Neural Network Models of Human Gloss Perception

Dissertation zur Erlangung des Doktorgrades der Naturwissenschaften der

Justus-Liebig-Universität Gießen
Fachbereich 06 – Psychologie & Sportwissenschaften

Vorgelegt von Konrad Eugen Prokott
am 22.12.2021

Erstbetreuer:
Prof. Dr. Roland W. Fleming (Justus-Liebig-Universität Gießen)
Zweitbetreuer:
Prof. Dr. Karl R. Gegenfurtner (Justus-Liebig-Universität Gießen)

ACKNOWLEDGEMENTS

I want to thank my supervisor Roland Fleming for his effort, time, advice, patience, encouragement, and general support over these last five years and before then since my bachelor thesis. Thank you also to my second supervisor, Karl Gegenfurtner. Thank you, Hideki for your collaboration and discussions during the second chapter.

I also want to thank the other members of the Fleming Lab, past and present for many helpful discussions and questions and for making it fun to be a part of this lab. I especially want to thank Kate, Rob, Yaniv, Filipp, JanJaap, Vivian, Lina and anyone else who I came to for advice or tech support. Thank you also to Sandra, Frau Aichner and Anouk for taking care of paperwork.

I want to thank my family and friends for their support, encouragement, food, open ears and babysitting. Special thanks to Isabel for finding that one error in my code that one time. (I still think that was a lucky guess). Also thank you for all the rest.

ABSTRACT

The human visual system can identify materials and their properties at a quick glance. One property that contributes to a material's appearance is gloss – the physical tendency of a surface to reflect light in a single direction, causing the appearance of sharp distinct reflections. The glossiness of a material allows us to infer further material properties, such as the freshness of food, the cleanliness of a surface, and whether the floor is wet and slippery. Despite its importance human perception of gloss remains poorly understood.

Artificial neural networks have been very successful over the last decade as powerful tools in computer vision and have garnered substantial interest from human vision science. Using a convolutional architecture and machine learning these models can extract useful features in large datasets containing millions of images in order to recognize objects, materials, faces and for many other tasks in computer vision. However, despite their power and versatility, or perhaps because of these properties, the applications and interpretations of neural networks in human vision science remain challenging and complex and new methods are still emerging. This thesis has two aims - to investigate human perception of gloss, and to investigate the application and usefulness of artificial neural networks in human vision science.

The first study explores different feed-forward architectures of convolutional neural networks (CNNs) to replicate human responses in discriminating between high gloss and low gloss textured materials. We found that CNNs of different depths may correlate well with human responses, but that networks with 3-5 layers most typically tend to respond similarly to humans. We also trained Deep Convolutional Generative Adversarial Networks (DCGANs) of different depths to recreate images showing low- and high-gloss materials and showed these to human observers. Observers were able to tell apart low- from high-gloss materials in images created by DCGANs with two or more layers, while DCGANs with three layers produced images that were as discriminable as renderings. Our findings show that CNNs of relatively shallow depths can replicate successes and failures of gloss classification that are typical for humans and can generate images that convincingly depict glossy materials.

In the second study we investigated human perception of gloss highlights – sharp and bright reflections on a glossy surface. Humans classified individual pixels in grayscale images

of textured glossy surfaces as containing a highlight or surface texture. We trained a neural network to identify pixels containing highlights in a large set of such images. In a second fitting stage we pruned the network to find a subnetwork that responds more like humans. We found that we can indeed find pruned configurations that explain the human data better than the full network. Further analysis suggests that representations in the network mostly resemble simple directly image computable features, while only some show similarity to certain complex geometric features. The network appears to be only weakly sensitive to violations of photo-geometric constraints.

Taken together our findings support a view of gloss perception as a process of mid-level vision. We find that relatively shallow neural network architectures of 3-5 layers are sufficient to model human gloss perception. We also find that our model for highlight detection learned features and representations that resemble both complex geometric predictors as well as simpler directly image computable features. In both projects our most human-like models appear to be only weakly sensitive to photo-geometric constraints.

INDEX

Chapter 1:

INTRODUCTION

1.1 Gloss Perception

Visual perception of materials is an everyday part of our lives. At a glance we can get an impression of what material an object is made of (Sharan, Rosenholtz & Adelson, 2014), a useful prerequisite for interacting with our environment. Adelson (2001) pointed out the importance of material perception in addition to previously prevailing research into object perception. One important attribute of material perception is the perception of surface gloss. It allows us to tell whether a piece of floor is slippery or wet, to determine whether a piece of food is fresh, raw, or cooked, and how a surface has been processed (for example polished or sanded). Gloss is also one of several factors that contribute to the characteristic appearance of material categories (Fleming, Wiebel & Gegenfurtner, 2013). Despite how easily and often correctly our visual system can determine the glossiness of a material, this is no trivial task, and we still have a poor understanding of the computations the human brain performs to perceive gloss.

1.1.1 Physical descriptions of gloss

Gloss is an aspect of material appearance that is caused by a material's tendency to display sharp *specular* reflections of light falling on the surface. While gloss is associated with certain physical reflectance properties of materials, gloss itself is a perceptual, visual property that cannot be directly measured. Some of the reflectance related terms I use in this thesis have specific meaning when used to describe the physical reflectance properties of materials. Here I will use these terms more loosely to describe the image characteristics relevant for vision that are associated with different physical reflection events. Broadly speaking perceived gloss is related to the amount of light reflected specularly. An ideal mirror material reflects incoming light in one direction only, causing specular reflections that appear to shift over the surface when the viewing angle is changed. In contrast to this, an ideal matte material reflects light *diffusely* in such a way as to display the same luminance at a given point of the surface regardless of the viewing angle. See **Figure 1**. Luminance variations on a matte surface appear as shading.

Most materials are neither perfect mirrors nor perfectly matte but reflect some light diffusely and some specularly. We can simulate many materials as combinations of these two components (e.g. Ward, 1992). However, measuring and describing the physical reflectance of naturally occurring materials is more complicated than that. The full reflectance properties of a material can be described by the Bidirectional Reflectance Distribution Function (BRDF) (Nicodemus, 1965). This function specifies how for any angle of light falling onto a surface, light is reflected into any direction. Both the incident and viewing angles are described for any point on a hemisphere above a surface and may be described by two angles each (azimuth and zenith), making the BRDF a 4 – dimensional function. The angular resolution or density of measurements can be arbitrarily high.

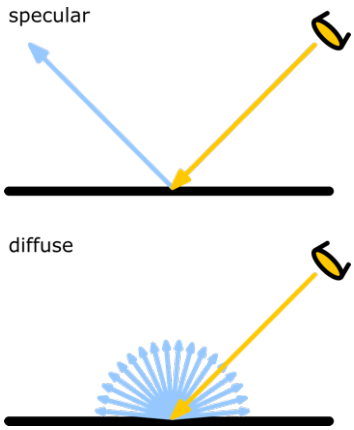

Figure 1: *Glossy materials* ***(top)*** *reflect light focused into one direction. Matte materials* ***(bottom)*** *reflect light diffusely, displaying the same luminance in all directions.*

1.1.2 Research directions in gloss perception

Humans do not perceive gloss in the same detail as a BRDF, although the exact dimensionality of human gloss perception has been the subject of a long debate and much research (see Chadwick & Kentridge, 2015 for a review of gloss perception research). One focus of studies of gloss perception has been to link physical measurements and descriptions of surface reflectance to human perception, including the study of perceptual dimensionality. This has been of interest for industrial applications (for example to assess the appearance of paper or paints; Billmeyer & O'Donnell, 1987; Harrison, 1949; Hunter, 1937; Ingersoll, 1921; Leloup, Hanselaer, Pointer, & Dutré, 2012) and more recently for applications in material design for computer graphics (Matusik, Pfister, Brand, & McMillan, 2003; Pellacini, Ferwerda, & Greenberg, 2000; Serrano, Gutierrez, Myszkowski, Seidel, & Masia, 2016; Wills, Agarwal, Kriegman, & Belongie, 2009).

A second focus of research into gloss perception that has gained attention in the last 25 years has been to identify which features of an image or a scene cause the impression of gloss and to better understand the mechanisms by which these are evaluated in the human brain. (Adelson, 2001; Nishida & Shinya, 1998). This is a computational challenge since the glossiness of a material can never be directly observed, only in the context of an object or surface shape that it takes and the environment it is seen in. A material can take on a wide range of shapes and, in the context of different environments (illuminations) and viewing angles, an even larger range of appearances (See **Figure 2**; Fleming et al. 2003). One of the fundamental questions in the study of visual material perception is how the human visual system infers material properties such as gloss with a high degree of consistency from the vast range of images they can cause under different viewing conditions. This thesis aims to better understand the mechanisms and features in static, monocular images used in human gloss perception.

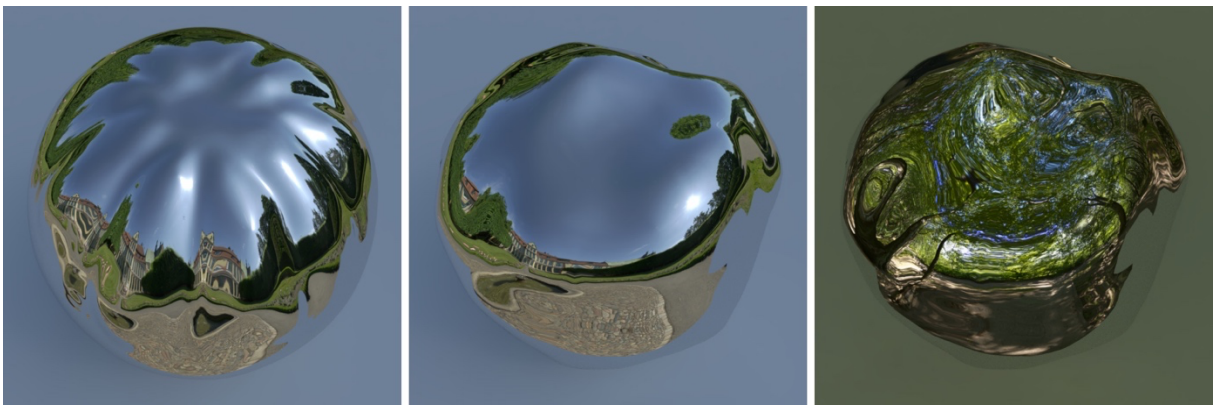

Figure 2: *A glossy material can take on a wide range of appearances due to changes in shapes (center and left) or illumination (center and right), yet we can still perceive it as the same material. Figure inspired by Fleming et al. (2003)*

1.1.3 Theoretical mechanisms of gloss perception - inverse optics

A seminal theory of vision by Barrow and Tenenbaum (1978) proposed that early stages of visual processing decompose a retinal image into causal attributes, such as shape, shading, color and materials. This theory is known as *intrinsic image decomposition*. In the context of this theory, reflectance is a physical property of a material that interacts with illumination and the shape of a surface to cause the overall appearance of an object we may observe. The idea of intrinsic image decomposition into veridical components is associated with the theory of *inverse optics* – the idea that the human visual system explicitly calculates the physical processes by which illumination, material and shape interact in reverse, to arrive at a physically accurate representation of the intrinsic factors of the object and the environment. This view

implies that the visual system has an explicit understanding of the physical interactions between the intrinsic components of a scene.

There exist several arguments and experimental observations against an inverse optics approach. The most fundamental argument is that any 2D view of a scene confronts the visual system with a loss of information and an unresolvable ambiguity as to its underlying causality. As stated earlier, a material can only be observed in the context of a shape and the illuminating environment. Calculating one of these intrinsic components without knowing the others is not possible, yet neither of the components can be known from looking at a single image. Without this knowledge there are multiple possible solutions to any retinal image. For example, an apparent reflection in a surface could be due to an actual reflection or it could be due to an image painted onto a matte surface. Similarly, distortions in the retinal image of an object could be due to the object's shape, irregularities in a reflected environment, or a distortion in a texture painted onto the object. A seemingly matte surface could either be truly matte or illuminated by a very blurry or foggy environment (Pont & te Pas, 2006). From one single view of an object, the contributing factors cannot be told apart. Barrow and Tenenbaum (1978) already pointed this out and suggested that the visual system uses a set of assumptions about the physical world that act as constraints. Examples of such assumptions could be that we rarely or never encounter completely diffuse illuminations, making a diffuse object the more plausible interpretation. A surface convincingly painted to imitate reflections may be created by a skillful artist but also occurs only rarely and needs to be viewed from a specific angle. The idea that the visual system uses knowledge from prior experiences to make plausible inferences of distal scene variables from an ambiguous visual input can be traced back to at least von Helmholtz (1866; as cited by Gregory, 1997). While many authors acknowledge the infeasibility of inverse optics, it remains an important theory (eg. Marlow & Anderson, 2015).

1.1.4 Informative misperceptions

Another key argument against an inverse optics theory of gloss perception is that humans display certain characteristic failures of gloss constancy. Gloss constancy is the theoretical ability to perceive the reflectance of a material unvaryingly under varying viewing conditions. As stated earlier, a veridical decomposition according to inverse optics would require the visual system to disentangle an object's shape, material, and illumination. Over the past two decades several studies have shown that the shape of an object and the illumination it is viewed under can influence perceived gloss, indicating that gloss perception falls short of such a veridical decomposition of the intrinsic factors.

Regarding the entanglement of gloss and illumination, (Pont & te Pas, 2006) showed spheres of systematically varied materials under different illuminations to observers and found that humans have difficulties identifying identical materials under varying illumination ('material constancy') and identifying identical illumination conditions when applied to different materials ('illumination constancy'). The authors conclude that there is confusion between perception of reflectance and illumination. (Fleming et al., 2003) presented subjects with two spheres of varying gloss rendered under different illuminations and asked subjects to adjust the glossiness of one of the spheres to match the other. They found that subjects were able to perform the match more accurately when the test spheres were presented under natural illumination than under artificial illuminations such as point light sources or noise illumination maps. They also demonstrate that the context has little effect on the perception of gloss by cropping spheres rendered under naturalistic light maps and exchanging backgrounds. The authors conclude that the statistics of natural illuminations cause specific image features that are diagnostic of the reflectance properties of a material they are reflected in. However, they also acknowledge that while the complex cues provided by naturalistic illuminations increase realism and the ability to accurately match gloss, they are not necessary to achieve a perceptual impression of gloss. Olkkonen and Brainard (2010) used a similar experimental paradigm and let participants match the intensity of the diffuse and specular components of spheres rendered under natural illuminations to match a target sphere. The authors found no evidence for gloss constancy, and the results of the study show that the diffuse intensity (albedo or brightness) is influenced differently and independently from specular intensity. Doerschner, Boyaci, and Maloney (2010) investigated the transfer function, describing how perceived gloss translates from one illumination to another. Observers were asked to make pairwise comparisons between spheres of different gloss rendered under two different illumination maps. From this, the authors estimated the transfer functions between two pairs out of three illumination maps and showed that they could calculate the third transfer function from the first two. This transitivity of gloss transfer functions is an important finding, because it shows that perceptual differences in gloss under different illuminations are not random, but rather due to some lawful mechanism. The authors interpret this result as an indication that different cues present under different illuminations are integrated according to a fixed rule rather than a varying or even conflicting interaction of different cues. Motoyoshi and Matoba (2012) used a material matching task to investigate the influence of systematic manipulations of the illumination by changing mean intensity, gamma and contrast of natural light fields. They find that the mean intensity has little

influence, while gamma and contrast both have large effects on the matched material reflectances.

Taken together, there is evidence that the human visual system is tuned to detect gloss cues that are caused by natural illuminations. While variations in illumination have an effect on perceived gloss, it appears that the perception and integration of image features caused by these variations follow fixed rules.

Regarding interactions between surface shape and perceived gloss Vangorp, Laurijssen and Dutré (2007) found that the shape of objects influences the discriminability of materials with different complex BRDFs. In this study subjects were asked to compare the material of two objects in an image under the same naturalistic illumination and respond whether the material is identical. The authors found that subjects were less accurate when the two objects had different shapes. In comparisons of two objects of the same shape, the shape made a difference to the accuracy. Comparisons between materials of tessellated spheres (i.e. polyhedral with flat surfaces) were less accurate than for other shapes, emphasizing the importance of curvature for accurate gloss perception. Spheres showed a comparatively high threshold for difference detection, causing the authors to question the suitability of spheres as stimuli in other material studies. The study showed no one best geometry for material discrimination. Ho, Landy, and Maloney (2008) showed rendered images of bumpy glossy surfaces with varying levels of bumpiness and gloss to observers, asking them to rate both properties. The authors found an additive interaction between the two perceptual properties – surfaces with higher relief appeared more glossy and glossier surfaces were rated to be bumpier. Marlow, Todorović, and Anderson (2015) and Marlow and Anderson (2015) demonstrated how depth and shape cues can cause the visual system to interpret the same intensity gradient as specular or matte reflections. In their experiments the perceived 3D shape was manipulated by either cropping a gradient to different contours or by adding stereoscopic depth cues.

There is also evidence for an interaction between the influences of illumination and shape on perceived gloss. Wijntjes and Pont (2010) found that renderings of perturbed matte surfaces can appear glossy when the 3D surface relief is deeper, and they are illuminated frontally. Olkkonen & Brainard, (2011) also found that shape and illumination changes interact to influence gloss perception, and that these effects could not be easily explained by the individual effects of each factor.

Together these findings suggest that the visual system does not disentangle shape, illumination, and material veridically and entirely independently from another. Failures in gloss constancy indicate that there are remaining interactions between perceived shape, gloss and

illumination, as physical changes in one component can influence perception of the others. These remaining interactions may be due to heuristics that the visual system uses based on statistical regularities in the environment. However, the nature of these heuristics is not entirely clear. They could be assumptions about physical factors that are used as constraints to approximate a solution of inverse optics. Another possibility is that these heuristics concern the directly visible appearance of glossy materials. More recent theories have focused on the latter possibility, proposing that the visual system relies on regularities in images. These theories are described in the following sections.

1.1.5 Theoretical mechanisms of gloss perception - image statistics

One alternative to inverse optics that has been proposed is the theory that the visual system uses *image statistics* to arrive at estimates of reflectance properties of scenes. These image statistics are regularities in 2D images that are used as a proxy of physical reflectance. Motoyoshi et al. (2007) found that a positive skew of the intensity distributions of surfaces is positively correlated to human perception of glossiness and negatively correlated with perceived surface lightness or albedo. Manipulating the intensity distribution skewness of images of various material surfaces shifted subjects' perception of gloss and lightness in these expected directions. They proposed a neural mechanism for detecting local skewness and support their hypothesis by demonstrating an adaptation effect. Subjects reported perceiving an image of a stucco surface as more glossy and darker after adapting to negatively skewed images, and as less glossy and lighter after adapting to positively skewed images. Boyadzhiev, Bala, Paris and Adelson (2015) have offered a similar explanation of gloss perception through the distributions in specific frequency subbands, demonstrating how material properties can be changed by deliberately manipulating subband distributions. Sawayama & Nishida (2018) also demonstrated a connection between intensity distributions and material perception. They arrive at the conclusion that shape perception is based on the order of intensities, while reflectance perception uses the magnitude of intensity gradients. Image statistics have also been shown to contain useful information for other aspects of material perception, such as wetness (Sawayama, Adelson, & Nishida, 2017) or subresolution density of textures (Sawayama et al., 2017; see also Nishida, 2019 for a review of image statistics in material perception).

Anderson and Kim (2009), in response to Motoyoshi et al. (2007) have shown that intensity distributions alone without fulfilling very specific constraints of the relationship between shading and specular components of a surface, are not enough to cause an appearance of gloss. The authors showed that shifting the specular and matte components of an image of a glossy surface so that the highlights are no longer aligned with the diffuse shading decreases

the impression of glossiness – more so with increasing angular offset or distance of a translational offset – while leaving intensity distributions unchanged. In another demonstration they created a positively skewed image by inverting pixel intensities of an image of a matte surface, which also did not make the image look glossy. Kim and Anderson (2010) replicated the aftereffects described by Motoyoshi et al. (2007) but also demonstrated that these aftereffects occur after adaptation with different zero-skew adapters, indicating that they are not due to a neural skewness detection mechanism.

The results of Anderson and Kim (2009) suggest that gloss perception involves more specific features than the overall luminance distribution in an image. The spatial arrangement of highlights with respect to the surrounding shading gradients seem to play an important role. This could be related to higher-level regularities in image features, such as the orientations and alignment of highlights and gradients as are proposed to be involved in mid-level vision.

1.1.6 Theoretical mechanisms of gloss perception - statistical appearance models

A third approach to gloss perception that has emerged recently is what Fleming (2014) called *statistical appearance models*. According to this view the visual system does not assess glossiness directly, but rather measures qualities and features of an image that are directly observable and that are correlated with glossiness under typical viewing conditions (Fleming, 2012 referring to results from Marlow, Kim, & Anderson, 2012). In this case gloss to the visual system would not be represented as a physical property. Instead, perceptual gloss is a range of appearances that are typical for glossy materials and can be detected and described by frequently co-occurring image features. In other words, the visual system does not try to infer the distal properties of how a material reflects light, but only the proximal attributes of how a material looks in a given image. A statistical appearance approach would not seek to explain gloss perception and failures of gloss constancy as described in the previous section in connection with underlying physical factors, but in connection with what is there and visible in an image. In contrast to an image statistics approach, the image features would not act as a proxy to some deeper understanding or representation of material reflectance.

According to Fleming (2014), the visual system would first identify candidate features or elements in an image, such as glossy highlights. The perception of a specific material would arise from similarities and shared visual properties of such elements in an image, that tend to co-vary systematically between different materials. This could be for example the contrast, size or blurriness of specular highlights.

