Figure A.6: Results of Round A1-A3

Each panel shows result of each round as histogram of images. See also 2.4.

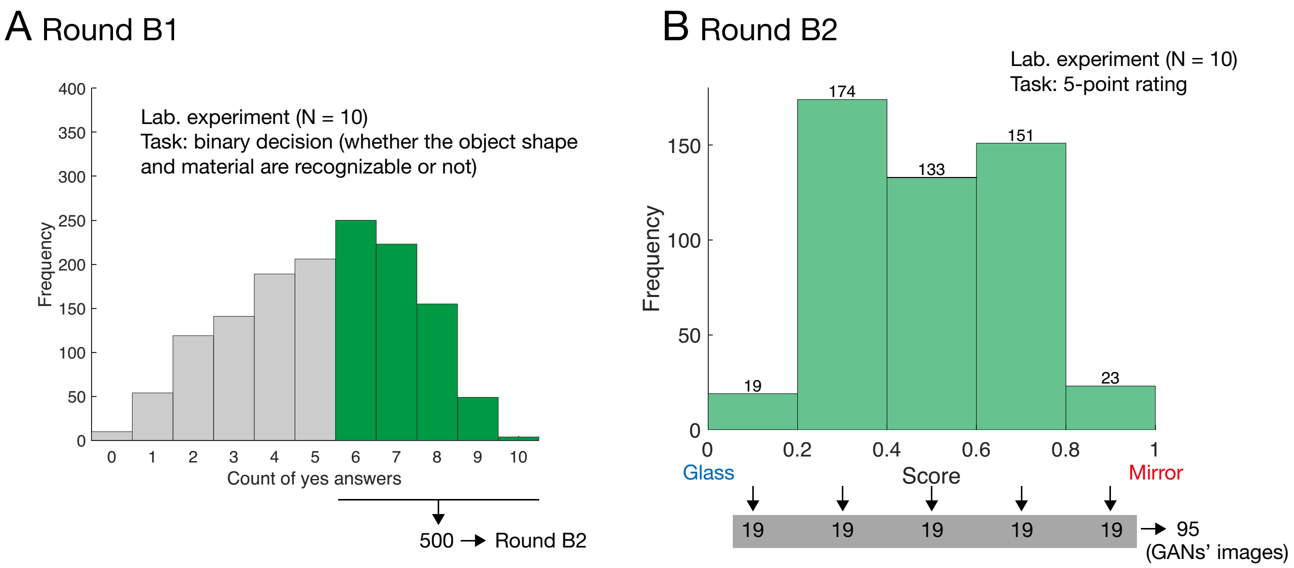


Figure A.7: Results of Round B1-B2

Each panel shows result of each round as histogram of images. See also 2.4.

Appendix B

Static visual cues

B.1 Supplement figure

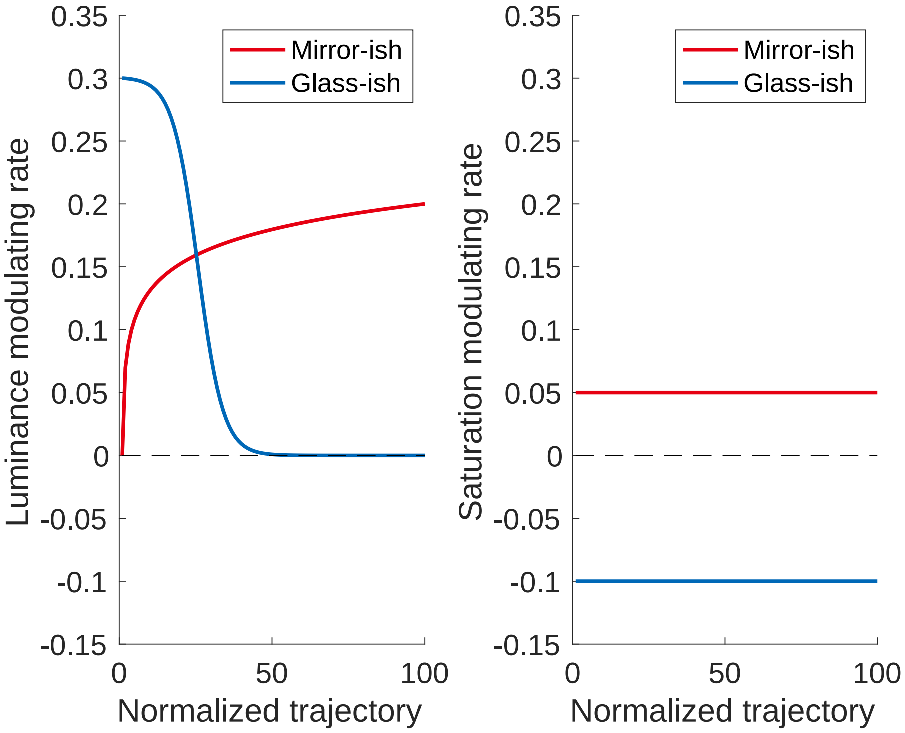


Figure B.1: Luminance and color saturation modulating rate

The horizontal axis indicates the normalized trajectory from the object contour (0) to its center (100). The vertical axis indicates the modulating rate of luminance and saturation (in the left and right panel, respectively). The zero means no change, and the positive and negative rates mean increasing and decreasing pixel values, respectively.

Appendix C

Dynamic visual cues

C.1 Optic flows in different materials

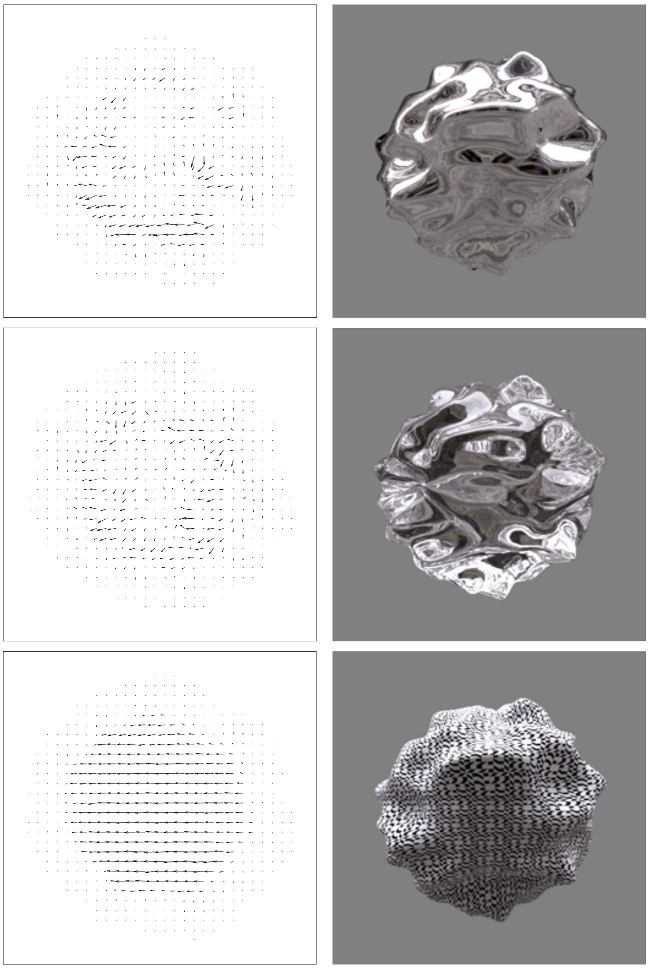


Figure C.1: Optic flows in different materials

The optic flows (in the left column) were generated from between the first frame (in the right column) and the second frames of the videos, which featured three different materials, i.e., mirror, glass, and matte (textured).