Evidence for a statistical appearance view of gloss perception comes from Marlow et al., (2012), who provide an explanation of confounding effects of illumination and shape on

gloss perception in terms of image properties of highlights. Participants viewed bumpy surfaces of different relief heights under different illuminations, showing different effects of illumination and relief height on perceived gloss. In a second experiment they asked participants to rate the sharpness, contrast, and coverage of specular reflections in a subset of the same images (as well as depth in a parallel version of the stimuli with stereoscopic disparity). A linear combination of these cues could predict the effects of relief height on perceived gloss under different illumination conditions. Contrast, sharpness and coverage had the same relative weights in conditions with or without binocular disparity. Histogram skew was included in simulations as a possible cue but was not part of the optimal model. Marlow and Anderson (2013) expanded these findings, deliberately manipulating the proposed image cues through changes in illumination, shapes and viewing angle and again finding perceived gloss to be well explained by a combination of these image cues. Another important finding from this study is that the relative contributions of these image cues may vary for different stimuli. Leloup et al. (2012) found evidence that several cues (the sharpness of reflections and the difference in brightness) are used by human observers in perceiving and comparing real-world material samples. The authors also observe that participants show different strategies, placing more emphasis on either one of these cues. The observation that different strategies are used by different observers was also made by Sève, (1993; as cited by Chadwick & Kentridge, 2015). van Assen, Wijntjes and Pont (2016) investigated the influence of highlight shapes in photographs and real-world examples on perceived gloss and found that samples with simple highlight shapes, such as a square or disk were perceived as most glossy.

Statistical appearance models provide a possible explanation of human failures of gloss constancy in terms of several easily accessible image properties of highlights. Interactions between shape, illumination and perceived gloss can be explained by the way these different physical factors can affect the same image features. However (so far), statistical appearance models leave an important question unanswered: How does the visual system identify highlights?

1.1.7 Localizing glossy highlights

While the importance of highlights for gloss perception is widely acknowledged and plays a central role in statistical appearance models, it is not yet understood how the visual system identifies highlights and what makes them appear as highlights rather than bright texture markings. Some studies have identified conditions of the context and placement of highlights that influence whether they are perceived as highlights and contribute to a perception of surface gloss. Beck and Prazdny (1981) first observed the importance of highlights, noting that the

presence of both an intensity gradient (causing an impression of surface curvature) and highlights causes a surface to appear glossy. The authors find that consistency between highlights and shading gradients is not necessary but does increase the perception of gloss. Todd, Norman and Mingolla (2004) also observed the importance of consistency between highlights and shading, noting that elongated highlights need to be aligned in the direction of minimum surface curvature in order to produce a perception of a glossy surface. Anderson and Kim (2009) showed that offsetting the specular highlight component with respect to the shading component of a surface through rotation or translation both reduced perceived gloss, emphasizing the importance of congruence between diffuse and specular components. In a more detailed investigation, Marlow, Kim and Anderson (2011) manipulated individual highlight patches in images of glossy surfaces by translating or rotating them. They found that perceived glossiness depends on rotational congruence between highlight patches and the surrounding gradient as well as congruence of location between highlight patches and the brightest spot in the gradient.

These findings underscore that specific conditions of the spatial arrangement and orientations of bright spots on a shaded surface need to be fulfilled for bright patches to appear as highlights. However, the findings listed above could be explained in two different ways. One possible explanation is in terms of congruence between highlights and the shape of the surface. These conditions are referred to as photo-geometric constraints. Alternatively, the observations could be explained in terms of congruence between highlights and visual intensity gradients caused by the shading on the surface.

This second explanation would mean that gloss perception can possibly be explained entirely by mid-level visual processes. Mid-level vision concerns the grouping of elements in an image according to covariations of low-level image features, such as edge orientations, gradients and colors. This grouping enables the organisation of individual elements in an image into perceptions of coherent surfaces. According to Kubilius, Wagemans, and op de Beeck (2014) mid-level vision mainly uses processes for estimating similarities and pooling features to segment an image into surface-like representations of objects and backgrounds. Adelson (2000) points out that mid-level vision is defined only roughly, being somewhere between low-level vision (retinal processes) and high-level vision (involving semantic knowledge). Anderson (2020) describes mid-level vision as a hypothetical stage of pre-semantic visual processing that is the earliest stage where properties of the environment such as shapes and materials are explicitly represented. In this sense, the purpose and content proposed for mid-level vision are not entirely unlike representations suggested by an inverse optics approach. The

crucial difference lies in the assumed underlying computations in terms of complexity and degree of tacit physical understanding as well as the explicitness of representations of physical factors. Adelson (2001) proposed that similar concepts and features used for texture analysis might prove useful in modelling material perception. Sharan, Liu, Rosenholtz, & Adelson, 2013) demonstrated that a computer vision model using perceptually inspired low and mid-level features could rival human performance in material recognition. An explanation in terms of mid-level visual processes would indicate that the localization of highlights can be explained by image features as well, without the need for an explicit understanding of geometry or other physical factors. However, since shading gradients and geometry are closely connected this is not an easy distinction to make.

There is some evidence however, that gloss perception involves explicit 3D shape perception. Berzhanskaya, Swaminathan, Beck and Mingolla (2005) investigated the propagation of perceived gloss due to highlights and found that perceived gloss decreases with distance on a surface rather than distance on a presented image. Discontinuities in a surface caused by a gap or an occluding element further decrease the propagation of gloss. This provides further evidence of an influence of shape on gloss perception, but does not address the causal attributes for highlight perception. It has also been demonstrated that depth cues – either bounding contours (Marlow et al., 2015) or stereoscopic cues (Marlow & Anderson, 2015) – can cause humans to perceive the same gradient as being caused by either a matte or metallic glossy material. At the same time the illumination direction appears differently. The authors take this as evidence that gloss perception does incorporate explicit calculation of shape and illumination. Localization of highlights and the underlying features is one of the main perceptual tasks I will investigate in this thesis.

In summary, there is evidence to suggest that human gloss perception is not veridical but rather based on heuristics. Finding and describing failures of gloss constancy - where the glossiness of materials is perceived differently when viewed under different conditions – has been useful for understanding the heuristics involved in gloss perception. These heuristics are closely related to the appearance of gloss under typical viewing conditions, and indeed, some researchers suggest that we do not perceive gloss, but rather typical appearance of gloss. However, there is a need for better understanding whether these heuristics concern physical factors – typical constellations of physical viewing conditions underlying the images we see – or purely visual factors – typical occurrences and regularities in the appearance of glossy

materials. Visual features have been shown to explain failures of gloss constancy. A particularly interesting question in this context is how the visual system identifies glossy highlights.

To investigate these mechanisms and features used for perceiving gloss we used a combination of response data from human subjects and artificial neural networks as image computable models of these tasks. These are a class of very powerful models in machine vision that have been shown to imitate certain aspects of human mid-level vision and can 'learn' features to solve specific visual tasks. From these we hoped to learn about the computations required to imitate human gloss perception in terms of complexity and computational requirements as well as the content of intermediate processing stages.

1.2 Machine vision with artificial neural networks

Artificial neural networks are a class of mathematical models in which information is processed by a number of artificial neurons (referred to as units) that perform a series of parallel computations (usually involving non-linear functions) in a specified order. Convolutional neural networks are a class of neural networks that have been very successful in computer vision tasks (eg. Krizhevsky, Sutskever, & Hinton, 2012; Simonyan & Zisserman, 2015; Szegedy et al., 2015). These networks make use of convolutional processing layers, which apply local image filters to every position in an image. This is a computationally very efficient way to scan an image for local features without specifying feature detectors for individual locations. Machine learning allows us to define the architecture of models and the type and order of computations taking place without specifying any of the parameters in the network. Parameters and internal weights are fitted during a training period in order to perform a specific task, the performance on which is operationalized by a loss function. Importantly for the aims of this thesis, modern neural networks represent trainable multi-layer models that contain hierarchically organized features, making them very promising candidates as image computable models for investigating mid-level processes of vision. This thesis focusses only on feed-forward networks - that is, networks in which the information is processed in one direction only, without lateral or feedback connections.

Over the last decade, artificial neural networks have gained a large surge of interest in machine vision but also from human vision science (Kietzmann, McClure, & Kriegeskorte, 2018). Since the introduction of AlexNet by Krizhevsky et al. (2012), deep convolutional neural networks (CNNs) have dominated competitions for computer vision, matching and recently exceeding human performance levels in image recognition (He, Zhang, Ren, & Sun, 2016; Szegedy et al., 2015; VanRullen, 2017). While much of the computational groundwork has

been laid over the past several decades (eg. LeCun et al., 1989; Rosenblatt, 1958), in recent years neural networks have been very succesfully applied to a range of different tasks in computer vision and other domains. Cox and Dean (2014) attribute this sudden gain in popularity more to developments in hardware technology than to recent theoretical discoveries, observing similarities between successful modern CNNs and networks that were developed 30 years ago. These technological developments have enabled faster computing, especially of parallel operations and the availability of large amounts of data (Krizhevsky et al., 2012). It is this training on large amounts of data from which we hoped to learn about features and regularities that are diagnostic of gloss over a large variety of images and appearances.

1.2.1 Supervised and unsupervised learning

Broadly speaking machine learning can be categorized into supervised and unsupervised learning. In supervised learning, networks are provided with labeled data and are trained to learn a mapping between input data and labels. Labels may contain various data and data types to be extracted from an input – for example categorical labels describing the object shown in an image (Russakovsky et al., 2015), continuous labels of physical properties shown (van Assen, Nishida, & Fleming, 2020), or detailed image descriptions such as pixel-wise labels for image segmentation (Bell, Upchurch, Snavely, & Bala, 2013), or saliency data obtained from human fixation behavior (Kümmerer, Wallis, & Bethge, 2018).

In unsupervised learning (also called self-supervised learning) there is no target label information and the target output is the same as the input or derived from the input. For example, autoencoders are trained to reproduce input images after compressing (encoding) them into representations with fewer parameters than the input and subsequently decoding them again (Hinton & Salakhutdinov, 2006). A successful autoencoder will accurately reproduce the image given as input, learning compression, representation and reconstruction that loses as little information of the input as possible. Another class of unsupervised networks that are being used in **chapter 2** are Deep Convolutional Generative Adversarial Networks (DCGANs) (Goodfellow et al., 2014; Radford, Metz, & Chintala, 2015). These consist of two networks that work against each other – a generator and a discriminator. A generator takes a low dimensional noise input and creates images. The discriminator is fed these images mixed with images from a target dataset. Both networks learn simultaneously – the discriminator learns to identify the generated images among the target images, and the generator learns to create images that are more likely to mislead the discriminator. The result is that the generator learns to create images that look as if they could be part of the target image set (according to the discriminator). It should be noted that DCGANs include a supervised element - the

discriminator is trained on explicitly labeled fake or real images, making use of supervised training. Radford et al. (2015) consider them as unsupervised networks, which could be argued, since no other data than the target images need to be provided to train the network. They could also be considered self-supervised, as the network generates the labels for training the discriminator by itself. An important difference to autoencoders is that the network does not replicate individual target images. It creates images that replicate certain shared higher-level statistics from the target image set, creating new images that – if training is successful – convincingly look as if they could be part of the target image set.

1.2.2 CNNs as models of human vision

While many aspects of modern neural networks are inspired by findings from neuroscience (Cox & Dean, 2014), they are heavily abstracted and simplified in favor of computability (Kriegeskorte, 2015). Some aspects, such as artificial neurons ('units') integrating input over a receptive field and activating in a non-linear function are loosely based on biology (Kietzmann et al., 2018). Other functions such as the back-propagation algorithm very commonly used for training neural networks (Rumelhart, Hinton, & Williams, 1986; Werbos, 1974, 1981) have little resemblance to biological mechanisms (Cox & Dean, 2014; Kietzmann et al., 2018) but have been useful in training multi-layer neural networks. The hierarchical processing of image information through a series of convolutional layers with non-linear activation has been compared to the human ventral visual stream (Kietzmann et al., 2018), but CNNs are not bound to biologically plausible architectures.

The rapid increase in computational power of modern neural networks has brought about a necessity and opportunity for new research approaches, and research designs and methods are rapidly emerging (Kriegeskorte, 2015). As opposed to 'hand engineered' machine vision models that are constructed based on known or proposed mechanisms and computations, we are now able to train highly performing complex models without predefining parameters, whose inner workings we do not fully understand. Parameters and features are 'learned' by fitting a model to perform a vision task on a (usually very large) dataset. While we have access to all weights and parameters in a network, reading these out and visualizing them is still a challenging task with many proposed approaches (Kietzmann et al., 2018). For example, Zeiler and Fergus (2014) proposed a method for approximately reversing the operations a CNN has performed to arrive at the activation of a certain unit. This visualization is performed in the context of a specific image and gives a visualization of a specific location and the content of the image that caused the activation. Famously, Google's DeepDream algorithm (Mordvintsev, Olah, & Tyka, 2015) is one of several approaches to use backpropagation to iteratively optimize

an image to maximize the activation of a desired unit caused by this image in a network. Such optimization approaches (see Olah, Mordvintsev & Schubert, 2017 for a review) work without the context of a specific image and can be initialized from a noise filled image. Other approaches have focused on finding images or regions in example images that activate a certain unit (for example by monitoring unit activation while systematically erasing parts of the image). In addition to the challenge of understanding internal representations of networks, the possible configurations of a single network architecture are vast, and several instances trained on the same task can learn very different features (Li, Yosinski, Clune, Lipson, & Hopcroft, 2015; L. Wang et al., 2018) (but see also Kornblith, Norouzi, Lee, & Hinton, 2019). Despite these challenges, neural networks offer us unprecedented opportunities to explore how perceptual features and behavior emerge in a learning system.

Several studies have reported finding similarities between behavior of neural networks and human observers. Ward (2019) found that a VGG19 network (Simonyan & Zisserman, 2015) trained on object recognition developed sensitivity to the Müller-Lyer illusion. Watanabe, Kitaoka, Sakamoto, Yasugi, and Tanaka (2018) reported that PredNet (a video prediction network implementing the theory of predictive coding; Lotter, Kreiman, & Cox, 2017) trained in videos of self-motion predicted motion in a display of the rotating snakes illusion. The authors take this as evidence that predictive coding theory provides a basis for explaining visual motion illusions. Gomez-Villa, Martin, Vazquez-Corral, & Bertalmio, 2018) found that CNNs trained on denoising or deblurring images or filtering images for color constancy show sensitivity to different brightness and color illusions. The susceptibility to certain illusions changed with model architecture and the task the CNNs were trained on. The observation that visual illusions like those documented in human observers arise as a by-product of other tasks, suggests certain similarities in processing in CNNs and the human visual system. It also provides an opportunity for insight, which illusions arise from which tasks and how efficient visual processing may be connected to illusions (Gomez-Villa et al., 2018).

To human vision scientists, the learned representations of neural networks are especially interesting for investigating intermediate processing stages and mid-level aspects of vision like material perception. Every subsequent layer learns to identify joint regularities in earlier layer representations. By fitting these networks to large datasets, they learn statistical regularities and features that are present in the data and are useful for solving the posed vision task. After fitting, early level filters in networks trained on object recognition are tuned to simple image features, such as colors, edge orientations and frequencies (Krizhevsky et al., 2012) and have been compared to receptive fields in early stages of processing in the human visual system

(Kietzmann et al., 2018; Kriegeskorte, 2015). Receptive fields in subsequent layers have been associated with texture, with late layers becoming sensitive to complex features such as parts of objects and eventually entire objects or scenes (Zhou, Bau, Oliva, & Torralba, 2018). This hierarchy is similar to organization in the human brain (Kriegeskorte, 2015). Several studies have found similarities between activations in artificial neural networks and the human brain (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014).

There are also notable behavioral differences between CNNs and the human visual system. For example, it has been shown that CNNs' object recognition performance decreases much more rapidly than that of humans when confronted with deteriorated images (Geirhos et al., 2017; Geirhos, Temme, et al., 2018). Similarly, CNNs trained to recognize objects in photographs show poorer performance compared to humans when confronted with abstract representations such as sketches or line drawings (Singer, Seeliger, Kietzmann, & Hebart, 2021).

The strategies and features used by CNNs can also be very different from human observers. Szegedy et al. (2014), showed that very specific small changes to images that are effectively imperceptible to humans, can cause CNNs to predict wrong object categories with high confidence. These effects are known as adversarial attacks and have since been shown to work for facial recognition (Sharif, Bhagavatula, Bauer, & Reiter, 2016) and photographs of printed manipulated images (Kurakin et al., 2019). Geirhos, Rubisch, et al. (2018) showed that CNNs trained on the ImageNet object recognition dataset (Russakovsky et al., 2015) are biased towards identifying textures rather than the contour shape of objects.

1.2.3 Gloss and material perception with neural networks

Research into material and specifically gloss perception with artificial neural networks has for the most part concentrated on achieving high veridical performance in material classification or image segmentation tasks rather than understanding or replicating human perception (eg. Bell, Upchurch, Snavely, & Bala, 2015; Fu, Zhang, Lin, Zhu, & Xiao, 2020; Schwartz & Nishino, 2020; T.-C. Wang et al., 2016; Zhang, Ozay, Liu, & Okatani, 2016). Intrinsic image decomposition is a long-standing problem in computer vision and approaches using CNN models have achieved state of the art results (Baslamisli, Le, & Gevers, 2018; Bonneel, Kovacs, Paris, & Bala, 2017). In the context of intrinsic image decomposition, specular reflections and highlights are often treated as interference in image processing that need to be identified to be removed or discarded (eg. Attard, Debono, Valentino, & Castro, 2020; Madessa, Dong, Gan, & Gao, 2020). Georgoulis et al. (2018) have demonstrated an effective method for extracting a

BRDF and natural illumination maps from renderings of complex shapes using a series of CNNs.

Storrs, Anderson & Fleming (2021) trained an unsupervised neural network to generate images of glossy surfaces from a combination of distributions of the image structure and a learned low dimensional summary encoding based on convolutional processing, learned from a large number of such images. The authors found that the representations in this network spontaneously grouped images according to glossiness, and illumination environment and direction. The network also displayed similar failures of gloss constancy as human observers, resembling effects of factors other than physical reflectance on perceived reflectance observed by Marlow et al. (2012). This indicates that such dimensions of gloss and lighting environment are effective summary statistics to characterize images of glossy surfaces by. It also provides evidence for unsupervised learning as a possible mechanism in perceptual learning.

van Assen et al. (2020) have demonstrated how investigating learned features in the context of material perception can give us insight into what intermediate steps and representations may be involved in human perception. The authors trained a neural network to predict viscosity of liquids visually from short animation sequences. In destroying network units artificially and testing their impact on network performance, the authors discovered that a particularly crucial group of units did not resemble any of the 18 hypothesised predictors, indicating that features not considered by the authors (or possibly not easily described) play an important role. This also demonstrates the potential for CNNs to help researchers identify novel perceptual cues.

In this thesis I will use CNNs as a tool for investigating human visual perception of gloss. Among the observed similarities between CNNs and the human visual system is the hierarchical organization and increasing complexity of features that both neurons and units respond to. This makes CNNs very promising candidates as image computable models, that contain similar mechanisms as are proposed for human mid-level vision. A key question about gloss perception is how the visual system is able to compute this attribute of material appearance from the large range of different images it can cause. Machine learning offers the opportunity to train CNNs on very large image sets, by which networks learn to extract features that are useful for performing a given visual task over a large range of different instances. We want to use CNNs to investigate how far a purely image based model can go in imitating human gloss perception, and to find out what internal features it learns to use.

1.3 Research aims and overview

Broadly speaking this thesis has two aims. The first is to investigate and better understand human perception of gloss. The second aim is to explore applications of machine learning and artificial neural networks as a tool to investigate human visual perception and especially mid-level vision. More specifically the aims are:

1. To find image computable models of two central tasks of gloss perception – gloss classification (**chapter 2**) and localization of highlights (**chapter 3**). The aim is to find models for both tasks that respond like human observers do. Such models should replicate specific errors and successes in human perception.
2. To investigate the architectural demands CNNs need to fulfill to qualify as a model of human gloss perception (**chapter 2**). The architectural specifications of human-like models could give insights into the computations that make a model react like a human. Depth of networks could act as a rough approximation of the complexity of computations.
3. To investigate how well human-like behavior of CNNs transfers to other tasks and stimuli. This has implications about the functions and features CNNs learn and whether these are specific to the task they are trained on or describe strategies that lead to more generally human-like behavior.
4. To investigate the internal representations learned by an image computable model that responds similarly to humans (**chapter 3**). Comparing internal representations of a network to candidate predictors such as intrinsic components of images, geometrical factors or local contrast filters can shed light on the intermediate stages a model learns to arrive at a human-like solution.

Chapter 2 focuses on a simple task in gloss perception – discriminating between high gloss and textured low gloss materials. The main goal of the project was to find an image computable model of gloss perception that responds similar to human observers and to identify factors of the architecture and training that produce such a model. We investigate feed-forward models for gloss classification using readout training based on networks previously described in literature, as well as using a large-scale Bayesian search of hyperparameter settings that yield networks that respond like humans. We determine models' similarity to humans in terms of correlation of networks to mean human responses on a diagnostic test set of images. This image set was carefully selected to include both successes and failures from human observers. We

also test network reactions to image manipulations of highlights that have been shown to influence human gloss perception. In addition to CNNs for categorization we train DCGANs of different depths to recreate images of low- and high-gloss materials to investigate the architectural requirements to create image features that are convincing to human observers. This project focusses on the architecture and especially depth of feedforward convolutional neural networks as a model of human gloss discrimination. In this project we sought to better understand the architectural requirements and computational complexity needed for a network to discriminate low- from high-gloss materials like human observers. The results of readout training and the Bayesian search place the most human-like models at a shallow intermediate depth (around 2-5 layers). This is also the range of shallowest network depths where we find a maximum effect of manipulations to specular highlights that resembles effects observed in humans. Similarly, using generated images we find that humans can classify gloss successfully in images generated by 2-layer DCGANs and that images from 3-layer DCGANs are as distinguishable as rendered images.