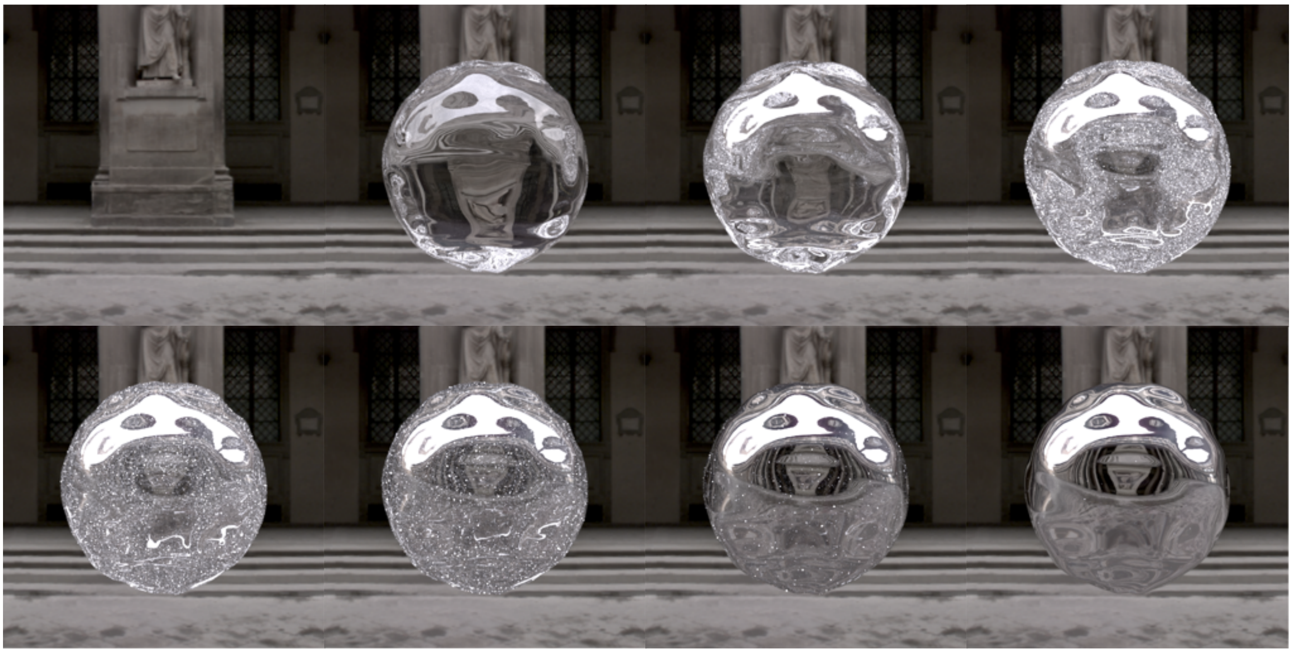


Figure C.2: Material appearance with changing refractive index

Eight glass objects with different refractive indices (1.0, 1.5, 2.5, 5.0, 7.5, 10, 20, and 50 from the top left to the bottom right) are shown.

C.2 Material appearance with changing refractive index

C.3 Supplementary experiment

Observers

Ten naïve observers participated in this experiment. One observer was excluded from analysis because their performance was almost equal to the chance level (50%) and quite different from the others. Thus, the final sample consisted of nine observers ranging in age from 23 to 26 years (average 24.2 ± 1.2 years).

Procedure and Task

The procedure was the same as in experiment 1, except the shuffle and the colour inversion conditions were included. The shuffle condition was set as the control condition. It was comprised of four frames extracted from the video presented in a random sequence, like a cut-off animation (see also Stimuli). All three presenting conditions (static, dynamic, and shuffle) were defined as 'natural'. An additional condition was defined as 'colour inversion' stimuli, in which colours were inverted so that positive values became negative in RGB colour space. This condition was created to measure the amount of information provided by static cues derived from the luminance polarity of the natural environment. Both natural and colour inversion stimuli were intermingled in one block. The experiment was composed of 720 trials (two materials × three shapes × five illuminations × two naturalness × 12 present conditions), and all trials were randomly ordered.