In **chapter 3** we investigated human perception and recognition of specular highlights. We asked human observers to classify individual pixels in images of glossy textured surfaces as containing a highlight or not. This task probes observers' ability to differentiate highlight reflections from texture markings, which is an impressive computational feat of the visual system as the two can produce very similar images. We trained a CNN as an image computable model of highlight localization in images of glossy textured surfaces. To model human responses, we used in depth probing of a single network. Rather than searching multiple architectures we focused on one architecture and on identifying what components of this network make it behave more like human observers. For this we use network pruning. Whereas pruning is commonly used as a tool to delete redundant or ineffective neurons from a network in order to reduce network complexity and processing requirements, we use it as a second fitting stage to modify network behavior to respond more like human observers. We then investigate the differences between the network before and after pruning and compare internal representations of the pruned model to various candidate predictors. In this project we wanted to better understand intermediate representations learned by a network that recognizes specular highlights similarly to humans. Our results indicate that the network has learned internal representations similar to different predictor categories, some most similar to geometric factors of the scene. However, in a lesion analysis we find that neither units sensitive to simple image features or units similar to geometric representations are more important for the network to perform well. We also find the network to be only very weakly sensitive to violations of photo-

geometric constraints. We find that geometric representations are not crucial for an image computable network of highlight localization to respond like human observers.

Both projects use supervised training of feed-forward networks with stimuli generated using computer graphics and labeled according to ground truth of the simulation. In both cases we train networks on veridical labels and investigate what makes a model behave like humans. The rationale is that the human visual system too aims to solve gloss perception robustly for variations in shape and illumination. To do this, it likely uses heuristics and statistical regularities, which generalize across a wide range of viewing conditions and often lead to correct results, but can also lead to very specific failures. A machine learning model will also find an approximate solution. In a human-like model, this approximate solution will be more similar to the solution provided by the human visual system, replicating failures that humans make. This would indicate that a model has learned to focus on similar features as human observers. In **chapter 2** we use such an approach to focus on the architecture and hyperparameters of a CNN, while in **chapter 3** we use it to learn more about the essential components that make one network more human-like. Another question that could potentially be investigated with CNNs is that of human optimality. If CNNs perform better at gloss perception tasks than humans, this would imply that the human visual system does not use the available image information optimally for gloss perception.

For the purposes of these projects gloss is operationalized as a single dimension in the underlying render engines. The stimuli in **chapter 3** use only one material, while the stimuli in **chapter 2** consist of two materials created from linear interpolations between an ideal specular and ideal diffuse materials. They only vary along one dimension – the relative intensity of the specular component. Furthermore, since the networks in **chapter 2** are trained to differentiate between the two materials, the network output allows only for a trade-off between the two categories, limiting the network to a one-dimensional representation of gloss. The stimuli used in **chapter 2** are created using the RADIANCE rendering software (G. J. Ward, 1994) based on the Ward reflectance model (G. J. Ward, 1992). The stimuli in **chapter 3** are created in Blender using the Cycles render engine but use a similar material definition – a (textured) matte component combined linearly with an ideal specular component.

Chapter 2:

GLOSS PERCEPTION: SEARCHING FOR A DEEP NEURAL NETWORK THAT BEHAVES LIKE HUMANS

A similar version of this chapter has been published as:
Prokott, K. E., Tamura, H., & Fleming, R. W. (2021). Gloss perception: Searching for a deep neural network that behaves like humans. *Journal of Vision*, *21*(12), 1–20.

The visual computations underlying human gloss perception remain poorly understood, and to date there is no image-computable model that reproduces human gloss judgments independent of shape and viewing conditions. Such a model could provide a powerful platform for testing hypotheses about the detailed workings of surface perception. Here, we made use of recent developments in artificial neural networks to test how well we could recreate human responses in a high-gloss vs low-gloss discrimination task. We rendered > 70 000 scenes depicting familiar objects made of either mirror-like or near-matte textured materials. We trained numerous classifiers to distinguish the two materials in our images—ranging from linear classifiers using simple pixel statistics to convolutional neural networks (CNNs) with up to 12 layers—and compared their classifications with human judgments. To determine which classifiers made the same kinds of errors as humans, we painstakingly identified a set of 60 images in which human judgments are consistently decoupled from ground-truth. We then conducted a Bayesian hyperparameter search to identify which out of several thousand CNNs most resembled humans. We find that while architecture has only a relatively weak effect, high correlations with humans are somewhat more typical in networks of shallower to intermediate depths (3-5 layers). We also trained deep convolutional generative adversarial networks (DCGANs) of different depths to recreate images based on our high and low gloss database. Responses from human observers show that 2 layers in a DCGAN can recreate gloss recognizably for human observers. Together our results indicate that human gloss classification can best be explained by computations resembling ‘early’ to ‘mid-level’ vision.

2.1 Introduction

Recognizing materials from their visual appearance is an important task for the human visual system (Adelson, 2001; Anderson, 2011; Fleming, 2014, 2017; Komatsu & Goda, 2018). One particularly interesting aspect of material perception is the perception of gloss (Chadwick & Kentridge, 2015; Marlow et al., 2012; Nishida & Shinya, 1998). From judging the freshness of food to recognizing a wet and slippery spot on the ground, gloss perception is an important daily task. Yet, despite its importance, the computations underlying human perception of gloss remain poorly understood (Anderson, 2020).

The appearance of an object results from an interaction between the object's shape, the illumination, and the object's optical properties. Material perception poses the visual system with the task of separating the contributions of these factors so that it can recognize the characteristic optical properties of a material between differently shaped objects and in a large range of environments. This is not a trivial task, as any given object can cause a wide range of retinal images depending on its viewing conditions, while objects made of different materials can yield very similar images.

Previous models of gloss perception range from very simple summary statistics—such as histogram skewness (Motoyoshi et al., 2007; Sawayama & Nishida, 2018) and contrast in particular frequency sub-bands (Boyadzhiev et al., 2015) — to the idea that sophisticated photo-geometric computations determine the causal origin of features (e.g., in distinguishing highlights from texture markings; Anderson & Kim, 2009; Kim, Marlow, & Anderson, 2011). Claims that statistics of the luminance distributions in an image can explain human gloss perception, have been contradicted by several studies showing the importance of spatial information and the congruence of specular highlights with shading patterns (Anderson & Kim, 2009; Beck & Prazdny, 1981; Kim et al., 2011; Kim & Anderson, 2010; Marlow et al., 2011; Todd et al., 2004). Indeed, identical image gradients can be interpreted as glossy or matte, depending on the apparent 3D surface structure (Marlow et al., 2015; Marlow & Anderson, 2015), reiterating in the field of gloss perception, what has been suggested for several decades in the field of lightness perception (Gilchrist, 1977). Like lightness perception, several authors place material perception—or more specifically gloss perception—as a task of 'mid-level' vision (Fleming, 2014; Kim, Marlow, & Anderson, 2012; Liu, Sharan, Adelson, & Rosenholtz, 2010; Sharan et al., 2013). Mid-level vision concerns the pooling and comparison of low-level image features such as orientation, color, brightness or scale with intermediate-level representations of surface structure, such as local surface geometry, usually with the assumption that the output of such processes disentangles physical causes that are comingled in the input.

In using mid-level features, the human visual system makes use of heuristics for gloss perception and becomes susceptible to misperceptions (see also Fleming, 2012). Marlow et al. (2012) have shown that perceived gloss varies with perceived contrast, coverage and sharpness of highlights. It has also been shown that shape and illumination can influence perceived gloss (Fleming et al., 2003; Ho et al., 2008; Olkkonen & Brainard, 2011). It is important for a good candidate model of human gloss perception to capture not just the visual system's broad successes, but also the misperceptions specific to humans.

Arguably the most fundamental gloss perception task is to distinguish categorically whether a given surface is glossy or not. Given the enormous diversity of images that can be created by glossy and non-glossy surfaces, this is a non-trivial inference. While previous studies have generally investigated gloss perception within relatively constrained stimulus sets, here, we took a 'big data' approach to the gloss classification task, using machine learning techniques to train different classifiers on a large set of images. This way, we fit our models to a dataset that encompasses large variations in appearance that high or low gloss materials can take, allowing our models to identify features that are diagnostic over a wide range of appearances. Using computer graphics, we rendered a large dataset of high- and low-gloss (near matte) textured materials under the same viewing conditions. We then compared model classifications to human judgments. We tested a range of different model classes to determine which ones best predict human responses. While research in machine learning typically focusses on achieving the best possible performance at a task, here our focus is on identifying which models reproduce the specific characteristics of *human* gloss judgments, spanning both their errors and successes of gloss classification.

To test how much of human gloss discrimination can be explained by simple image features we created two 'hand engineered' models based on summary pixel statistics and texture statistics. The first of these was a support vector machine (SVM) trained on 8 simple pixel summary statistics (mean, variance, skewness and kurtosis of pixel luminance and saturation histograms). The other was a logistic regression classifier trained on texture statistics from Portilla and Simoncelli (2000). These mid-level visual features capture color and higher-order wavelet coefficient statistics, which have been used to model human perception of texture and aspects of peripheral vision (Freeman & Simoncelli, 2011).

We also look at feedforward convolutional neural networks (CNNs). CNNs have dominated computer vision benchmark tests for object recognition for nearly a decade (Kietzmann et al., 2018) and have achieved approximately human level performance (VanRullen, 2017). There have been links drawn between CNNs and the human visual system

– from models that are architecturally inspired by our knowledge of the ventral stream (Kubilius et al., 2019; Simonyan & Zisserman, 2015; Spoerer, McClure, & Kriegeskorte, 2017) to analyses that compare responses of CNNs or representations within networks to human behavioral data (e.g., Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Kriegeskorte, Mur, & Bandettini, 2008; Rajalingham et al., 2018; Tripp, 2017; Yamins & DiCarlo, 2016). While there are many observed similarities—such as networks replicating some human visual illusions (Gomez-Villa et al., 2018; E. J. Ward, 2019; Watanabe et al., 2018)—there are also many striking dissimilarities, such as cases in which CNNs fail to replicate human behavior in simple tasks (Stabinger, Rodríguez-Sánchez, & Piater, 2016), react very differently from humans to slight changes in stimuli (Kurakin et al., 2019; Nguyen, Yosinski, & Clune, 2015; Sharif et al., 2016; Szegedy et al., 2014), and show weaker performance than humans in generalizing across different forms of image degradation (Geirhos et al., 2017; Geirhos, Temme, et al., 2018). It is important to note that merely finding comparable overall performance to humans in a given task is a weak basis for claiming equivalence, as there is a potentially infinite number of different models—with different architectures and internal representations—that could yield equivalent performance. Even comparisons based on correlations in responses across randomly chosen test images will tend to overestimate similarities between humans and models. This is because if both humans and models perform well at the task, they will tend to give similar responses to most stimuli. Because they get the answer correct in most cases, they will necessarily correlate strongly. Yet, these correlations would simply indicate that both systems perform the task well. The shared variance would be driven by the *ground truth*, not necessarily by inherent similarities in the way they arrive at their responses. As a result, comparisons based on random stimuli generally do not distinguish between different computations that achieve equivalently good performance.

In order to reveal the computations that are specific to the human visual system, it is therefore necessary to decouple (i.e., decorrelate) human perception from the ground truth. We need images for which humans make mistakes to provide a source of variation in performance that is *independent from the ground truth*, which can indicate human-specific computations. This is challenging as human perceptual errors are relatively rare. Nevertheless, here, we identified a set of images that consistently yield misperceptions, allowing us to test which models predict the specific patterns in human perception. Using these images, we then varied the CNNs' architecture in the hope to gain insights into which levels of computation are necessary and sufficient for reproducing human judgments. We systematically varied the depth (number of layers) of the networks, and for each depth searched architectural and learning

hyperparameters to identify networks that best matched humans. We reasoned that if human gloss judgments are driven by sophisticated high-level representations, deep (i.e., many-layered) networks would be required to reproduce human performance. In contrast, if human judgments are based on relatively simple image cues, shallow networks might be sufficient.

Broadly speaking there are two ways we use the term 'complexity'. One refers to stimulus characteristics. This could in principle be quantified in terms of information theoretic ideas, such as entropy. However, here, we use the term more loosely than this, in the sense of an everyday intuition about highly structured images. For example, an image of glossy surface in a natural environment containing varied and structured patterns of specular highlights, shading gradients and shadows is intuitively 'more complex' than one that contains only smooth gradients or uniform random noise (even if the entropy of the latter is actually higher than in the natural image). Our other usage of the term refers to the sophistication of computations. Simple image measurements, such as first order pixel statistics are in an intuitive sense 'less complex' than computations that involve multiple stages of operations, which pool and select information across the image in conditional or nonlinear ways to determine distal surface properties from the image. This usage is also related to stimulus complexity, as complex stimuli tend to require more sophisticated computations. Again, the term is used loosely to indicate an intuitive sense of the degree of sophistication – e.g., the number of layers in a CNN and the number of non-linearities in a network required to achieve a given response.

In a second approach we also used Deep Convolutional Generative Adversarial Networks (DCGANs; (Goodfellow et al., 2014; Radford et al., 2015) to generate new images with ambiguous material appearance after training them on our high or low gloss image dataset. DCGANs consist of two networks—a generator and a discriminator—that are trained by working against each other. Specifically, the generator network synthesizes 'fake' images, which the discriminator learns to distinguish from 'authentic' training images (in our case example renderings from the training set). The generator's goal is to create images that the discriminator cannot identify as 'fakes'. Both networks are trained simultaneously to improve at their respective tasks, causing the generator to create images that are progressively harder for the discriminator to distinguish from the training images. During training, the generator learns to synthesize images using new and different features, while the discriminator learns to identify these features to decide whether the image is generated or part of the training set. Again, we sought to identify at which network depths these images included features that humans can use to identify high or low gloss materials. We trained DCGANs of different network depths and showed the generated images to human observers to see at which depths observers can

accurately distinguish recreations of high gloss images from recreations of low gloss images. Again, if human gloss classification is driven by simple, low-level image statistics, then relatively shallow DCGANs should be sufficient to evoke compelling gloss percepts in humans. In contrast, deeper DCGANs can reproduce more sophisticated image structures, yielding impressions of bounded objects with internally coherent surface and image structure. If humans require such cues, then the deeper DCGANs would be necessary to yield reliable gloss judgments.

2.2 Methods

2.2.1 Stimuli

To train our classifiers and test our observers we used 128×128-pixel computer renderings created with Radiance (Ward, 1994). We gathered a database of 1834 3D object meshes which included natural objects and manmade artifacts, as well as a set of 214 HDR illumination maps ('light probes'). The light probes came from various sources, among them two scientific databases (Adams et al., 2016; Debevec, 1998; see **Appendix A** for a list of all sources). We rendered random combinations of objects and illuminations, picking random viewing angles from a hemisphere above the object, with the object centered in the image. We rendered the object in a completely specular and a completely diffuse material for each scene and used a linear combination of these two to create different levels of gloss. For all experiments reported here, we only used two levels of gloss—'high gloss' (specular component × 0.98 + diffuse component × 0.02) and 'low gloss' (specular component × 0.02 + diffuse component × 0.98). Because images of completely smooth diffuse materials could be easily distinguished from images of specular materials based on trivial cues like overall brightness and contrast, we added textures to the diffuse component. The textures were created by mapping marbled distortions of randomly selected illumination maps onto the surface of the object. They were multiplied with the diffuse component. The overall formula for combining the components of our images was as follows:

specular image × specular weight
+ diffuse image × texture image × (1-specular weight)
+ (1-alpha map) × background

We discarded any images where the object covered less than 20% or more than 90% of the image. The total number of images was 149922 (74961 high-gloss and 74961 low-gloss).

To train the DCGANs, we created another dataset with the same procedure and based on the same object meshes and illumination maps. The only difference was an increased viewing distance to ensure that the bounding box of the objects was completely contained in the viewing window. Initial tests showed that DCGANs produced images more resembling objects on a background rather than patches of material interspersed with patches of background when the objects in the training images were completely within the image boundaries. Again, we discarded images where less than 20% or more than 90% of pixels were covered by the object. This image set contained 187 630 images (93 815 high- and low-gloss images each).

2.2.2 Experiments with Human Observers

For all lab experiments with human observers, we presented 128×128-pixel images from our set of renderings on a black background. Participants were shown the images for as long as they took to respond. Before each experiment, participants saw 12 example stimuli at 512×512 resolution (6 low gloss and 6 high gloss images). These were not included as stimuli in the experiment. All observers had normal or corrected to normal vision and signed a consent form in accordance with the declaration of Helsinki (2008).

2.2.2.1 Random Images Experiment

10 observers (all female; age m = 26.1, sd = 4.04) were shown 150 randomly selected images from each ground truth material, one at a time. They were asked to classify the images as either ‘high gloss’ or ‘low gloss’ by pressing one of two different keys. There were four repetitions of the image set, each time in a different random order. Every 100 trials participant were asked to take a short break. The experiment lasted between 1 and 1.5 hours.

2.2.2.2 Selecting a Diagnostic Image Set

Responses to a randomly chosen image set are not well suited to assessing a model’s similarity to human observers. Since we expect both humans and at least some models to solve the task well, there will be a misleadingly high similarity in their responses to randomly chosen images. This similarity would be due to both sets of responses being similar to ground truth. Any model that solves the task better would therefore seem to be more similar to humans while it is actually only more similar to ground truth. We therefore selected a *diagnostic set* of images, in which human responses were decorrelated from ground truth. This diagnostic image set was selected in two steps. We first did two series of pre-screening experiments in the lab where observers

would categorize images as either high- or low-gloss images. This resulted in 500 candidate images. The second step was an online crowd souring experiment in which participants judged the pre-selected images on a 5-point rating scale from low to high gloss. Based on these responses we selected the final diagnostic set comprising 60 images—30 from each ground truth material—in which images from both materials were evenly distributed across three bins of perceived gloss.

2.2.2.3 Image Pre-screening for crowd sourcing

Over the course of 5 experiments, we used subject responses to select a set of 500 candidate images starting from 31 500 randomly selected images. Participants were shown images one at a time and were asked to classify them as 'high gloss' or 'low gloss' by pressing one of two keys. In addition, they could flag an image using the space-bar if they found there was no recognizable object in the image. Every participant saw 1500 images. In the first round we showed 15 000 images selected randomly from the overall set, divided among 10 subjects (8 female, 2 male; age m = 24.8, sd = 4.8), so each subject saw 1 500 images (750 from each ground truth material) and each image was judged by one subject. For the second round, we removed all images that were flagged as unrecognizable. We selected all of the remaining images that were classified incorrectly (587 low gloss and 1817 high gloss images) plus correctly judged images to total 2250 images from each ground truth category (1663 and 433 correctly judged low- and high gloss images respectively). These were judged by 15 participants (12 female, 3 male; age m = 23.7, sd = 3.8). Again, every participant saw 750 images from each ground truth material, resulting in 5 judgements per image. These results were combined with the classifications of these images from the first round. From these results—6 binary judgements on each image—we divided the images from each ground truth into 7 bins according to the mean responses. For ground truth high-gloss images, we picked 750 images—107 from each bin, and 108 from the most incorrectly judged bin. For ground truth low-gloss images there were not enough images in each bin to pick the same amount. Where this was the case we picked all images from that bin, and added the difference between the actual bin size and the target number of images to the target number for the next bin. We did this procedure starting with the bin of most incorrectly judged images. The resulting set of 1 500 images was then judged again by 4 participants (3 female, 1 male; age m = 22.5, sd = 2.1), which resulted in 10 classifications per image after combining these results with those of the first two rounds.

Because the number of incorrectly perceived high gloss images was much larger than that of low gloss images, we repeated the search progress. This time we tested in two stages. In

the first stage we showed 16 500 images (12 000 high gloss and 4 500 low gloss) to 16 subjects (14 female, 2 male; age m = 23.8, sd = 3.2)—1500 images each (750 low gloss and 750 high gloss), resulting in one classification response per low gloss image. High gloss images were included to balance the stimulus set, but the data was not used to identify candidate images. These were shown to several subjects, while low gloss images were seen by only one subject each. For the second stage we again removed all images that were flagged as unrecognizable and from the remainder took 750 low gloss and 750 high gloss images (favoring incorrectly judged low gloss images) and tested 9 more subjects (6 female, 3 male; age m = 23.9, sd = 2.9) on these 1500 images, resulting in 10 binary judgements for each low gloss image. We did not use the high gloss images from this experiment because we already had enough to fill our diagnostic set from the first set of experiments.

The images from the final stage of the first set of experiments, and the low gloss images from the final stage of the second set of experiments were combined and divided into 5 bins from "seen as low gloss" to "seen as high gloss". We picked 50 images from each ground truth material per bin, except for the most "seen as high gloss" bin, where there was only one ground truth low gloss image. We filled this bin up with ground truth high gloss images. These 500 candidate images were the set we used in the crowd sourcing experiment. These images and the images of the same scenes from the other material category were withheld from training the classifiers, leaving 148922 images for training and validation.