Results

Figure C.3 shows the performance for each condition. We performed two-way repeated-measures analysis of variance (ANOVA) for the presenting condition and the naturalness. The main effect of the presenting condition was significant ($F(2,\ 16) = 13.234$, $p < 0.001$), and the performance under the dynamic condition was the highest for all conditions (multiple comparison test; dynamic condition vs. static condition ($p < 0.005$); dynamic condition vs. shuffle condition ($p < 0.05$)). Not only image information from various viewpoints but also the consecutive images from motion provide us dynamic cues, which supports our hypothesis. The main effect of naturalness was significant ($F(1,\ 8) = 29.001$, $p < 0.005$) and indicated that the colour inversion decreased the performance of perceptual material discrimination. This is the same tendency as the rotating condition (Figure 4.2A). There was no significant interaction ($F(2,\ 16) = 0.348$, $p = 0.712$). The performance for the colour inversion stimuli was significantly lower than that for the natural stimuli, similar to the upside-down stimuli. These results suggest the luminance polarity of the natural illumination also contributes to the static cue for perceptual material discrimination.

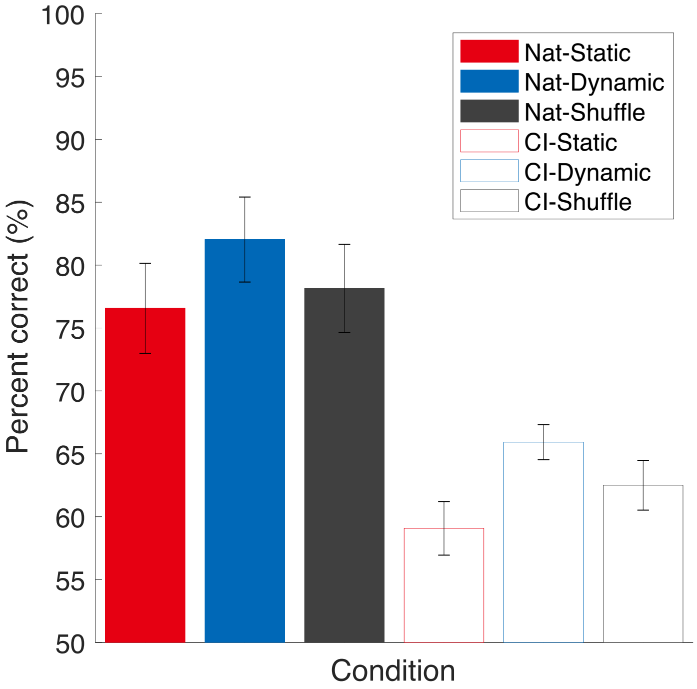


Figure C.3: Perceptual material discrimination with the shuffle and colour inversion conditions

The horizontal axis indicates each condition combined with the naturalness and presenting condition. 'Nat' signifies natural illumination, and 'CI' signifies colour-inverted illumination. The vertical axis indicates the percentage of correct answers. Averages and standard errors among observers were obtained. The error bars represent the standard error of the mean across all nine observers.