2.2.2.4 Online Crowd Sourcing Experiment

For our crowd sourcing experiment, we recruited participants through an online platform, Clickworker. Instructions were shown in German and English at the same time, and participants were recruited to be between 18 and 60 years old. 99 people participated and judged the 500 images that resulted from our image prescreening experiments. Each participant was shown the same high-resolution example images we used in our lab experiments, and then judged each image from our test set on a 5-point rating scale from 'low gloss' to 'high gloss'. We included two photographs—one of a sandcastle and one of a silver teapot—as catch trials at the end of the experiment. Participant data was excluded if their dataset included too many or too few trials (resulting from using the 'back' and 'forwards' buttons on their web browser) or if they failed to judge the teapot in either of the two highest gloss categories or the sandcastle in either of the two lowest gloss categories. Response data from 35 participants was excluded based on these criteria, leaving 64 participant responses. We divided the 500 candidate images into 5 bins based on mean gloss ratings of the 64 subjects. The diagnostic set needed to be balanced across these bins and was therefore limited by the number of low gloss images in the highest

gloss bin (0) and in the second highest gloss bin (10). We chose 10 images from the three central bins for both materials (randomly for bins that contained more than 10 images) to make up the diagnostic image set. Example images are shown in **Figure 3a**.

2.2.3 Classifiers

2.2.3.1 Experiments with Diagnostic Image Set

As an initial test of CNN performance on our task and similarity of responses to humans, we applied readout training to the AlexNet (Krizhevsky et al., 2012) and VGG16 (Simonyan & Zisserman, 2015) networks. These networks were pre-trained on the ImageNet object recognition task. We took the networks up to a certain layer and added a linear output layer (or dense layer), which we trained to perform our classification task, leaving the rest of the network unchanged. We did this for each convolutional and dense layer in each of the architectures, taking the readout after the subsequent pooling and ReLU layers (or just before the next convolutional or dense layers). For each network and each readout layer we trained five instances of the classifier, each time with a random initialization of the final dense layer. Models were trained on 90% (134 030) of our images, while 10% (14 892 images, balanced for both materials) were randomly picked and withheld as a validation set. Loss on the validation set was calculated every 100 training steps. Training would stop if there was no improvement for 5 consecutive validation steps. We chose this criterion rather than a fixed training duration to balance training for different architectures with different numbers of parameters.

To investigate how well a CNN trained from random initial weights could model human gloss perception we wanted to train and test networks that span a large space of possible hyperparameter values. To train and test such a large number of networks we decided on a general architecture, as shown in **Figure 4a**. We also picked a number of hyperparameters to optimize. These were parameters of both the architecture (size and number of filters in each convolutional layer) as well as the training (learning rate, learning momentum, L2-regularization). For networks up to 6 convolutional layers we included a max-pooling layer with a stride of 2 after each convolutional layer. For more than 6 layers the image size became too small, so we added any further layers without subsequent pooling. In addition, we defined hyperparameters that allowed the search algorithm to change the position of these convolutional layers without pooling within the network.

We varied the hyperparameters by letting a Bayesian search algorithm search for those parameters which—after training—would result in a network with a high correlation with

human observers on our diagnostic image set. Networks were trained on 90% (134 030) of our images. Again, the training progress was monitored by calculating the accuracy and loss on a validation set every 100 training iterations. The validation set consisted of 10% of the rendered images (14 892 images) that were picked at random, balanced for both materials, and withheld from training. If there was no improvement in validation loss for five consecutive validation steps, the training would stop. The trained network was then tested on the diagnostic image set and its responses correlated with the mean judgements of human observers. The Bayesian search program used this correlation as the objective to be maximized. A Bayesian search approach to hyperparameter optimization has been shown to be effective in finding hyperparameter settings that yield a well-performing network (Snoek, Larochelle, & Adams, 2012). Here we use an optimization approach to look for those combinations of hyperparameters that cause a network after training to correlate highly with human observers.

In addition to these CNNs we trained two linear models using 'hand engineered' features: A Support Vector Machine (SVM) using pixel statistics and a logistic regression using dimensionally reduced Portilla-Simoncelli color texture statistics (Portilla & Simoncelli, 2000). The pixel statistics we used were the mean, variance, skewness and kurtosis of pixel luminance and saturation histograms. To the Portilla-Simoncelli statistics we applied PCA to reduce the dimensionality from 3381 to 817 dimensions, which explained over 99% of the variance in our images captured by the 3381 parameters. 58 images were excluded from the training data (29 high gloss images and renderings of the same scenes with the low gloss material) for causing errors in the Portilla-Simoncelli color texture analysis, leaving 148864 images. These classifiers were each trained 20 times by splitting our image set in half, training one network on one half and testing it on the other and vice versa. This resulted in 10 predictions for each image from these classifiers.

2.2.3.2 Experiments with manipulated specular components

To test our CNN models' reactions to factors that have been shown to influence human gloss perception (Anderson & Kim, 2009; Marlow et al., 2012) we prepared a test set of images for which we manipulated the specular component. See **Figure 5a** for examples. Specifically, we manipulated the *contrast* of highlights by changing the relative weight of the specular component (the levels we used were 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0), the *size* of highlights by applying erosion and dilation to the specular component (with radii from 2 to 5 pixels each), and the *orientation* of highlights by rotating the specular component of the images in steps of 10° up to 90° in both directions. For size and orientation manipulations we chose an intermediate specular weight of 0.1. The training set contained images with specular weights

of 0.02 and 0.98. We chose an intermediate level for these manipulated images expecting intermediate responses for the unmanipulated images, so there would be no ceiling effects limiting network responses in either direction. For the orientation manipulations we used an alpha channel to limit the specular component to the area that overlapped with the diffuse-textured component. To control for this reduced area of the specular component we rendered parallel images to go with each 'rotated' image, containing reflections at the correct orientation, but cut to the same shape as was caused by the rotation. We applied all manipulations to 120 images, which were generated according to the same principles as the original training set, but not included in the training or test image sets.

2.2.4 DCGANs

As a starting point for our DCGAN architecture we used the architecture described by Radford et al. (2015). We added a 5th convolutional layer, because the original DCGAN is designed to generate 64×64 px images. From there we created architectures for shallower networks based on the following principles:

- Image resolution doubles between deconvolutional layers in the generator network and is halved between convolutional layers in the discriminator.
- Processing depth (number of filters) decreases by half for later deconvolutional layers and doubles for later convolutional layers.
- for shallower networks we would 'skip' processing at lower resolution. For a 5-layer generator network, deconvolutional processing starts on a 4×4 px representation, for a 4-layer generator at 8×8 px, for a 3-layer generator at 16×16 px, etc.
- The latent space was the same for all network depths (100×1)

An overview of the resulting architectures can be seen in **Figure S2**. We trained two instances of each architecture – one on low gloss renderings and one on high gloss renderings. DCGANs were trained using the MatConvNet toolbox for Matlab (Vedaldi & Lenc, 2015). Since DCGANs have no objective function that captures model performance and image quality (Salimans et al., 2016), we include an assessment of image realism in our experiment with human observers (see below).

2.2.4.1 Human responses to DCGAN images

We selected 75 images generated from each of our 10 DCGANs, resulting in 150 images from each network architecture, half of which were based on low gloss and half on high gloss

renderings. In addition, we added 75 randomly picked low gloss and 75 high gloss images to the stimulus set, making a total of 900 images. See **Figure 6a** for example images.

A group of 15 subjects (3 female, 12 male; age m = 24.2, sd = 4.0) were shown these images one at a time, and were asked to respond on a triangular rating field. The triangle corners were labeled 'low gloss' and 'high gloss' along the horizontal bottom edge, and 'unreal / not an object' on the top corner. **Figure 6b** shows the rating scale. Observers moved the position of the cursor within the rating field with the mouse. The experiment lasted between 1 and 1.5 hours.

2.3 Results and Discussion

2.3.1 Human performance on random and diagnostic images

We created a dataset of 74 961 scenes, showing a familiar object under image-based illumination. Each scene was rendered once with the object made of a mirror-like ('high gloss') material and once with the object made of a near-matte textured ('low gloss') material, yielding a total of 149 922 images.

Human observers were mostly able to discriminate between our two material categories. We asked 10 participants (all female; age m = 26.1, sd = 4.0) to classify 300 randomly selected images based on their glossiness. There were 150 images from each material category and every observer saw each image 4 times. On average human observers judged 87.6 % of images correctly (sd = 4.6%).

Responses to such a randomly chosen image set are not a sufficient criterion to evaluate how well a model replicates human perceptual judgements. It is not enough for a model to make the same number of correct or wrong decisions as humans, but rather a good model should also make correct or wrong decisions on the same images that humans do. On a randomly chosen image set, humans perform well above chance and their responses are highly correlated with ground truth. Any model that solves the task well will therefore also correlate highly with human observers. Thus, to decorrelate model accuracy from similarity to humans we assembled a *diagnostic set* of images in which mean human judgements are decorrelated from ground truth. In a series of lab experiments we showed a total of 31 500 images to 54 participants (43 female, 11 male; age m = 23.9, sd = 3.5) in a binary classification task and used their responses to identify 500 candidate images that frequently yielded errors. The experiments were conducted in several rounds, where we used subjects' responses to narrow down the set of candidate images for later rounds. Some images were seen by only one observer while the final candidate images were seen by 10 observers (see **Methods** and **Figure S1** for details). From

these candidate images we selected 500 images which we showed to 99 participants (aged 20 – 64, no gender information available) in a crowd sourcing experiment to judge each image on a 5-point rating scale from high to low gloss. We excluded data from 35 participants based on double trials or skipped trials (resulting from using 'back' or 'forward' buttons in their web browsers) or failing at least one of two catch trials at the end of the experiment (see also **Methods**). From the ratings of the remaining 64 participants we identified the 60 final images that make up our 'diagnostic image set'. The diagnostic set contained 30 images from each of the two reflectance categories. These were selected so that the mean responses across crowd sourcing participants classify an equal number of images wrongly, correctly and half-way between high and low gloss. This dataset allows us to test to what extent a model makes similar perceptual decisions as humans, independently of the model's accuracy. Thus, on this diagnostic image set, human performance was by definition chance (53.3% accuracy of mean human responses). Correlation between mean human response and ground truth was $r = 0.13$; $p = 0.32$. Example images from the set are shown in **Figure 3a**.

2.3.2 Read-out networks

As a pilot experiment to test different model complexities we looked at read-out networks of two well-researched CNN architectures, AlexNet (Krizhevsky et al., 2012) and VGG16 (Simonyan & Zisserman, 2015) that had been trained on the ImageNet object recognition task. At the time this study was conceived, VGG 16 was state of the art in imitating the architecture of the human ventral stream. We used two architectures to ensure that observations we make about readout networks generalize between source networks and are not specific to either one. We trained a linear classifier on representations at different stages throughout both networks to perform the high gloss *vs* low gloss classification task. For each layer from which we took read-out features, we trained five instances of the linear classifier. The performance of these networks in terms of accuracy and their correlation to humans on the diagnostic image set is shown in **Figures 3b** and **c** respectively. There are two notable trends: The accuracy improves for read-out networks from later layers (Pearson correlation between mean accuracy and readout layer for AlexNet: $r = 0.933$, $p = .002$; for VGG16: $r = 0.8907$, $p < .001$); and read-out networks based on VGG16 representations from earlier layers show more variance in their correlations to human observers between instances (Pearson correlation between variance in correlation to humans and readout layer: $r = -0.7235$, $p = .002$). Read-out networks based on AlexNet representations showed a non-significant correlation in the same direction ($r = -0.593$, $p = 0.160$). Yet, crucially, we also found that single instances of read-out networks with the highest correlation to humans for both AlexNet and VGG16 were trained on representations of

early layers (the maximal correlations for both were achieved from second layer representations). This led us to expect that for CNNs trained from random initial weights, we can find exemplars of shallow networks that correlate well with human observers. Early layer features of object recognition CNNs tend to capture more localized spatial regularities than later layers and have been associated with textures (Zhou et al., 2018). This could also mean that texture statistics may already provide a basis for human-like gloss discrimination—at least for the types of images in our training and test sets—even if such image properties are insufficient to account for all phenomena in gloss perception (Anderson & Kim, 2009; Kim et al., 2011; Marlow et al., 2011; Marlow et al., 2015; Marlow & Anderson, 2015).

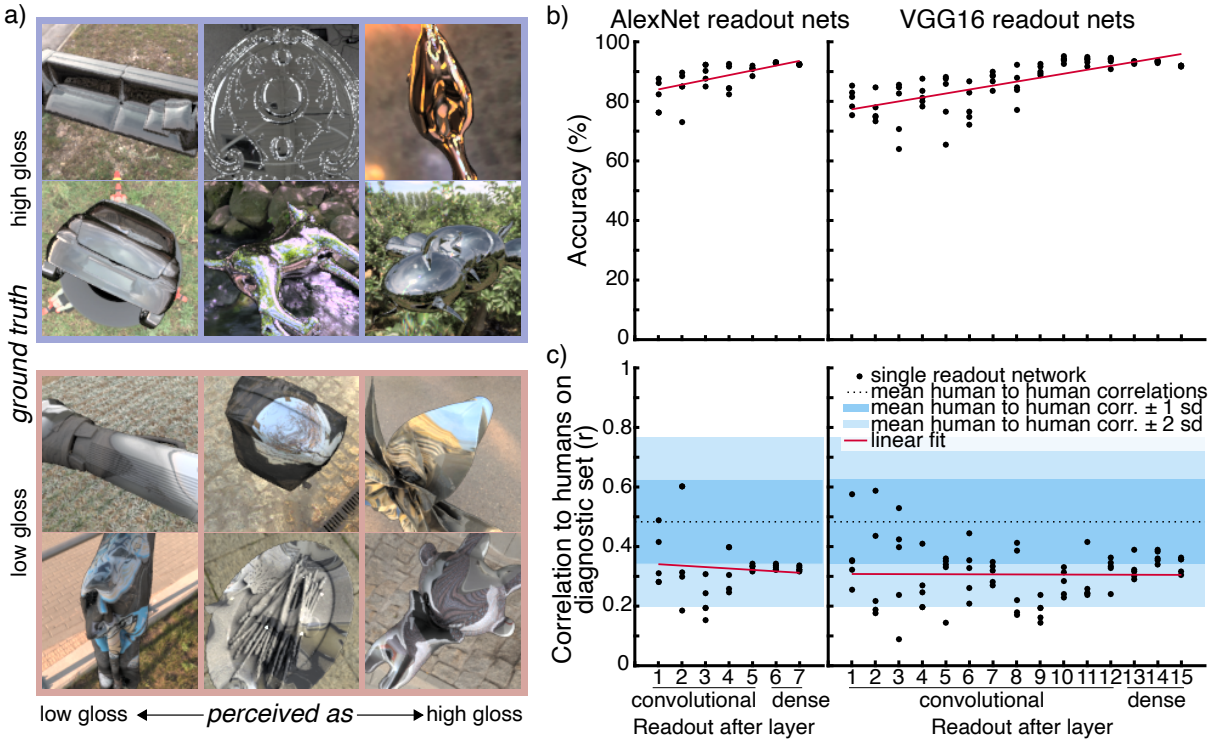

Figure 3: (a) *Example images from the diagnostic image set. Images are sorted in columns according to increasing perceived gloss from left to right. Images in the top two rows were rendered in a high gloss material, images in the bottom two rows were rendered in a low gloss textured material.* ***(b)*** *Classification accuracy of read-out networks from AlexNet and VGG16 (% correct). For every read-out network we trained 5 instances from random initialization.* ***(c)*** *Correlation of read-out networks to humans on the diagnostic image set (r). The dotted line shows the mean of correlation between individual human observers and the mean of the remaining observers. The blue areas show the first and second sd.*

2.3.3 CNNs and linear classifiers

For a detailed sampling of CNN architectures that had not previously been trained on other data, we conducted a Bayesian search, training 2700 networks from random initial weights to identify the ground truth class (low vs. high gloss) of 134 030 images from the rendered training set. We used the Bayesian search algorithm to optimize hyperparameters of the training procedure and the architecture of networks with 1, 2, 3, 4, 5, 6, 7, 8, and 12 convolutional layers, training 300 models for each network depth. The Bayesian search optimized model correlation with human responses on the diagnostic image set. Models were trained on 90% of our image set, 10% (14 892 images) being selected at random and withheld as a test set. The images in the diagnostic set were always withheld from training. For our general model architecture see **Figure 4a**; see **Methods** for details on the training.

We also trained two 'hand engineered' classifier models. One was a logistic regression trained on features we derived from a texture model of human mid-level vision (Portilla & Simoncelli, 2000; see also **Methods** for details). Lastly, we trained a support vector machine (SVM) on eight pixel statistics (mean, variance, skewness and kurtosis of the pixel luminance and saturation histograms) to differentiate high-gloss from low-gloss images. We expected linear models trained on pixel statistics to insufficiently model human responses. At the same time this is a useful benchmark to test more complex models against and to ensure that the discrimination task provided by our stimuli is not trivial. For training these 'hand-engineered' classifiers, we split our image set in half, training one classifier on each half and testing it on the other (two-fold cross validation). We repeated this ten times for each model.

An overview of how all CNNs and the two linear classifiers correlated with human observers on the diagnostic set can be seen in **Figure 4b**. We excluded all 'dead' networks (i.e. networks that resulted in a failed training or constant predictions to all stimuli). These were 61 in total (27, 23, 2, 0, 1, 1, 3, 2, 2 for 1, 2, 3, 4, 5, 6, 7, 8 and 12-layer networks respectively). We compare these correlations against the distribution of correlations of individual human observers with the mean of the remaining 63 human observers. The mean, as well as the first and second standard deviations of this distribution can be seen in the dotted line and shaded regions in **Figure 4b**. The SVM on pixel statistics clearly correlates with humans much less than the other classifiers. The logistic regression on Portilla-Simoncelli statistics is close to the mean of human-to-human correlations. Interestingly, there are examples of CNNs from all depths that correlate well with humans and there is no obvious trend or difference between the depth groups.

On average individual human responses to the diagnostic stimulus set correlate at .47 with the mean of the remaining observers, explaining only 22% of the variance in the mean responses. This means that human responses are quite idiosyncratic for the diagnostic stimulus set. The logistic regression based on Portilla-Simoncelli statistics reaches a similar correlation to the mean human response vector. However, the best of the CNN models correlates to 0.7 with the human mean, explaining 50% of the variance. The most human-like CNNs therefore explain more of the central tendency in human responses than the responses of most individual human observers do. This could reflect the fact that the human visual system has internal noise (e.g., Pelli & Farell, 1999)), while CNNs do not. This is an interesting topic for speculation but the current data do not allow for strong conclusions about the nature of noise within individual observers' data so further research on the origins of individual differences in gloss perception is necessary.

An ANOVA revealed a main effect of network depths ($F(8,2638) = 34.4$, $p < .001$), indicating that on average networks of different depths correlate differently to human observers. 1-layer networks had the highest mean correlation (mean $r = 0.49$; for the other depth groups mean r ranged from 0.41 to 0.47). However, our Bayesian search algorithm was not designed to characterize the *mean* similarity to humans for an entire hyperparameter space. Rather, it searches for those settings and individual networks that correlate *highly* with humans, choosing new settings to investigate particularly interesting regions of the hyperparameter space, rather than sampling the space in a grid-like fashion. Looking at the top 10% of most human-like classifiers we find CNNs from all depths (see **Figure 4c**), indicating that single exemplars from different network depths may result in high correlations to humans. The most human-like of all networks was a 1-layer network and indeed 27% of the top 10% most-human networks were only 1 layer deep, followed by a further 18% of 2-layer networks. This suggests that using the Bayesian search approach we applied, it is easier to find relatively shallow networks that resemble humans than deeper ones. This may reflect a use of relatively simple cues by humans, although it may also be related to the much smaller parameter space to be searched for shallow networks.

Having decorrelated human responses from ground truth, it is important to look at the performance of models as well as their similarity to humans. We sought to answer whether there is a systematic relationship between performance on the objective function the networks are trained on (i.e., accuracy at gloss classification) and their tendency to reproduce human patterns of gloss judgments. We therefore investigated how similarity to human responses and model performance are connected, and whether those CNNs that respond most like humans are

outliers in terms of their performance or show typical levels of performance for their network depth. To do this, we looked at the relationship between network accuracy (on the randomly picked 10% of images that were kept from training and used as a validation set) and the correlation coefficients to humans on the diagnostic set. Overall, these two factors barely correlate (r = -0.1 $p < .001$), confirming that using the diagnostic image set, human-specific response characteristics can be measured independently of overall performance. Yet, looking at the groups of networks with the same numbers of layers (see **Figure 4c**) revealed a clear trend: For shallower networks there is a positive correlation between accuracy and correlation with humans, meaning those exemplars that correlate well with humans are also the ones that perform better within the range of possible networks. For deeper networks there is a negative correlation. In other words, of the deep networks, the ones that correlate well with humans perform the task badly relative to other networks of the same depth. This also indicates that there is an intermediate range of network depths, where there is little or no correlation between accuracy and correlation with humans, and where a high correlation with humans is more typical. **Figure 4d** shows the correlation coefficients of the plots in **Figure 4c** as a function of network depth, showing the trend of decreasing correlation with network depth. A quadratic fit to the trend reveals an intersection with zero correlation at a depth of approximately 4-5 convolutional layers. Taken together, these analyses suggest that relatively shallow networks tend to be those that, typically and independently of their performance on the training objective, tend to correlate most closely with humans.

Our results suggest that the task of distinguishing high gloss from low gloss textured materials in a similar way to humans does not require the complexity of very deep convolutional networks. Indeed, even a linear classifier using texture statistics can match human perceptual judgements on the task at the level of human-to-human mean correlations. However, CNNs are able to explain mean human responses even better. There is no improvement of deeper networks over shallower ones, despite their increased objective accuracy at the task. Analysis suggests that networks with approximately 4-5 convolutional layers tend to typically correlate well with humans.

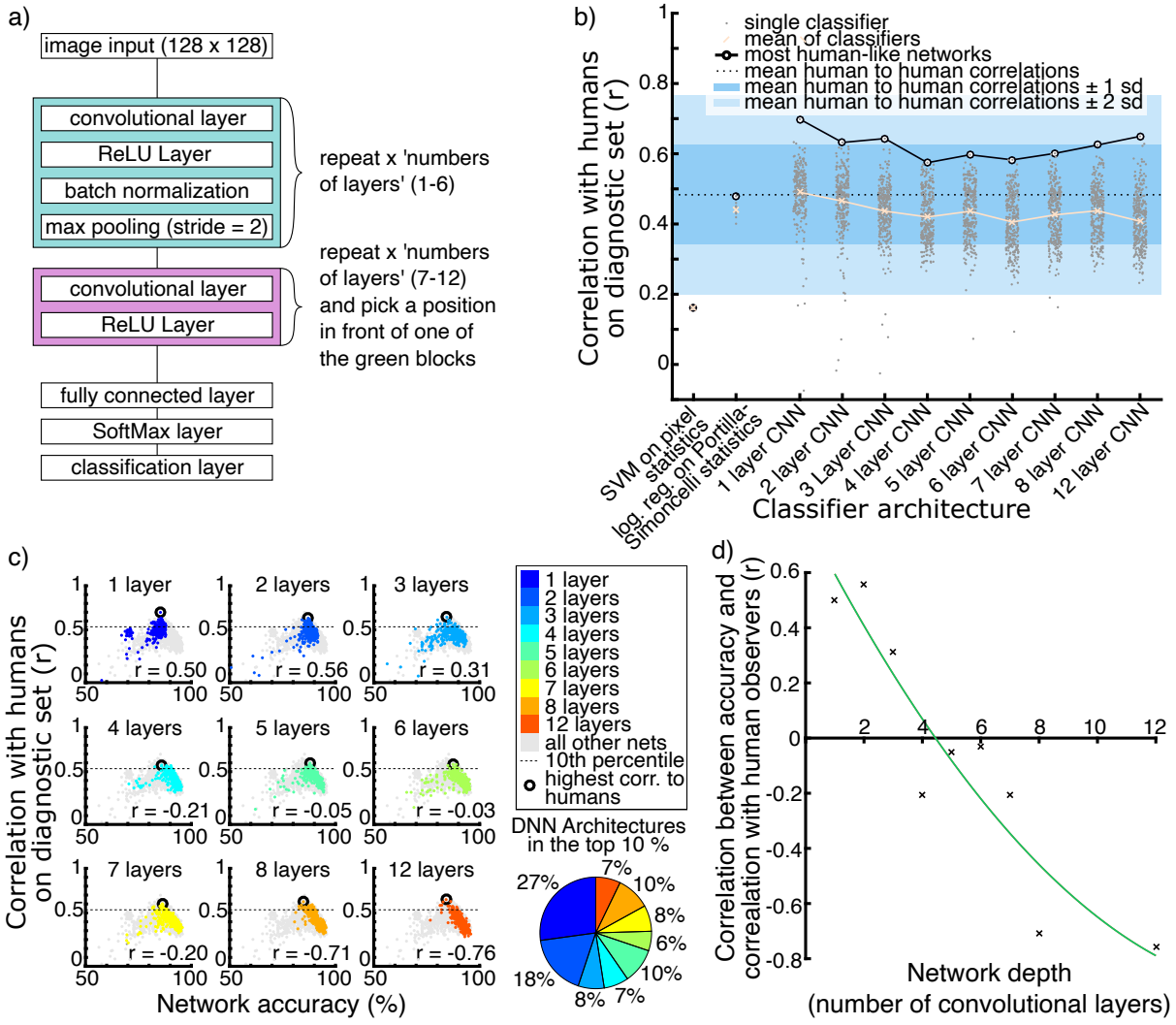

Figure 4: (a) *General architecture of the CNNs in our Bayesian search.* ***(b)*** *Overview of the performance of linear classifiers and CNNs from the Bayesian search. Performance is plotted in terms of correlation to human observers on the diagnostic image set (r). For every linear classifier we trained 20 instances. For every CNN depth we trained 300 networks, using a Bayesian search algorithm to optimize training parameters as well as the number and size of convolutional filters within each layer.* ***(c)*** *Accuracy of each network plotted against correlation to humans on the diagnostic set for each depth of CNN. Grey points show all networks, while colored points show one network depth each. Correlation coefficients are shown. Bottom right: Proportion of different CNN depths in the 10% of CNNs that correlate highest with humans.* ***(d)*** *Correlation coefficients from c) plotted against network depth, showing a trend from positive correlations for shallow networks to negative correlations for deep networks.*

To further investigate similarities in network responses to human observers, we rendered a set of images with manipulations to the specular component that have been shown to influence human gloss perception (Anderson & Kim, 2009; Marlow et al., 2012) see also **Methods** for a detailed description and **Figure 5a** for examples). These manipulations were changing the *contrast* of specular reflections by changing the specular weight, changing the *size*

of highlights by eroding or dilating the specular component, and *rotating* the specular component. The mean responses from networks in each depth group for 120 images with these manipulations are shown in **Figures 5b – d**. Networks appear to be very sensitive to specular contrast, predicting images with higher contrast reflections to be glossier. This is not surprising as specular weight is the primary manipulation in the training data. Networks also react to highlight size, predicting images with eroded highlights to be less glossy and images with dilated highlights to be glossier. Rotations of highlights appear to have little influence on network responses. On average networks predict images with rotated highlights to be less glossy than images with correctly oriented highlights. This effect does not appear to increase with the angle of rotation.

To compare the effects these manipulations have on network responses, we calculated an effect size for each manipulation and compared this between networks. For the contrast manipulation we took as an effect size the differences in responses between the highest and lowest specular weight conditions we tested (1.0 and 0.01). For size manipulations we calculated two separate effect sizes for erosion and dilation. For each we took the differences in responses between images containing unmanipulated highlights and images with the most dilated or most eroded highlights we tested (both with a radius of 5 pixels). For highlight rotations we took the difference in responses between images with rotated reflections and parallel images with non-rotated reflections for the orientation with the largest effect per depth group. **Figure 5e** shows these effect sizes for each depth group of networks. For all four manipulations the effect is smaller for very shallow networks, increasing with network depth, but also reaching a maximum. This maximum appears to be reached after network depths of 3 - 6 layers.

Our CNNs show different degrees of sensitivity to different manipulations to the specular components of images. It is not surprising that networks react strongly to changing levels of specular contrast or specular weight. This is the primary difference between the two material classes in the training set. It is however interesting to see that intermediate specular weights also cause intermediate responses and that on average responses ordinally match specular weights. Size of highlights also changes network responses – more eroded highlights causing lower gloss predictions and more dilated highlights causing higher gloss predictions on average. Rotations of the specular component appear on average to cause networks to judge images as less glossy. This effect is very small, however. A possible explanation lies in the training data and task. The training data contains only high gloss mirror-like material or low gloss material showing shading and texture and some specular highlights. While congruence

between shading and highlight components is a part of the typical appearance of the low gloss material, learning this aspect of appearance is of little consequence for the network to improve at the objective. Images containing both shading and specular highlights (and texture) are already at the low end of the learned gloss scale.

The effect sizes of highlight manipulations change with network depth. All effects are smaller for very shallow networks and increase with network depth until they appear to reach a maximum after about 3 - 6 layers. The directions of these effects are in line with what we would expect from human observers - higher specular contrast, larger highlights and correct orientation all lead to higher gloss judgements on average. However, we cannot draw any conclusions about the absolute size of these effects. It seems likely that our training data and objective, representing only a limited part of gloss appearance and perception, provided only limited opportunity for learning some aspects of gloss perception. The levelling out of effect sizes with increasing network depth however seems to imply that in so far as these features can be learned by networks in our training, they are maximally learned after about 3-6 layers. Deeper networks may learn more complex features, but they do not appear to be more sensitive to these manipulations. This is in line with our previous observation that it is more typical for networks of intermediate depths to correlate highly with humans and that further increased network complexity does not necessarily increase network similarity to human observers.

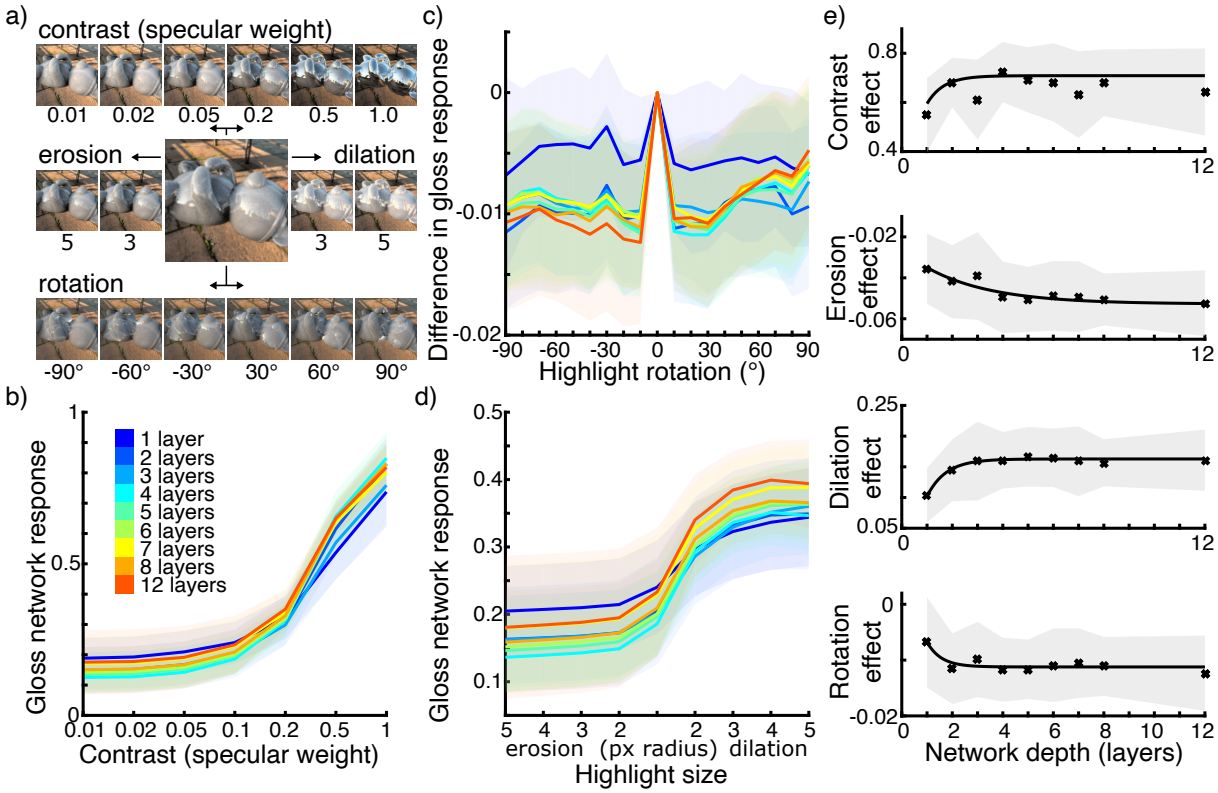

Figure 5: (a) *Examples of gloss manipulations. The larger central image shows an image with unmanipulated specular component and a specular weight of 0.1. The top row shows manipulations of contrast by changing the specular weight. The middle row shows manipulations of the size of highlights by erosion and dilation. The bottom row shows examples of images with rotated specular components.* ***(b)*** *Mean responses of networks of different depths to images with different gloss contrasts (specular weights). Shaded areas show sd.* ***(c)*** *Mean response differences for networks of different depths to images with rotated highlights. Note that the y-axis shows mean response differences rather than absolute responses because network responses for each image are compared to a parallel image with specular components that are not rotated from the original, but corrected for coverage according to the overlap in rotated images (see Methods for details). Shaded areas show sd.* ***(d)*** *Mean responses of networks of different depths to images with differently sized highlights (through erosion and dilation). Shaded areas show sd.* ***(e)*** *Mean effect sizes of gloss contrast, erosion, dilation and rotation for networks of different depths. Shaded areas show sd per network depth.*

2.3.4 Generative Models

One limitation with using renderings to evaluate human gloss perception is that ambiguous stimuli are relatively rare. To increase the number of stimuli that could be diagnostic of human perception, in a second experiment we turned to Deep Convolutional Generative Adversarial Networks (DCGANs; Goodfellow et al., 2014; Radford et al., 2015) to generate images that share certain high-order statistical characteristics with renderings, but which elicit less distinctive surface appearance. This also allows us to compare the necessary ingredients for networks to create images that appear glossy to humans, with those required by discriminator networks to match human judgments. Whereas in the previous experiments we searched for

architectures of networks that classify low and high gloss images in a similar way that humans do, in this experiment we looked at different architectures of networks that generate images of low and high gloss materials that are distinguishable to human observers. For this purpose, we trained 5 DCGAN architectures to replicate low and high gloss images ranging from 1 to 5 convolutional layers. Of each architecture we trained two exemplars – one on high gloss renderings and one on low gloss renderings. By training the DCGANs separately on each class of images we have 'ground truth' labels for which material is being recreated in each image.

From each DCGAN we generated 75 images. In addition, we randomly picked 75 high gloss and 75 low gloss renderings from the image set we used for training DCGANs. Overall this gave us a set of 900 images (5 architectures × 2 image classes × 75 generated images + 75 × 2 renderings). We showed these to participants one image at a time in random order and asked them to rate the images within a triangular rating area. The labels of the three corners were 'high gloss' and 'low gloss' (along the horizontal axis) and 'unreal / not an object' on the top corner. See **Figure 6a** for example stimuli.

15 subjects (3 female, 12 male; age m = 24.2, sd = 4.0) participated in this experiment. Their results are shown in **Figures 6b – d**. We separated subject responses into glossiness and realness—the horizontal and vertical components of their responses within the triangle's area respectively. We conducted a two-way repeated measures ANOVA of glossiness responses. Mauchly's test indicated that the sphericity assumption was violated ($\chi^2(65) = 218.4$; $p < .001$). Since Greenhouse-Geisser Epsilon $\varepsilon = 0.246$ we report p corrected according to Greenhouse-Geisser. The ANOVA reveals two within-subject main effects: The image generation method ($F(5,70) = 8.2$; $p = .003$) and the image ground truth ($F(1,14) = 97.9$; $p < .001$) These imply that images from different generation methods (networks of different depths or renderings) are perceived differently in terms of their glossiness. Similarly, overall, observers can tell low gloss images and high gloss images apart. The ANOVA also revealed an interaction term ($F(5,70) = 40.0$; $p < .001$), meaning that the difference between high and low gloss images changes between different image generation methods. To better understand this interaction, we conducted a series of comparisons of estimated marginal means of glossiness ratings of low and high gloss images for each image generation method separately. These comparisons revealed significant differences in glossiness ratings between low and high gloss images for all image generation methods except 1-layer networks ($t(14) = 0.93$; $p = 0.368$; for all others $t(14) = 5.61$ or larger; all $p < .001$)

To quantify the difference between low and high gloss images from different generation methods we looked at the classification accuracy. We define accuracy as the proportion of

images that are rated on the correct half of the glossiness scale. These are shown in **Figure 6b**. In terms of accuracy, images from two-layer networks can already be discriminated as well as those from 5-layer networks (76% vs 77%). However, the main effect of image generation method on glossiness ratings indicates that the mean ratings across low and high gloss images differ between network depths, making the 50% criterion questionable. Visual inspection of **Figure 6b** shows that for 3 to 5-layer networks the image ratings are shifted overall towards the low gloss end of the axis. As a measure of how well glossiness ratings of low and high gloss images are discriminable independent of a threshold criterion, we calculated the sensitivity index, d', for each network depth. These were d' = 0.17; 1.1; 2.3; 1.4; 1.4 and 2.2 for 1-5 layer DCGAN images and renderings respectively. According to these values images from 2-layer networks were less discriminable than those from 4- or 5-layer networks. However, images from 3-layer networks showed the same discriminability as renderings. An interesting observation is that by visual inspection of **Figures 6b and 6c** it appears that perception of high gloss images shifts for deeper DCGANs, while low gloss images (red dots in the figure) remain mostly in the same position. This could indicate that high gloss perception is more specific, and that perception of low gloss material can be achieved by a wider range of image quality and features.

We also performed a two-way repeated measurements ANOVA of 'realness' responses. Mauchly's test again indicated violation of the sphericity assumption ($\chi^2(65) = 276.6$; $p < .001$). We report p corrected according to Greenhouse-Geisser, since Greenhouse-Geisser's Epsilon $\varepsilon = 0.205$. The ANOVA of 'realness' responses revealed a within-subject main effect of image generation method ($F(5,70) = 34.1$; $p < .001$). A main effect of image ground truth was not significant ($F(1,14) = 0.1$; $p = 0.808$), indicating that low and high gloss images were not, overall, judged differently in their realism. An interaction was also significant ($F(5,70) = 4.9$; $p = .022$), meaning that for different image generation methods realness differences between low and high gloss images vary. A series of comparisons of estimated marginal means of low and high gloss image realness for each image generation method separately revealed a difference only for images from 2-layer networks ($t(14) = 3.66$; $p = .003$). This indicates that for all network architectures except 2-layer networks quality of low and high gloss images in terms of realism was not perceived differently by observers. Together these findings suggest that the lowest-level texture-like statistical image structures that can be reproduced by one-layer DCGANs are insufficient to create distinct and realistic impressions of low- and high-gloss materials. Increasing depth led to more realistic and more distinct impressions, confirming that

mid-level image organization factors play an important role in the perception of surfaces and their reflectance properties.

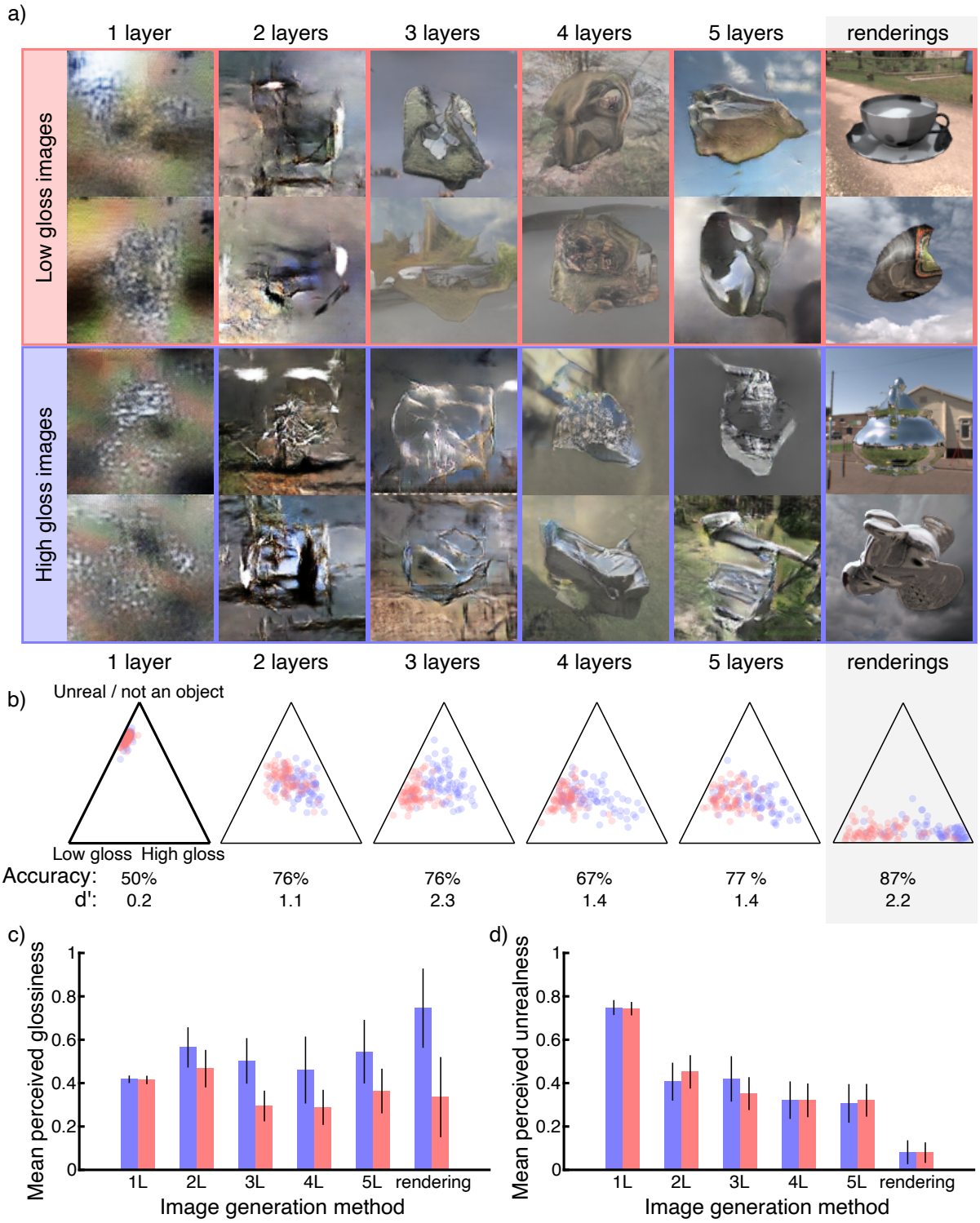

Figure 6: (a) *Example images generated by DCGANs of different depths and renderings in the right most column. The top two rows show images from networks that were trained on low gloss textured images, the bottom two rows show images by networks trained on high gloss images.* ***(b)*** *Mean ratings from 15 human observers. The left most triangle shows the labels that were displayed during the experiment.* ***(c)*** *Mean and sd of gloss ratings for each network depth* ***(d)*** *Mean and sd of unrealness / 'not an object' ratings per network depth*

2.4 General Discussion

There has been a long-running debate about which level of processing or type of information observers rely on to identify surface gloss—relatively low-level image statistics (Boyadzhiev et al., 2015; Motoyoshi et al., 2007; Sawayama & Nishida, 2018), or more sophisticated mid-level representations that capture relationships between image features and 3D surface structure (Anderson & Kim, 2009; Kim et al., 2011; Marlow et al., 2011; Marlow et al., 2015; Marlow & Anderson, 2015). Yet, the classes of information relevant for gloss judgments could plausibly vary depending on the task. Here, we focused on one of the most basic gloss-perception tasks: classifying a whole image as either high or low gloss. We asked observers to classify a diverse set of low-resolution renderings of textured low-gloss and untextured high-gloss surfaces. To our knowledge, this is one of the largest scale gloss-perception experiments performed to date, spanning tens of thousands of images.