C.4 Image morphing method

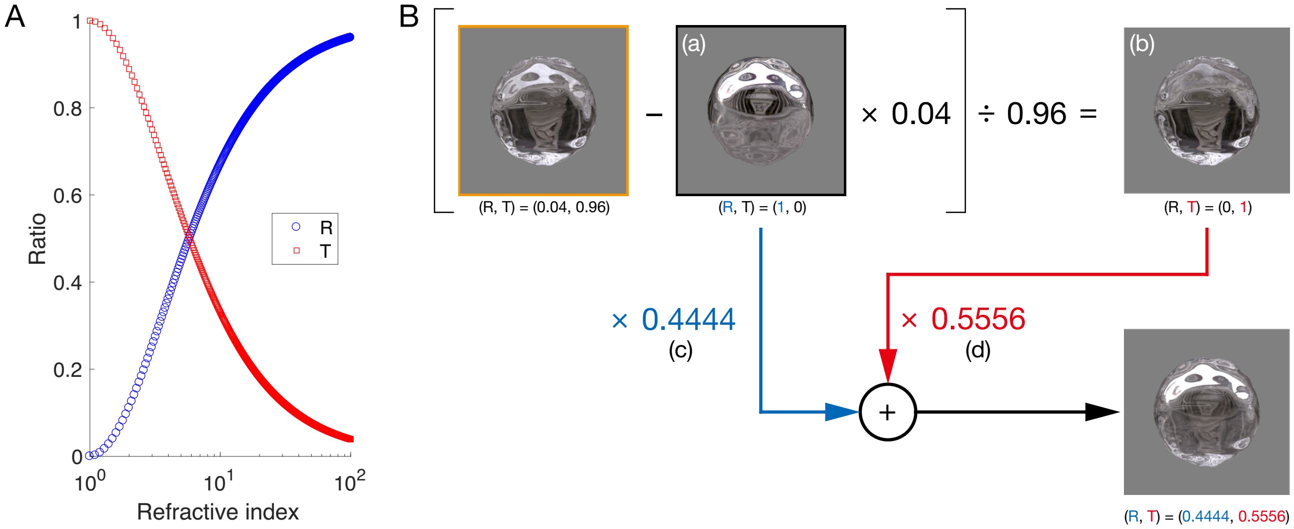


Figure C.4: Image morphing method

(A) Characteristics of reflectance and transmittance versus refractive index. The horizontal axis indicates the refractive index. The vertical axis indicates the ratio from 0 to 1. (B) Description of image morphing method. This example shows how to create an image with an arbitrary refractive index of 5. In this case, the reflectance and transmittance are 0.4444 and 0.5556, respectively. The components from (a) to (d) are consistent with the main text.

C.5 Movie

Movie 5.1

https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-26720-x/MediaObjects/41598_2018_26720_MOESM1_ESM.mp4

Movie 5.2

https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-26720-x/MediaObjects/41598_2018_26720_MOESM2_ESM.mp4

Movie 5.3A

https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-26720-x/MediaObjects/41598_2018_26720_MOESM3_ESM.mp4

Movie 5.3B

https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-26720-x/MediaObjects/41598_2018_26720_MOESM4_ESM.mp4

Movie 5.3C

https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-26720-x/MediaObjects/41598_2018_26720_MOESM5_ESM.mp4

Movie 5.4A

https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-26720-x/MediaObjects/41598_2018_26720_MOESM6_ESM.mp4

Movie 5.4B

https://static-content.springer.com/esm/art%3A10.1038%2Fs41598-018-26720-x/MediaObjects/41598_2018_26720_MOESM7_ESM.mp4

Appendix D

The glare illusion

D.1 Brightness estimation with reference to a black annulus

The method was the same as that used in Experiment 1 (see 5.2.4), except for the reference stimulus. In this experiment, the reference had a uniform black (0.56 cd/m^2) annulus instead of gray (44% of the center region). Samples included Glow, Halo, Unif44, and Unif0. Unif44 was the same as Unif in Experiment 1. Unif0 was the control condition, in which the sample and reference were the same. Five subjects who also took part in the main experiment were the participants in this additional experiment. We observed robust and significant brightness enhancement in the Glow condition across the luminance range of the samples ($p < 0.05$, binomial test; see Figure D.1). Unif44 was darker than the control, corresponding to the effect of simultaneous contrast.

D.2 Effect of viewing angle

The method was the same as that used in Experiment 2 (see 5.2.4), with the exception of the eye control. Seven observers participated in two different view angle sessions. Firstly, in the peripheral observation session, the sample stimulus was presented on the peripheral visual field while the observers fixated on the fixation point on the screen, as in Experiment 1 in the main text. Secondly, in the foveal observation session, no fixation point was presented and the observers were asked to gaze at the center patch of the sample stimulus. The stimulus was displayed until a response was obtained. The sample stimuli were the same (Glow, Halo, and Uniform). There were six luminance levels: 0, 40, 80, 120, 160, and 200 cd/m2. Each session was composed of 144 trials (4 trials × 2 positions × 6 luminance levels × 3 profiles), and all conditions were randomly intermingled. The results are shown in Figure D.2. The categorical responses were almost the same across viewing conditions, including the original condition of the main experiment. If differences existed, the thresholds of the original condition were slightly lower. This lowering due to eye control could help to support the main conclusion that brightness is enhanced even when its category is gray. Thus, we conclude that the difference of eye position is not critical for the categorical responses.

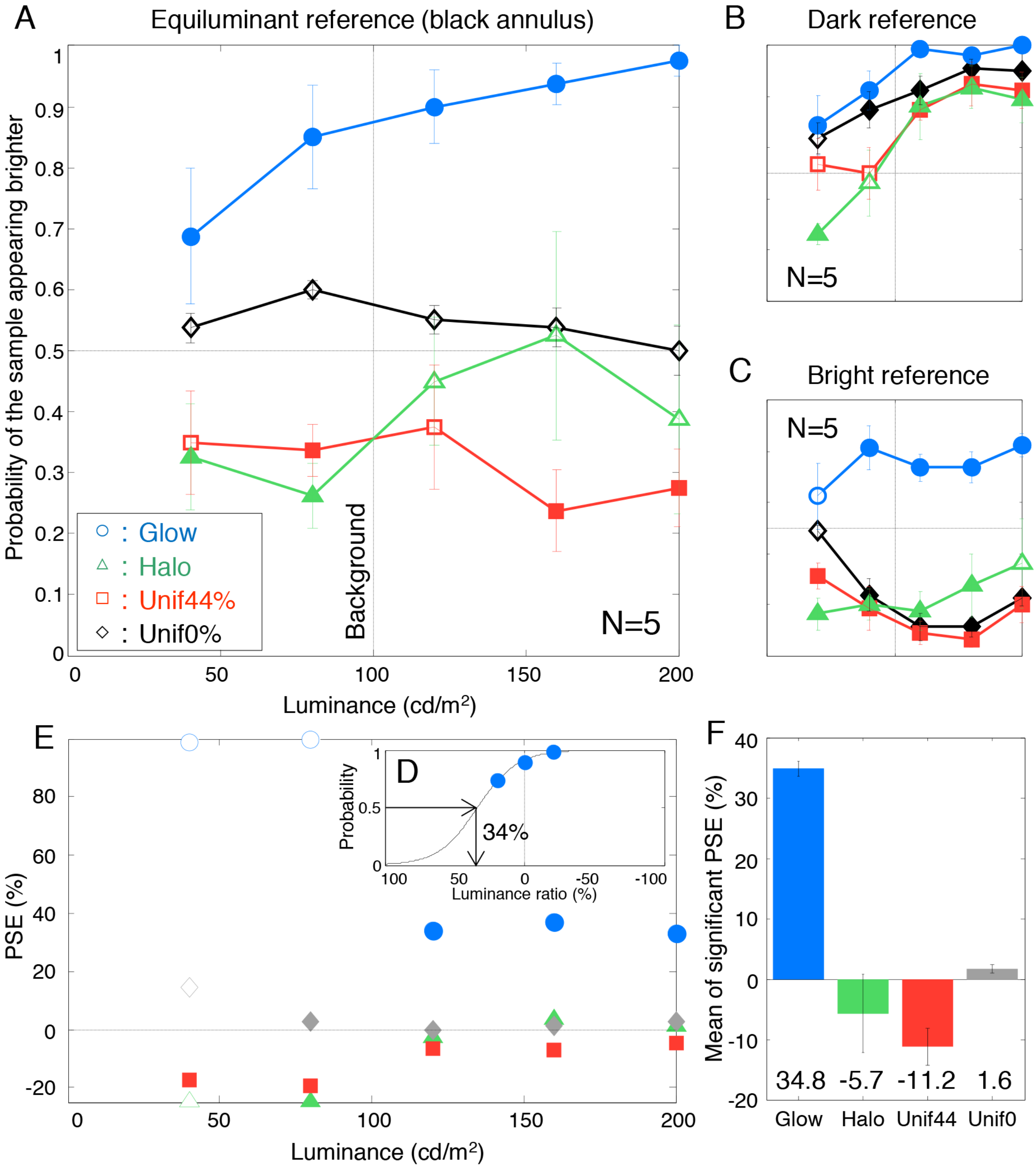


Figure D.1: Brightness enhancement in the illusion in an additional experiment

The format is the same as Figure 5.2 in the main text, except for the two uniform conditions.