Overall, we found that observers were good but far from perfect at distinguishing them—providing ample possibilities to identify the human-specific patterns of responses. To gain insights into the kinds of cues and computations observers used, we identified a set of diagnostic images in which human responses were systematically decoupled from ground truth, which we then compared against a variety of models, based on different types of computation. While we could rule out simple intensity and saturation statistics, features from a well-known texture analysis/synthesis algorithm (Portilla & Simoncelli, 2000) could predict mean human judgments roughly as well as individual observers could, and a range of CNNs even better. Together, these findings hint that while extremely simple image statistics cannot account for human gloss categorization, relatively straightforward mid-level image statistics—without any explicit representation of 3D structure, might be sufficient to account for human performance in this simple task. This raises the intriguing possibility that there might be a typical overall 'look' of matte and glossy surfaces, which can be captured by texture-like image statistics, and which is sufficient for matte-glossy decisions.

By systematically varying CNN architectures and hyperparameters, we sought to identify certain CNNs that more closely resembled human performance than others. However, we found that the extent to which CNNs correlated with human judgments varied surprisingly little across changes in network depth. While an ANOVA revealed a main effect of network depth, with 1-layer networks correlating highest with humans on average, there are several points to be made against considering only the mean correlation to humans per network depth. The first is that the Bayesian search algorithm is not intended to investigate the mean correlation to humans. Rather it searches the hyperparameter space for settings or a range of settings that

maximize correlation to humans, as opposed to an evenly spaced grid search. Our data shows that there are human-like networks for every depth we investigated.

Another point to be made about the mean correlations of depth groups is that we cannot discriminate between variance in model performance due to random initialization and variance due to changing hyperparameters. It is possible that random initialization causes more variance than changes in particular hyperparameters for our training set, objective function and network architectures. We have also seen effects of this in our read-out networks, where the random initialization of only a final linear layer led to a wide range of correlations with humans.

We therefore use the analysis of network performance typicality shown in **Figures 4c** and **d** as summary statistics. We found that very deep networks that correlate highly with humans tend to perform relatively poorly in terms of classification accuracy. Shallow networks that respond very human-like tend to perform relatively well. Networks of intermediate depths (4-5 layers) typically show human-like responses, independent of their accuracy. While the most human-like single network we observed was a 1-layer network, we suggest that the networks of intermediate depths are of particular interest in modelling human gloss perception, because they more typically respond similarly to humans, not just in outlier exemplars. It is also worth noting that the high prevalence of shallow networks in the top 10% most human-like networks may also be related to the size of the search space. Deeper networks have more hyperparameters and thus may require additional optimization to identify human-like exemplars. Taken together the analyses suggest that very deep representations are not required to predict human-like gloss categorizations.

These results are supported by CNNs' responses to images with manipulated highlights. We showed images with changes in specular contrast, highlight size and highlight orientation to all CNNs from our Bayesian search. Humans have been shown to be sensitive to similar manipulations (Anderson & Kim, 2009; Marlow et al., 2012). The results (**Figure 5**) show that networks are sensitive to these manipulations to different degrees. While we cannot compare the magnitude of the effects between manipulation conditions, we can compare them between networks. On average, network responses shift as we would expect from human observers. Effect sizes in responses to all manipulations increase for very shallow networks and reach a maximum after a network depth of about 3-6 layers. Deeper networks show no further increased sensitivity to these manipulations, indicating that intermediate networks learn these features as much as can be learned from our training data and task. This broadly agrees with our previous observation that intermediate CNN depths are sufficient to model human gloss perception.

Our experiment with DCGAN images showed that human observers were able to discriminate high gloss from low gloss images generated by 2-layer networks. 3-layer networks were able to generate images that were essentially discriminable as renderings to human observers. Although 4 and 5-layer DCGANs showed a decrease in discriminability, this cannot be easily explained by image quality as quantified by subjects' judgements of image realism. Possibly the complex features enabled by the models' increased depth do not contribute to gloss perception. The generator learns features (up to a certain limit) in order to replicate the image as well as the discriminator can identify, while the discriminator learns to identify features that the generator has not yet learned in order to discriminate generated images from training images. We find that starting in 2-layer networks and very much so in 3-layer networks the replicated features include at least some that human observers perceive as belonging to high or low gloss materials. While this on its own does not tell us exactly which features are necessary and sufficient, it provides converging evidence that very shallow representations are insufficient to capture the structures humans rely on, while very deep representations are not necessary.

Another interesting observation is that subjects can distinguish the glossiness of images even when they do not report perceiving an object in the image. This again hints that perception of gloss—at least at the level of simple binary classification—might be possible without perception of a coherent 3D surface. To fully evaluate this, it would be necessary to test shape perception for the DCGAN stimuli, which is an interesting avenue for future studies offered by these intriguingly ambiguous stimuli. Due to the processing differences between shallow and deeper networks it could also mean that more local cues are sufficient to perceive glossiness than are necessary to perceive an object. This is also in line with the observation from our results with read-out networks, that those single instances of classifiers that best predicted human gloss responses were trained on features from early intermediate layers. AlexNet and VGG16 from which the features for our read-out networks were taken, were trained to recognize objects. Yet the features that yielded the single most human-like read-out networks for the gloss-perception task were also from an earlier stage of processing than those necessary for image recognition, suggesting again that gloss perception does not require object perception. Taken in sum we believe our analyses provide convergent evidence that relatively low- to mid-level statistical image representations might be sufficient to account for human visual low gloss – high gloss category decisions. Taking into consideration work on cortical representations (Freeman & Simoncelli, 2011; Hiramatsu, Goda, & Komatsu, 2011; Komatsu & Goda, 2018; Nishio, Goda, & Komatsu, 2012; Okazawa, Goda, & Komatsu, 2012; Sun et al., 2016; Wada, Sakano, &

Ando, 2014) it is interesting to speculate that ventral stream areas spanning V1 to V4 might be those most important for such judgments.

2.5 Conclusions:

Our work has narrowed down and identified some architectural conditions under which CNNs (as discriminators or as DCGANs) learn features that cause similar responses to humans or that allow humans to perceive generated images as high or low gloss. There is convergent evidence from our experiments that relatively shallow architectures are sufficient to model human gloss perception with a CNN. This is supported by results of read-out networks based on AlexNet and VGG16 representations, in which single instances of linear classifiers trained on representations at early stages of these networks showed the highest correlations to human observers. Correlating the accuracies of our CNNs to their correlation coefficients with human observers showed that CNNs of approximately 4-5 layers more typically correlate well with humans on our task than deeper networks. We also find that networks of about 3-6 layers reach a maximum mean sensitivity to manipulations of highlight contrast, size and orientation. These intermediate network depths might be good candidates for further studies on modelling human gloss perception with supervised networks. Human ratings on images generated by different DCGAN architectures show that generative networks with 2 convolutional layers were enough to recreate high and low gloss materials in a way that humans could tell apart, while 3 convolutional layers were enough for human observers to distinguish as well as renderings. Linear classifiers using pixel intensity and saturation statistics were not enough to imitate human observers, and while the correlation to humans of logistic regressions trained on texture statistics came close to mean human-to-human correlations, they were surpassed by CNNs. Overall, these results support the view of gloss classification as a computation of low to mid-level-vision.

Chapter 3:

IDENTIFYING SPECULAR HIGHLIGHTS: INSIGHTS FROM DEEP LEARNING

A similar version of this chapter has been submitted as:
Prokott, K. E. & Fleming, R. W. (*under review*). Identifying specular highlights: insights from deep learning.

Specular highlights are the most important image feature for surface gloss perception. Yet, recognizing whether a bright patch in an image is due to specular reflection or some other cause (e.g., texture marking) is challenging, and it remains unclear how the visual system reliably identifies highlights. There is currently no image-computable model that emulates human highlight identification, so here we sought to develop a neural network that reproduces observers' characteristic successes and failures. We rendered 179 085 images of glossy, undulating, textured surfaces. Given such images as input, a feedforward convolutional neural network was trained to output an image containing only the specular reflectance component. Participants viewed such images and reported whether specific pixels were highlights or not. The queried pixels were carefully selected to distinguish between ground truth and a simple thresholding of image intensity. The neural network outperformed the simple thresholding model—and ground truth—at predicting human responses. We then used a genetic algorithm to selectively delete connections within the neural network to identify variants of the network that approximated human judgments even more closely. The best resulting network shared 68% of the variance with human judgments – more than the unpruned network. As a first step towards interpreting the network, we then used representational similarity analysis to compare its inner representations to a wide variety of hand-engineered image features. We find that the network learns representations that are similar to directly image computable predictors, but also to more complex predictors such as intrinsic or geometric factors, as well as some indications of photo-geometrical constraints learned by the network. However, in a lesion analysis we find these not to be important in order for the network to respond similarly to humans.

3.1 Introduction

Humans easily perceive and distinguish materials visually. One important optical aspect of materials is glossiness. It is useful in determining whether a piece of food is fresh, the floor is slippery or whether a surface is clean or greasy. Arguably the most important image feature for gloss perception is the presence of highlights (Beck & Prazdny, 1981; Fleming, 2012; Marlow et al., 2012) – direct reflections of light sources or other bright elements in the environment.

The exact computations underlying human perception of highlights and gloss remain poorly understood. Factors other than the physical reflectance properties of a material itself, such as shape or illumination can impact perceived gloss (Doerschner et al., 2010; Fleming, 2012; Fleming et al., 2003; Ho et al., 2008; Olkkonen & Brainard, 2011; Wijntjes & Pont, 2010). This has been taken to indicate that the human visual system does not accurately estimate the physical reflectance of surfaces, but rather arrives at a subjective impression of gloss through the use of heuristics, in which properties of highlights, such as shape, contrast and size play an important role (Fleming, 2012; Marlow et al., 2012). The importance of highlights is well known (Beck & Prazdny, 1981; Todd et al., 2004), but it remains unclear what exact computations the visual system uses to recognize highlights—i.e., to distinguish specular highlights from other bright patches in the image, such as texture markings. The perception of materials and glossiness is typically associated with mid-level vision (Anderson, 2011; Fleming, 2014; Komatsu & Goda, 2018) in which low-level image features such as edge orientation, color, brightness and gradients or scale are pooled and compared to arrive at surface representations. Several studies have shown the importance of 3D surface representations and the need for specular highlights to be congruent with surface geometry and shading patterns to elicit a perception of gloss (Anderson & Kim, 2009; Beck & Prazdny, 1981; Kim et al., 2011; Kim & Anderson, 2010; Marlow et al., 2011; Todd et al., 2004). Anderson and Kim (2009) showed that images in which the highlight component has been rotated with respect to the matte component are less likely to be perceived as glossy. Marlow et al. (2015) and Marlow and Anderson (2015) have demonstrated how identical image gradients can be interpreted as blurry highlights on a glossy surface or shading on a matte surface depending on perceived 3D surface structure. Yet, despite progress on many of the factors that influence the identification and interpretation of highlights, there is still no image-computable model that emulates human judgments.

To address this need, in this study we took a big data approach to modelling highlight perception. Specifically, we sought to develop a model that distinguishes whether bright markings in images of surfaces appear as highlights rather than texture or some other non-

highlight feature. We used machine learning as a method that allows us to train a model on thousands of images with random variations in geometry, texture and illumination, to capture those features that are useful for identifying highlights over a wide range of surfaces and appearances.

This is essentially an 'Intrinsic image decomposition' task—a well-known problem in the computer vision literature that has been studied since the late 1970s (Barrow & Tenenbaum, 1978). For a recent review see Bonneel et al. (2017). Most computational models on highlight detection however focus on removing highlights as they interfere with identifying other intrinsic components such as shading or surface reflectance. Here, we focus not on a full decomposition into intrinsic components, but specifically on isolating the specular component of images. Importantly, rather than solving the engineering goal of identifying highlights as accurately possible, we focus on reproducing the pattern of behavior that humans exhibit—both successes and failures.

Previous work comparing convolutional neural networks (CNNs) to humans has shown both striking similarities (Gomez-Villa et al., 2018; Ward, 2019; Watanabe et al., 2018) but also discrepancies where CNNs react very differently from humans to slight changes in a stimulus (Kurakin et al., 2019; Nguyen et al., 2015; Sharif et al., 2016; Szegedy et al., 2014), show different generalization behavior to humans (Nguyen et al., 2015), or have difficulties solving visual tasks that are very simple for humans (Stabinger et al., 2016). Networks are often susceptible to being fooled by specific small changes that are almost imperceptible to humans called adversarial attacks (Szegedy et al., 2014) and CNNs performance often decreases catastrophically when confronted with degraded stimuli (Geirhos, Temme, et al., 2018), unlike humans. In addition to picking up on pixel artifacts, there are also reports of CNNs learning different cues and mechanisms than humans. For example, CNNs have been found to make different use of scene context than human observers, outperforming humans in recognizing objects from their backgrounds only (Zhu, Xie, & Yuille, 2017). Geirhos, Rubisch, et al. (2018) found that CNNs trained on the ImageNet classification task (Russakovsky et al., 2015) tend to rely heavily on texture rather than object shape.

As one approach to mitigating these tendencies, here we use pruning as a method for fine tuning a trained neural network to make it respond more similarly to humans. Although many other approaches are possible, pruning is straightforward and does not require enough human data to train a network from scratch. It has been used for over three decades as a method for reducing network complexity and computational requirements (Janowsky, 1989; LeCun, Denker, & Solla, 1990; Mozer & Smolensky, 1989). In simple terms, the rationale for pruning

is that a complex network will typically contain both necessary and unnecessary units. Identifying and pruning unnecessary units can reduce network complexity while retaining high performance. There is usually a trade-off between network performance and reducing network complexity. The exact criteria for evaluating the importance of single units and the pruned network overall are a subject of much research (see Blalock, Ortiz, Frankle, & Guttag, 2020 for a recent review). It has been observed that pruning can improve generalization performance (Bartoldson, Morcos, Barbu, & Erlebacher, 2020; Hassibi & Stork, 1993; LeCun et al., 1990) and in some cases that pruned networks outperform the original unpruned networks (Han, Pool, Tran, & Dally, 2015; Suzuki et al., 2020) on the training objective.

Yet here, we use pruning as a method for optimizing the network functionality not in terms of the original training criterion, but to *human responses* on the same task. We expect that a neural network trained to identify physically accurate highlights will learn an approximate solution that includes features similar to those the human visual system uses, but also includes different features. We therefore hypothesized that in a second fitting stage we can prune the trained network to identify a variant that emphasizes the similarities to human responses while de-emphasizing the differences. Given this pruned network, we can also test whether it is possible to gain insights into the processes that make the model respond similarly to human observers.

To do this, we created a large dataset of 164 085 images containing glossy surfaces with varying textures or without texture. To limit the number of human responses required, we investigated the similarities in predictions only in certain pixels that we expected to be particularly informative. To identify these pixels, we used two extreme predictions as baseline models. One is a very simple model that we trained to learn a global intensity threshold value for classifying brighter points as highlights. Although crude, such a heuristic can correctly identify highlights in many conditions. However, as it lacks any knowledge of surface or image geometry, it can also be readily fooled by bright texture markings. The other extreme was the physical ground truth from our rendering simulations, representing a physically correct solution—the upper limit on performance that any observer system could achieve. We expect the human visual system to make more sophisticated decisions than an intensity threshold but also simpler inferences than a fully accurate inverse physics estimate. We therefore expect to find informative image locations that are particularly descriptive of human highlight perception where our two baseline models contradict each other. We chose image locations based on these two predictors, including some where both models agree, to check whether our stimuli and experimental method yield meaningful perceptual responses in conditions where we have a

clear expectation of subjects' responses. We ran two parallel experiments, the stimuli being constructed and chosen according to the same principles but chosen from a different set of images.

To anticipate, we find that observers respond to pixels where both models agree mostly in line with model predictions. For the two disagreement categories, subjects' responses are mixed, and neither ground truth nor the threshold model better predicts human highlight perception. This provides us with a promising basis to develop a better model of human perception.

We trained a novel CNN architecture (see **methods**) to predict for each point in the image whether it contains a highlight or not, using supervised training with the ground truth rendered specular component of the images as a label for each pixel. This model predicted human responses on the probe locations better than ground truth or the threshold model. To further fit our model to human responses we pruned network connections using a genetic algorithm in order to identify those configurations of pruning that maximize correlations to humans on the target dataset. A genetic algorithm allows us to test model fitness of various configurations of several units being pruned at the same time.

We find that indeed a large set of pruned configurations correlate higher with human responses than the full network. We examined one representative subnet in detail, investigating where the differences to the full network lie. We also conducted a representational similarity analysis (RSA; Kriegeskorte et al., 2008) to compare intermediate layers of the pruned network to various candidate predictors. We find that representations within the network are similar to various simple and more complex predictors suggested in human gloss perception literature. We also demonstrate that our network is weakly sensitive to violations of photo-geometrical constraints of highlights. In a lesion analysis however, we find no evidence that neurons with high similarity to geometric predictors are more important than other neurons for the model to react like human observers.

Our main result is that we have developed, the best of our knowledge, the first image-computable model that predicts human highlight judgments better than ground truth. While the network learned representations related to photo-geometrical predictors, surprisingly we find that these features are no more important than simpler image-computable features for the model to respond similarly to humans. Our findings also show an application of pruning neural networks based on a relatively small human dataset as a method for fine tuning neural networks that were previously trained on simulated physical data. This method could be beneficial in

other research areas where machine learning would be useful, but target data is slow or costly to obtain while simulated data is more readily available.

3.2 Methods

3.2.1 Training data and stimuli for experiment with human observers

To train and test our networks and to test human observers we rendered 164 085 grey scale images of glossy surfaces like those in **Figure 7**. The image size was 256×256 pixels. Every image was filled by a surface viewed at 45° and perturbed by waves and illuminated by a square light source parallel to and at a random position above the surface. Images were rendered in Blender using the "Cycles" render engine. We used 4 spatial scales of surface geometry. For every surface geometry we rendered a 'plain' surface consisting of an untextured specular and diffuse component combined, as well as 12 texture conditions – 4 spatial scales each of Voronoi patterns, marble patterns and checker patterns. In order to create texture patterns that could reasonably be confused with the highlights, we also rendered two versions of each scene where we used the specular reflections from a randomly chosen different surface as a texture map by multiplying it with the diffuse component (see examples in **Figure 7**, under 'false highlights'). We chose these conditions to include textures that consist of highlight-like patches in positions and orientations that are incongruent with the surface, as Anderson and Kim (2009) have shown that matte surfaces with physically correct highlights imposed at a wrong orientation or position are less likely to be perceived as glossy. Our intention was that a network trained on these stimuli would be forced to learn something about highlight positioning (rather than just intensity), possibly yielding a wider range of strategies that could show up during the pruning stage of our network fitting.

Of the rendered images, 7200 were withheld from network training, balanced for surface scale and patterns, leaving 156 885 training images. The withheld images were used as candidates in selecting images and probe locations for the first human experiment and as a validation set during training. We also rendered a second set of 15 000 images constructed the same way as the initial dataset. These images were also not shown during training. They were used to select the images and probe locations for the second human experiment and as stimuli for additional network analysis. The complete set of images will be made available for download from Zenodo upon acceptance.

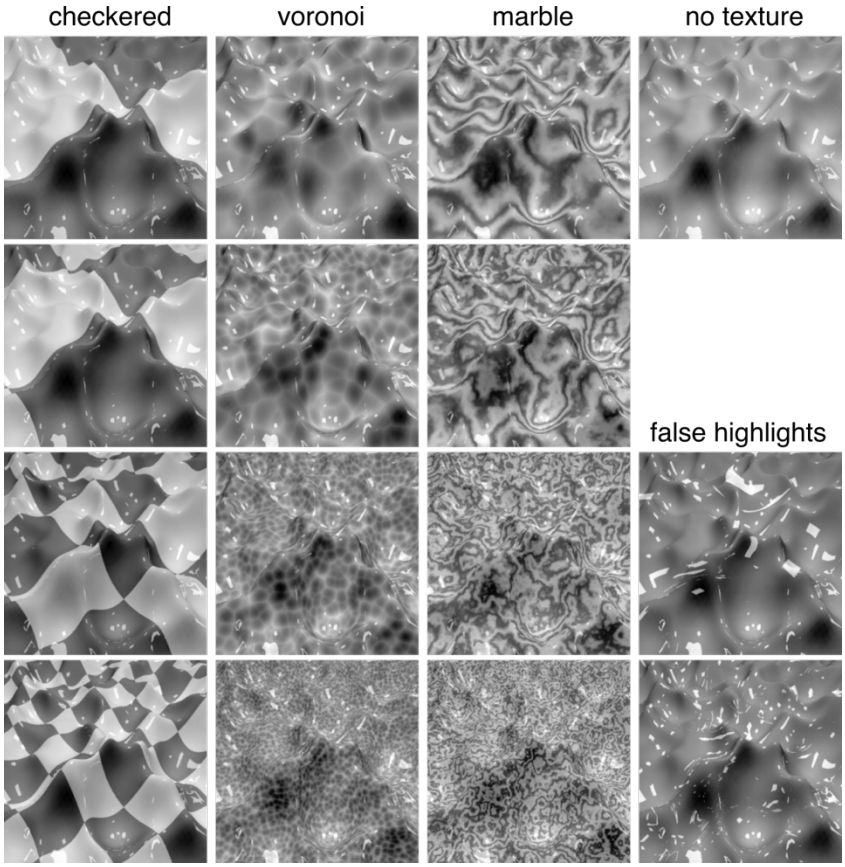

Figure 7: *Example images showing all different texture conditions for one scene*

3.2.2 Experiment with human observers

Human observers were asked to judge specific pixels in an image and to respond whether the pixel contained a highlight or not. To make these pixelwise responses more informative, we did not select the locations at random (also because only about 3% of all pixels contained highlights). To select probe locations (see Figure 8 for an example), we trained a simple model that used a single global intensity threshold to identify highlights and obtained a prediction for each of 7200 test images. We also looked at the specular map (ground truth) for these images. From these two response maps we could sort each pixel into one of four categories: (a) Both ground truth and the threshold predictor agree there is a highlight, (b) threshold predicts a highlight but there is no highlight according to ground truth, (c) there is a highlight but threshold does not predict one, and (d) threshold and ground truth agree that there is no highlight. For each category we chose one pixel per image that maximized this function. For 120 images we chose one probe for each of the four categories and for another 120 images we chose only one probe for categories b and c where threshold and ground truth disagreed. These are likely to be more informative in discriminating between predictors than pixels where a simple model already agrees with ground truth. Pixels from category a and d provide a useful baseline to

determine whether our method of single-pixel judgements yields expected results for easy stimulus conditions and to confirm that observers perceive highlights in our stimuli as such.