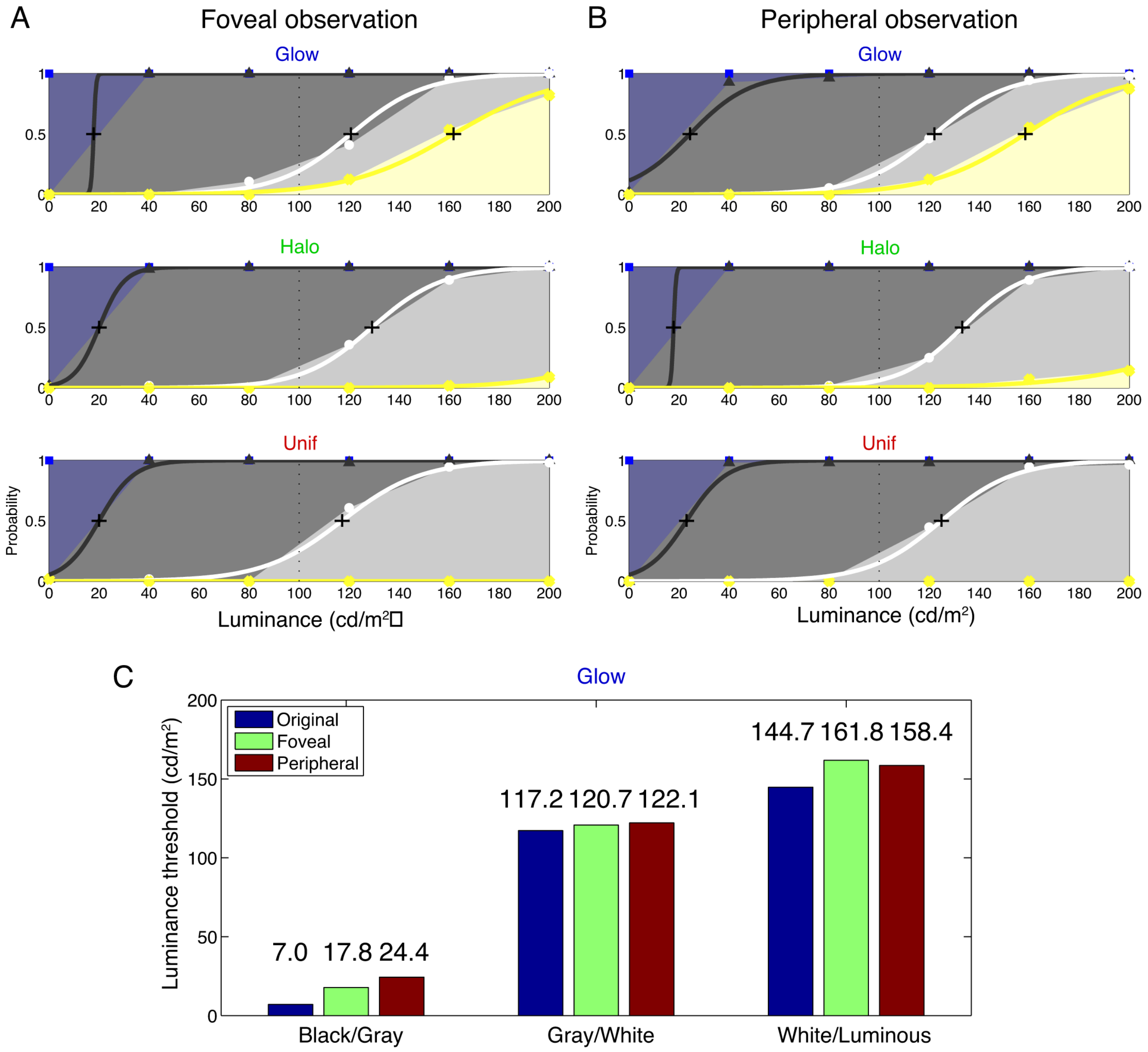


Figure D.2: The results of Additional experiment 2

(A) Categorical response probability in the foveal observation. (B) Those in the peripheral observation. The formats are the same as Figure 5.3 in the main text. (C) Luminance thresholds in Glow condition. Horizontal axis indicates three luminance thresholds: Black/Gray, Gray/White, and White/Luminous from the left. Each bar indicates a different view condition: Original (same as Experiment 2, free viewing), Foveal (foveal observation), and Peripheral (peripheral observation).

Appendix E

The rotating glass illusion

E.1 Additional experiment

When information on the top and bottom parts and their edges of the object are visible, the illusion does not appear (as in control condition; see