This way, we constructed two test sets. For the first—which we refer to as the *target set*—the images were chosen randomly from a pool of 7 200 candidate images. For the second test set, which we used as a validation set, images were chosen from a pool of 15 000 images. For both sets, images were chosen in such a way as to balance surface scale and texture conditions. In both cases, we tested subjects on 720 probe locations on 240 images, showing images in random order. Each trial consisted of a display of the image, with the current probe location marked by a red cross. Next to the image was a second display of a closeup of a 32×32-pixel image patch, magnified ×8. **Figure 8** shows an example display. By default, this was centered on the probe location but could be moved using the mouse. The cross indicating the probe pixel could be toggled on and off. Subjects were asked to judge the central pixel and respond with one of two keys whether this pixel contained a highlight/reflection or texture.

13 participants aged 20 – 39 (mean = 25.9) took part in the first experiment. In the second experiment 15 observers aged 19 – 33 (mean = 25.0) participated. They all had normal or corrected to normal vision and signed informed consent according to the declaration of Helsinki. The procedures were consistent with those approved by the local ethics committee of the Psychology Department at the Justus Liebig University of Giessen. Both experiments lasted about an hour.

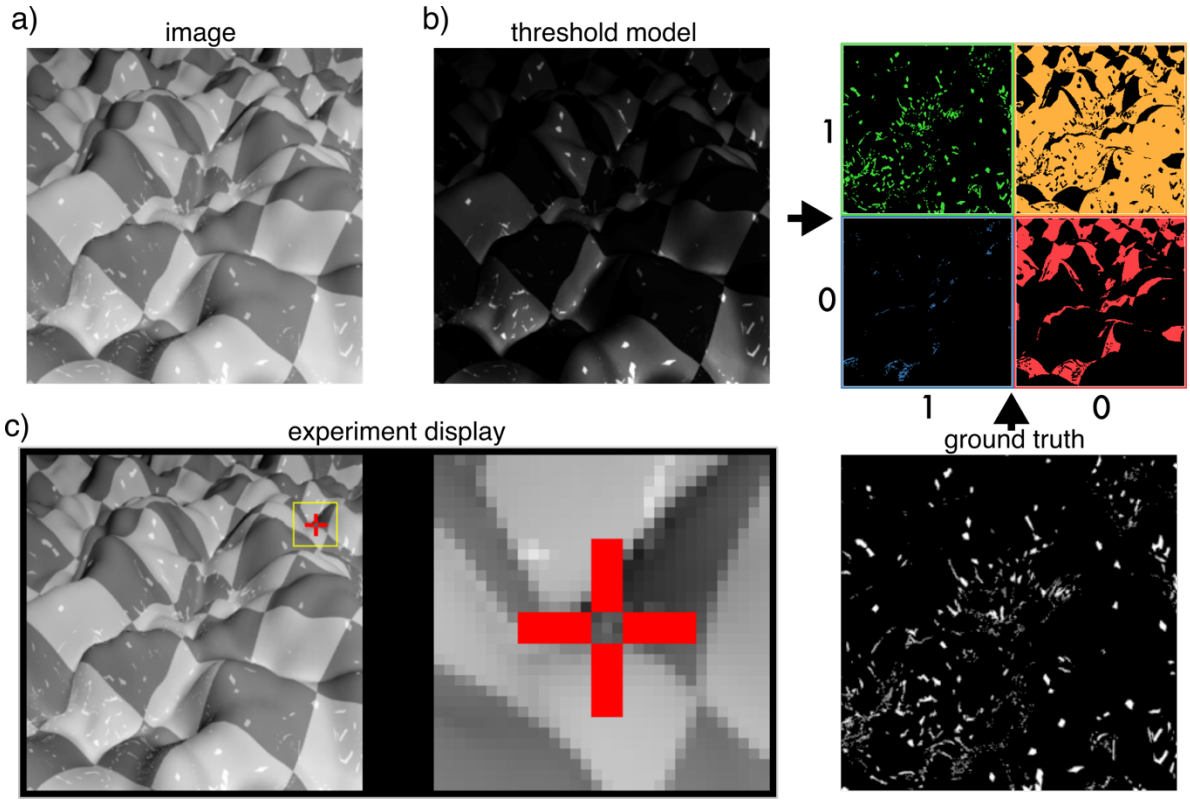

Figure 8: (a) *An example image as shown to the network and observers.* ***(b) Left:*** *Predictions by a simple threshold model (TM) of the image in a).* ***Bottom Right:*** *Ground truth specular component (GT) used as labels during network training* ***Top Right:*** *The four categories from which probe locations for experiments with human observers were chosen. Green: TM correctly identifies a highlight at these locations. Yellow: TM wrongly identifies a highlight. Blue: TM fails to identify a highlight. Red: TM correctly identifies no highlight.* ***(c)*** *Example of the display seen by observers during the experiment. Subjects saw an image with one pixel marked by a red cross and a magnified close up (x8) of a 32x32 pixel patch which they could move with the mouse. The location mark could be toggled on and off*

3.2.3 Ground truth and threshold model predictions

The ground truth information used for choosing probe locations for the experiments with human observers was obtained as a rendering of the untextured surface in a material reflecting only specularly in Blender. For the threshold predictor we trained a model to determine a global threshold for the 156 885 images that we also used for training the neural network. This model was implemented in tensorflow as a neural network with one convolutional layer consisting of a single pixel filter and a ReLU output. A second layer with a sigmoid activation function mapped the activations into the displayable intensity range. This way the model prediction was only based on each pixel's individual intensity – regardless of neighbours and context. The model contained a threshold below which predictions were 0 and otherwise gave a continuous prediction between 0 and 1. We trained the model for 50 epochs using binary cross entropy as

the objective loss function. This results in near-optimal threshold for approximating the ground truth.

To categorize candidate pixels into categories a-d described earlier, we used binary data from each predictor (whether a predictor gave a zero or non-zero prediction to each pixel). To choose pixels within each category we used the continuous prediction values to maximize the function for each category. For example, for category b (threshold predicts a highlight where there is none) we picked out of all candidate pixels in an image that pixel for which the threshold model gave the highest prediction. For an example of the model outputs and the probe location selection see **Figure 8**.

3.2.4 Network architecture

The network architecture we used is shown in **Figure 9a**. It was designed to give the network capabilities of performing image computations at different spatial scales and to exchange processing results between scales between layers. The network consists of four tiers of parallel convolutional layers. The first three of these consist of 7 parallel convolutional layers each – receiving input at different scales from 1/1 size to 1/64 size. Each of these contain 8 convolutional filters. Between the first and second, and second and third tiers there are scaling layers so that the output of all layers operating at different scales is scaled up and/or down in order that each of the parallel layers receives input from each of the parallel layers in the previous tier. The fourth tier consists of one layer with 8 convolutional filters at full image scale, after which comes an output layer of one channel.

3.2.5 Network training

We performed feedforward training on our network, using the specular maps as labels and with binary cross-entropy as our loss criterion. Training lasted for 50 epochs.

3.2.6 Network pruning

Pruning was performed on the trained network. Rather than pruning individual neurons, we deleted connections between the first, second and third tiers of parallel layers, where each layer contained 8 convolutional filters. Each of these tiers consists of 7 parallel convolutional layers processing the image at different spatial scales, each of which feeds into each of 7 parallel nodes in the subsequent layer. This amounts to 98 connections between the layers in the first and second, and second and third tiers.

To perform and evaluate pruning we used a genetic algorithm. Each network configuration can be described by a 98-parameter vector, where each parameter can either be a 1 (connection on) or a 0 (connection off). The ‘full network’ refers to the network configuration

vector of only 1s, i.e., where all connections are active. During training all connections were active. We started each run of the algorithm with a population of 99 random configurations and the full network. We applied each configuration to the trained network, obtained predictions for the test images and took the predictions for those pixels that we showed to humans in our experiment. From the response vector on these pixels, we calculated a correlation to humans and used it as a fitness criterion in the genetic algorithm. After calculating the fitness for an entire population (i.e., 100 configurations) we kept the 10 fittest members as survivors, added a mutated (5% chance of a switch between 0 and 1 for each connection) copy of each survivor to the next generation and added 80 configurations through cross-over from members of the previous generation (chosen randomly, weighted by fitness, 1% mutation chance). We repeated this process for 30 generations and ran 300 instances of the genetic algorithm.

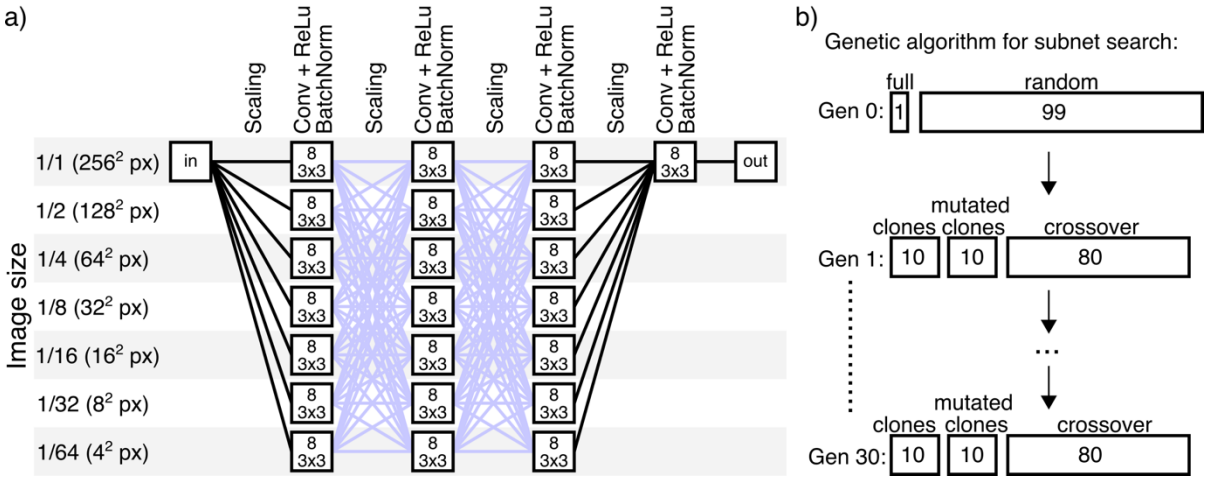

Figure 9: (a) *Network architecture. The first three tiers consist of parallel convolutional layers processing at different resolutions from full resolution to 1/64 resolution. The connections shown in blue are the ones that were subject to the pruning search.* ***(b)*** *A schematic of the search algorithm. The initial generation consisted of the full network and 99 random subnets. Each following generation consisted of 10 clones of the fittest networks of the previous generation, one mutated version of each of these 10, and 80 subnets created through crossover from two members of the previous generation, picked according to fitness. Fitness was defined as the correlation to humans on the target set.*

3.3 Results

3.3.1 Humans

Human observers were asked to judge whether 720 individual probe locations contained a highlight or not. These probe locations (single pixels) were picked based on ground truth (GT) and a threshold model (TM) that used only an intensity threshold to predict highlights. **Figure 8** illustrates the probe location selection process.

Mean responses from Experiment 1 grouped by the categories of probe locations are shown in **Figure 10a**. Results from the first experiment were later used as a target for our pruning algorithm and will be referred to as the *target set*. Mean responses from Experiment 2 are shown in **Figure 10b**. The probe locations in this experiment were selected from a different set of images according to the same criteria as the target set. Results from the second experiment were used for validation and will be referred to as the *validation set*.

Results show, as expected, that pixels that contain highlights and are brighter than the threshold (category a, see **Methods**) are most likely to be classified as highlights. Similarly, pixels without a highlight that are darker than the threshold (category d) are least likely to be classified as highlights. Interestingly, pixels from categories b and c (that either contain a highlight but are darker than the threshold or contain no highlight but are brighter than the threshold) are on average similarly likely to be judged as a highlight. This suggests that sheer relative pixel intensity does have an impact on human highlight perception, but that further factors play a role.

The pattern of results for the four pixel categories is very similar for both experiments. It shows that human observers perceived highlights in our stimuli, and that they were able to interpret and respond to single pixel probe locations. Both ground truth and threshold predictions seem to partially predict mean human responses equally well (correlation to mean human responses $r = 0.57$ and $r = 0.58$ for the target dataset and $r = 0.51$ and $r = 0.49$ for the validation dataset respectively). As a comparison we calculated the intercorrelation between human observers. Since human responses were binary, we randomly divided the observer group in two 10 000 times, correlating the mean responses of the two groups every time. The maximum correlations we observed were $r = 0.82$ for the target dataset and $r = 0.69$ for the validation dataset (mean correlations were $r = 0.73$ and $r = 0.57$ respectively). **Figure 11** shows the distributions of these human-to-human correlations. Human responses are highly idiosyncratic, since a large proportion of the variance in human responses remains unexplained by other observers.

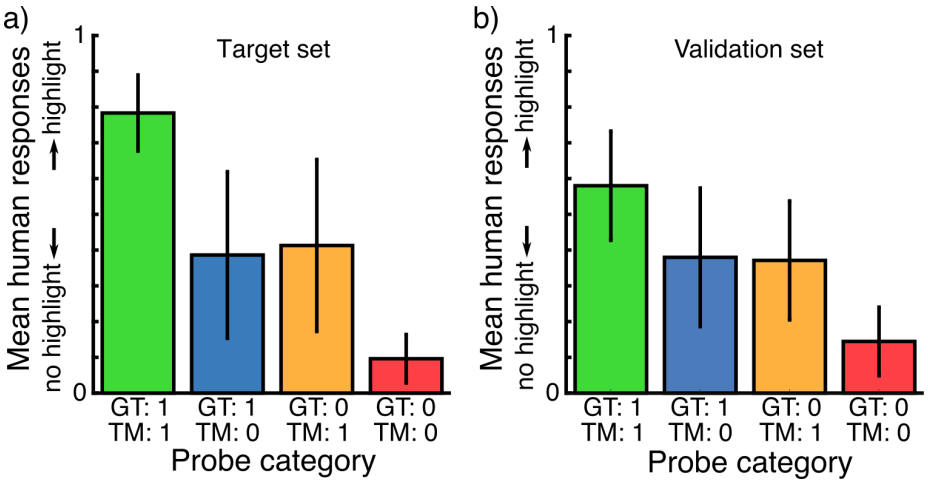

Figure 10: (a) *Mean results from 13 human observers on the target set – one point from each category described in* ***Figure 8b*** *selected for each of 120 images. One point from only blue and orange categories (disagreement categories) for another 120 images.* ***(b)*** *Mean results from 15 human observers on the validation set. A Second set of probe locations and images constructed and selected according to the same criteria as those in the target set*

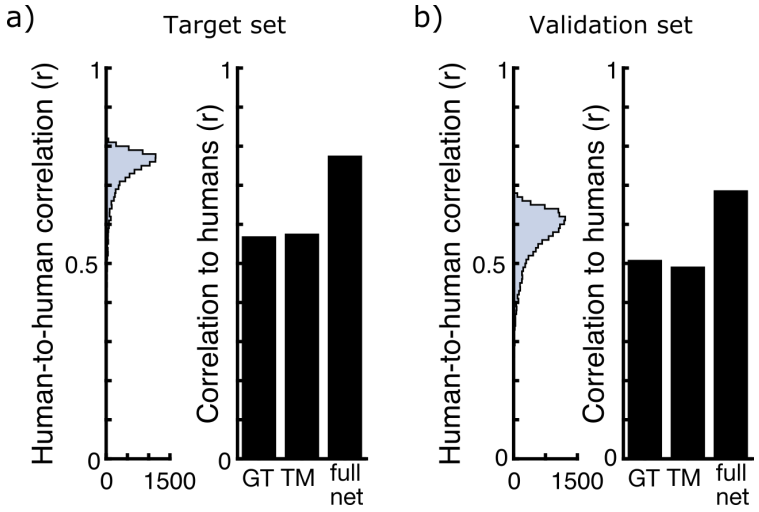

Figure 11: (a) Left: *Histogram of human-to-human correlations between responses to the target set calculated by randomly splitting the group of participants in two 10 000 times and correlating their responses.* ***Right:*** *Correlation of the threshold model (TM), ground truth (GT) and full network to the mean of human observers on the target set.* ***(b) Left:*** *Histogram of human-to-human correlations between responses to the validation set calculated by randomly splitting the group of participants in two 10 000 times and correlating their responses.* ***Right:*** *Correlation of the threshold model (TM), ground truth (GT) and full network to the mean of human observers on the validation set*

3.3.2 Network

The full architecture of the network is shown in **Figure 9a**. There are four tiers of convolutional processing. The first three of these consist of 7 parallel layers of 8 filters each, that perform convolutional processing at different scales of the image ranging from 1/1 scale to $1/2^6$. In between these are a number of parallel scaling layers, such that every one of the parallel convolutional layers receives the output from all processing scales as input. The fourth tier consists only of one convolutional layer of 8 filters that processes the image at full scale. We compared predictions by our full network as well as our threshold model (TM) and ground truth (GT) values to human responses (see **Figure 11)**. The full network after training correlated higher than TM or GT to the mean of human observers on the target set ($r = 0.78$) and on the validation set ($r = 0.69$).

3.3.3 Pruning

We then sought to improve further the fit of the network to human performance through pruning. We searched for pruned configurations of our network that responded more similarly to humans using a genetic algorithm. The genetic algorithm searched through configurations (different combinations of on / off settings) for the connections between the first and second, and second and third layers, shown in blue in **Figure 9a**.

The results of pruning in terms of correlation with humans on the target pixel set are shown in **Figure 12a**. While many pruning configurations reduced correlation to humans, in every run, the genetic algorithm discovered pruned configurations of the network that responded more like humans than the full network. The highest correlation between a pruned version of the network and humans was $r = 0.83$, at the upper limit of human-to-human correlations that we observed. **Figure 12b** shows subnet similarity to humans for all runs plotted against the number of active connections. Interestingly, there is a large range in the number of connections that are active in configurations that improve the network's correlation with humans.

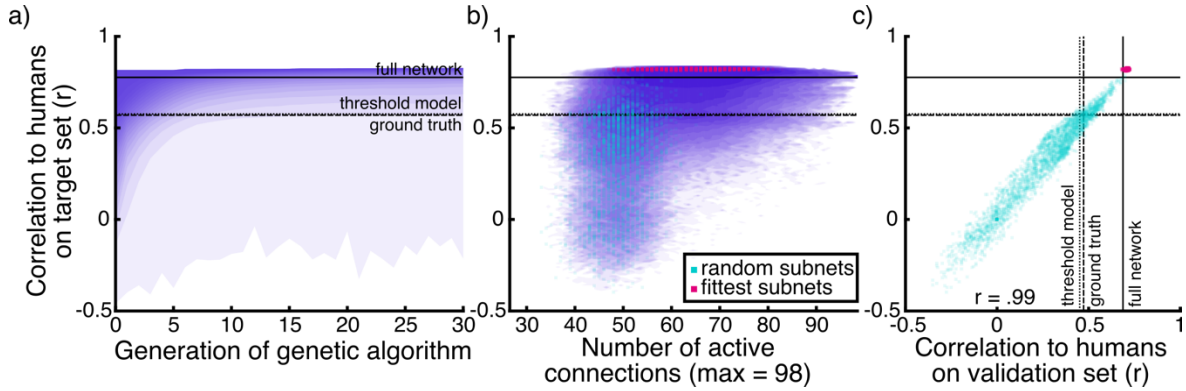

Figure 12: (a) *Distribution of correlations to human responses on the target set in each generation, summed across 300 runs of the genetic algorithm. Correlations between humans and ground truth, threshold predictor and the full network are plotted as lines.* ***(b)*** *Distribution of correlations to human responses on the target set plotted against the number of connections active in each subnet. Plotted in green and pink are 10 random and the 10 fittest subnets from each run of the genetic algorithm (3000 in total each) that are used in e) and other subsequent analyses.* ***(c)*** *Correlations of a subset of 3000 random and 3000 fittest subnets to human responses on the target set and on the validation set. Correlation between the fit on these two datasets is r = .99*

3.3.4 Validation performance

To validate whether pruning according to our target dataset yields generalizable results, we compared responses from a subset of the configurations against humans on the validation dataset. We picked the 10 fittest configurations from the last generation and 10 random configurations from the initial generation from each run of the genetic algorithm, resulting in 3000 fit and 3000 random configurations in total (shown in **Figure 12b**). The validation performance of these pruned versions of the network can be seen in **Figure 12c**, along with the validation fitness of ground truth (GT), threshold model (TM) and the full network (r = 0.51; r = 0.49; r = 0.69 respectively). The maximum correlation of a pruned network to humans on the validation set we observed was r = 0.73. Correlations of the pruned networks with humans on both datasets correlate at r = 0.99. This indicates that the similarity in responses to humans of the fit configurations is not limited to images in the target set. Of the 3000 fit configurations, all correlated with humans higher than the full network on the target set, and 21 did not improve on the full network's correlation with humans on the validation set.