Figure E.1). This suggests that the parts and edges of the top and bottom of the object help us to perceive the entire object shape and it leads to an accurate perception of material and direction. If the parts and edges (enough information to detect entire object shape) are not visible, the illusion appears even when the object has more homogenous bumpiness.

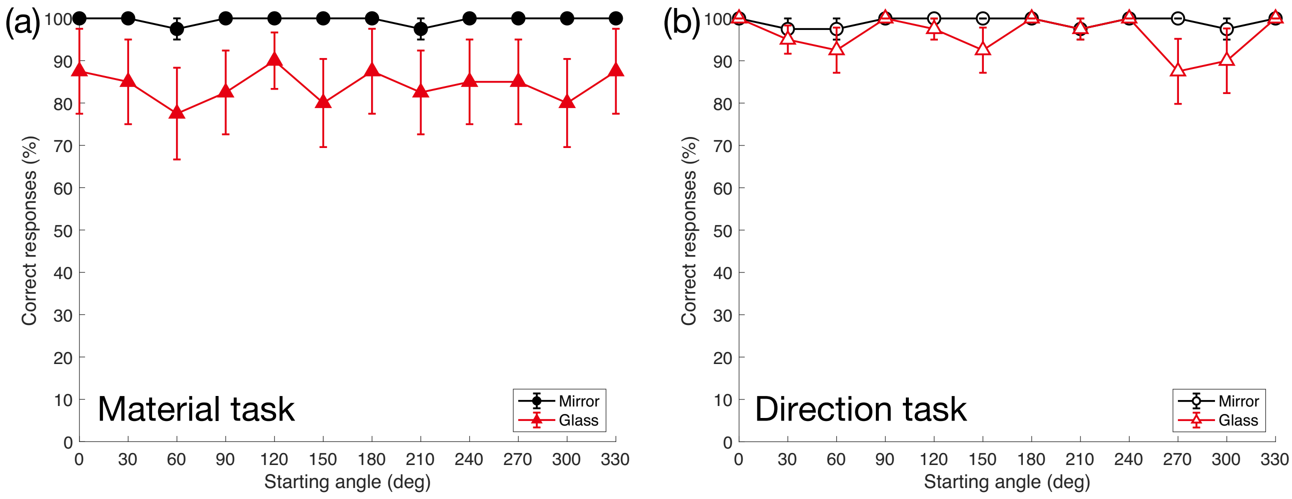


Figure E.1: Control condition

Control condition. Results of the material and the direction task in (a) and (b), respectively. Ten observers participated the experiment and the format is the same as that of Figures 6.1 (c) and (d).

E.2 Movie

Movie 6.1

https://players.brightcove.net/4988507115001/BJ5hvqqbQ_default/index.html?videoId=ref:sj-vid-1-ipe-10.1177_2041669518816716