3.3.5 Example pruned network

For the following analysis we picked a candidate pruned configuration of our network. Our choice was based on four criteria: (1) variance in human observer data explained (R^2), (2) variance in ground truth data explained (R^2), (3) variance in human errors explained by network errors (R^2 between the differences of prediction and ground truth and human responses and

ground truth), and (4) lowest RMSE to human observers. We rank-ordered the 3000 fit network combinations used previously in the validation analysis according to each criterion and picked the configuration that showed the lowest sum of ranks. The predictors were all highly intercorrelated (all r = 0.82 or more). We chose to use several selection criteria to avoid possible outliers, e.g., configurations that correlate well with humans but show very weak responses close to 0 (high correlation but high RMSE to humans). The network we picked according to the four criteria was also the third fittest in terms of R^2 to humans on the target set.

3.3.6 Differences to full network

To better understand what the pruned network does differently than the full network we compared network responses to human responses in more detail. We looked at both networks' responses to different spatial scales of surfaces. The correlation to human observers separated for surface perturbation size (**Figure 13a**) shows that the pruned network has a tendency to better predict human responses than the full network for all surface scale conditions. This improvement shows in the target and validation datasets. As a test of statistical significance we calculated the 95% confidence intervals for the difference between each pair of correlations (r between the pruned net and humans – r between the full net and humans) as suggested by Zou (2007). For the subsets per surface scale of the target dataset these were [0.03,0.09], [0.05,0.12], [0.02,0.08], [-0.01,0.06] respectively from smallest to largest perturbations (left to right in **Figure 13a**). For the subsets of the validation dataset the 95% confidence intervals were [0.05,0.12], [0.01,0.07], [0.00, 0.07], [-0.03,0.07] respectively. In the target dataset the pruned net correlated significantly higher than the full net with humans for all perturbation categories, except for the image category with largest perturbations. In the validation dataset the pruned net correlated significantly higher than the full net with humans for the two image categories with smallest surface perturbations. Where correlations were not significantly different, the observed direction of the difference was also that the pruned net correlated higher with humans than the full net. This indicates that the improvement of the pruned network is larger for images of surfaces with smaller perturbations, but not limited to specific spatial scales and includes a mechanism that is applicable over a wide range of geometries.

In another step we looked at the networks' responses to different categories of pixels to test how the pruning affected sensitivity to highlights (i.e., to answer the question whether pruning elicited a criterion shift). Specifically, we divided the pixels in every image into three groups: specular, bright texture patches (excluding areas that overlapped with specular), and other pixels. For 15 000 images we summarized the network responses as the mean response per pixel category. This result is shown in **Figure 13b.** Our pruned network appears to give

higher responses than the full network to pixels of all categories, but to different degrees. The largest increase in mean glossiness rating is for pixels containing a highlight with a lesser increase for texture pixels and a very small increase to pixels that contain neither highlights nor bright texture patches. In other words, the pruned network seems to make a criterion shift compared to the full network, that raises the responses to correctly identified highlights but also (to a lesser degree) to bright texture pixels.

This seems to suggest that the pruned network makes more or stronger true positive decisions at the cost of an increased false positive rate. Since the predictors do not make binary decisions, but rather give continuous predictions, we visualize this as the rate of true positive and false positive decisions if a certain predictor value served as the threshold for binary decisions. **Figures 13c** and **d** show network predictions of the 720 pixels in the target dataset in this way, compared to mean human ratings. The pruned network does indeed favor true positive, but also false positive decisions compared to the full network. In doing so the pruned network shows more similar behavior to the mean of human observers, where this tendency appears to be even more pronounced.

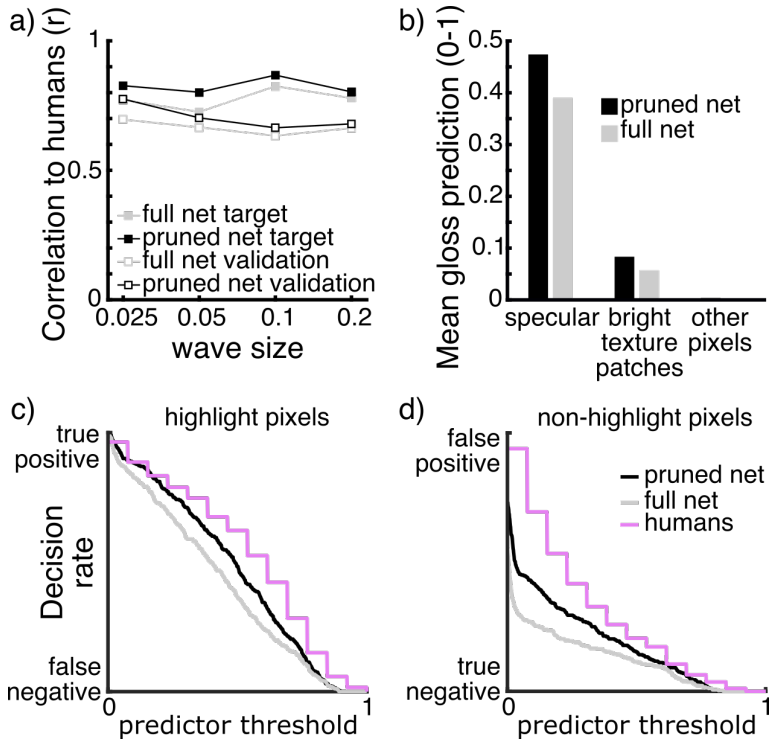

Figure 13: (a) *Correlation to humans of the full (black) and pruned (grey) network on the target and validation datasets (filled and empty squares respectively).* ***(b)*** *Mean predictions for all pixels in 15 000 images sperated into pixels that contain a highlight, bright texture patches or neither.* ***(c)*** *true positive and false negative decision rates for highlight pixels from the target dataset for the full network, pruned network and human observers if different predictor values act as a threshold in a binary decision.* ***(d)*** *False positive and true negative decision rates for non-highlight pixels from the target dataset*

for the full network, pruned network and human observers if different predictor values act as a threshold in a binary decision.

3.3.7 Network predictions for stimuli with modified highlights

To further investigate similarities between the pruned network and humans we constructed a set of images containing highlights that we manipulated in the image space. Specifically, we altered the *global rotation* of the entire specular component of the images in angles of 0°, 90°, 180° and 270. Anderson and Kim (2009) have shown that surfaces with displaced highlights are less likely to be perceived as glossy than when highlights are in their correct positions. These test images contained no texture. We fed the same 1000 images with every rotation condition through our pruned network and compared responses to the (manipulated) highlights contained in the images. We used RMSE per image between the network predictions and the highlight component (correctly or wrongly oriented) of the images as a measure of how much the network responses match the highlights. Lower RMSE values indicate a stronger tendency to recognize the manipulated highlights as highlights.

The RMSE between pruned network predictions and rotated highlight components is shown in **Figure 14**. Global rotations affect the RMSE, but to a very small degree. The original, non-rotated highlights produce the lowest RMSE indicating that globally rotated highlights are less likely identified as highlights by the network. It is important to note that while the direction of this effect is in line with what we expect from a human-like model, it is also very small. For comparison, the RMSE between 1000 pairs of random noise images of the same size as our stimuli is 0.412 (sd = 0.007). This means that most of the manipulated highlights are still mostly recognized as highlights, only slightly less than the original highlights. Interestingly, this seems to indicate that a highlight detection model can behave much like a human observer in conventional images, without being very sensitive to photo-geometrical constraints.

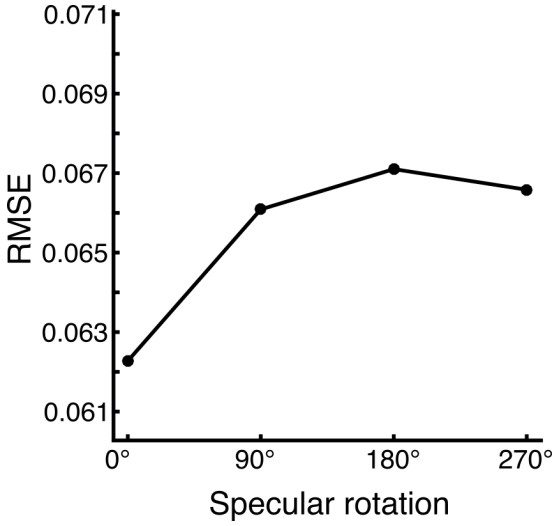

Figure 14: *RMSE between predictions by the pruned network and the ground truth specular component for 1000 untextured images with differently rotated specular components. A specular rotation of 0° indicates the correct alignment.*

3.3.8 Learned Representations

To investigate what kinds of representation the pruned network has learned, we compared the responses of units to a number of 'hand-engineered' image and surface descriptors ('predictors') in a representational similarity analysis (RSA; (Kriegeskorte et al., 2008). We calculated representational dissimilarity matrices (RDMs) consisting of the pairwise Euclidean distances between pairs of images as they are represented in the output of every convolutional neuron throughout the pruned network, as well as by the different predictors. RDMs were calculated based on 4500 images including all the texture conditions we used in the training set. We included 34 possible predictors ranging from image computable local filters and summary statistics to more complex geometrical and 'intrinsic image' (Barrow & Tenenbaum, 1978) components. We grouped the predictors into seven categories (for detailed descriptions see **Appendix B**):

Input Image: the original grayscale input image

Summary statistics: mean, sd, skewness and kurtosis of image intensity

Edge detection / direction: pixel gradients in x- and y-directions, local contrast, locally normalized image

Image gradients / anisotropy: orientation of the smoothed image gradient in x and y, anisotropy of the smoothed image

Geometry information: camera distance, angle to camera, light distance, angle to light source, convexity, pointiness (the magnitude of local curvature, as calculated by Blender), x normal, y normal, z normal, occluding edges, distance from occluding edges
Intrinsic components: texture, matte (shading), specular, specular direct, specular I ndirect, specular coverage, texture coverage
Scene information: surface scale, texture type, texture condition, scene

We correlated the top triangles of unit and predictor RDMs. **Figure 15a** gives an overview of where in our pruned network the internal representations show greatest similarities to each predictor category in terms of maximal total variance explained. Every layer contains 8 filters; shown in **Figure 15a** is the maximum per layer that any of these filters is explained by all predictors in a given category. In addition to the predictors mentioned above, we compared network representations to representations according to the texture analysis model by Portilla and Simoncelli (2000), also shown in **Figure 15a**. This is another 'hand engineered' model of mid-level vision that has very successfully been used in texture analysis and synthesis.

The results of the RSA broadly agree with the expectation that later network stages are associated with more complex representations. Simpler, directly image computable predictors tend to show similarities earlier in the network, becoming less relevant towards the output, while similarities to more complex predictors like geometric or intrinsic parameters emerge only late in the network. Summary predictors show greatest similarities to units that process lower resolution, more spatially summarized representations of the image. The low similarity to predictors from the *scene information* category means our network has likely not learned the categorical factors by which we constructed the dataset. Portilla-Simoncelli statistics are most similar to units in the first and second tier, especially those operating on spatially summarized representations. This indicates that our model indeed makes use of features similar to hand-engineered mid-level features. The placement of similar neurons in the network suggests that these features occur at intermediate processing stages, with other features in later layers computed based on them.

Figure 15b shows an arrangement of the individual units of the pruned network according to the similarity of their response characteristics. Specifically, we computed a second-order RDM of pairwise correlations of top triangles of first-order RDMs, and then visualized the similarities in two-dimensions using multi-dimensional scaling (MDS). Each point represents a single unit, colored according to the categories of their single most similar predictors. This reveals broad clusters of qualitatively similar units. Most filters show highest

similarities to the original input image or to image computable predictors from the categories *edge detection / direction* and *image gradients / anisotropy*. A small number of filters (the network output among them) is most similar to geometric predictors or to intrinsic image components. In the case of the output layer, this presumably reflects the objective function, which was to return a per pixel highlight map (i.e., an intrinsic image of the specular component). Thus, as expected, units that correlate with relatively high-level factors such as surface geometry or intrinsic images tend to be more prevalent in later stages of the network.

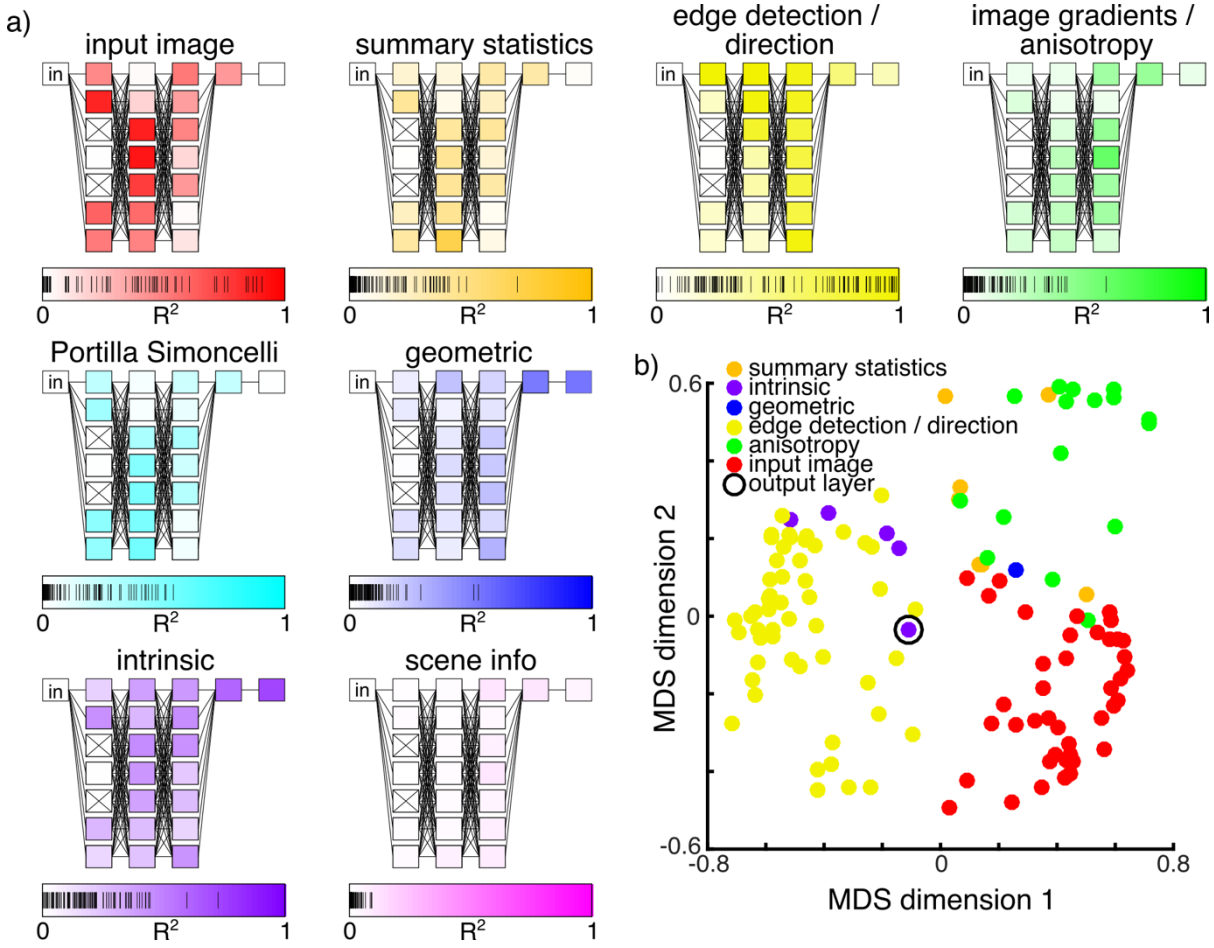

Figure 15: *(a) Representational similarities between single filters of the pruned network and predictor categories. Shown are the maximum variance explained of any filter's first order RDM (upper triangle) per layer (out of 8 each) by all predictors in a given category. Layers marked with an x contain only 'dead' filters. Markings on the scales below each network schematic show the variance explained (R2) of all network filters' representations by the predictors in the category. (b) The first two dimensions of a multidimensional scaling (MDS) representation of all active filters in the pruned network. MDS is calculated on the pairwise similarities between first order RDMs of representations of 4 500 images by individual filters (upper triangle). Each filter is shown as a dot colored according to the category of the most similar single predictor.*

3.3.9 Lesion Analysis

Having broadly classified the functions of individual neurons, to assess their relative importance for to the network's overall functionality we performed a lesion analysis. We lesioned every neuron in the pruned network individually by setting all weights of the respective neuron to 0. For every neuron we tested the lesioned network in terms of loss on the validation image set (15 000 images) and correlation to humans on the target set we used earlier (720 individual pixels). It is important to note that this analysis is different to the pruning used to identify network variations that more closely matched humans, where we pruned outgoing connections of neurons (of which there are several for each neuron in the first two levels).

We correlated both lesion scores with the variance of each neuron's representations explained by representations in each category of predictors (summarized in the bars under each network plot in **Figure 15a**). We find no substantial proportion of variance in loss or similarity to humans due to lesioning explained by the lesioned neurons' similarity to any predictor category (the largest being $R^2 = .03$ between individual neurons' similarity to edge detection / direction predictors and loss of the network with the respective neuron lesioned). The importance of neurons to network performance cannot be explained by similarity to any of the predictor categories. Put differently, the function of the network as a whole seemingly depends on all the different classes of function roughly equally. This can be contrasted to previous findings, where particular classes of node were of special importance to overall network function (van Assen et al., 2020).

3.4 Discussion

We trained a neural network to identify specular highlights in computer renderings of surfaces. This is a challenging mid-level visual inference about the causal origin of bright patches in images, which is considered a key stage in the perception of gloss (Beck & Prazdny, 1981; Berzhanskaya et al., 2005; Fleming, 2012; Kim et al., 2011; Todd et al., 2004). Unlike other recent work using deep learning to identify highlights in the machine vision literature (Attard et al., 2020; Fu et al., 2020; Lin, el Amine Seddik, Tamaazousti, Tamaazousti, & Bartoli, 2019; Madessa et al., 2020), our focus was on matching human performance, rather than achieving the best possible accuracy from an engineering perspective. To do this, we trained a neural network to identify highlights and used pruning to identify a subnetwork within the trained network that responds more like human observers than the full network. We found not one, but many configurations that do this (**Figure 12a**). The best of these significantly outperformed both a simple threshold operation and ground truth at predicting human judgments. To our

knowledge, this is the first image-computable model that predicts average human highlight perception judgments at approximately the same level as individual participants do.

Our network consists of 61,505 trainable parameters and is designed to respond to 65,536 pixels for one individual image. Importantly, we find that the similarity of pruned networks to humans is highly consistent on both target and validation datasets, correlating at r =.99 for a subset of 3000 random and 3000 fit pruning configurations. A small dataset of 720 individual pixels was sufficient to identify a component of network responses that transfers to another independent dataset with data from different observers responding to different stimuli. This indicates strongly that the component in the network that was emphasized by pruning is robust across random variations in our dataset. This opens up other possible applications for pruning where target data is hard to come by and a small target dataset could be used to fine tune a model that is pre trained on simulated data.

We found configurations that correlate better with humans than the full network in which more than half of all connections were pruned. This can have several possible explanations. One is that there are superfluous connections that make little or no impact to the behavior of the network overall, or do not influence its similarity to humans. Another possibility is that connections are redundant. A third explanation could be that there are many alternative routes that the network can take to act more human. It seems likely that superfluous connections and neurons play at least a part in this, due to the high number of dead neurons (55 out of 177 in the pruned network) that we have seen in later analysis. In the RSA we also saw that a high number of neurons show a high similarity in their representations to image computable features. This high similarity to the same predictor or groups of predictors also makes the possibility of alternative routes within the network seem plausible.

Picking one representative pruned configuration, we summarized the different responses of the pruned and the full network in terms of sums of pixels belonging to different image regions (specular patches, bright texture patches and all other pixels). The pruned network shows a higher prediction sum for all categories (i.e., it is more likely to recognize them as highlights), but the increase is largest for specular pixels, making the difference between overall predictor sum for specular and texture pixels larger than it is for the full network. Put another way, the pruned network acts as if it has a (liberal) criterion shift relative to the full network, accepting more highlights than the original. While it might be tempting to think that the pruned network performs better in terms of the original training criterion of identifying highlights, this is not the case. The mean per-pixel predictions in **Figure 13b** ignore the number of pixels in each category while the training loss does not. This criterion shift opens up questions about

whether human perceptual decisions (especially in ambiguous cases) are driven by maximizing true positives or correct reject decisions and indeed our results indicate that humans are more likely than the full or pruned networks to make true positive decisions, at the same time also increasing the rate of false positives compared to our networks. Perhaps false positive decisions in recognizing highlights are not as important to humans and it is questionable whether the exact area covered by highlights plays a similar role to humans in this context as it does to our networks as defined by the loss. If a measurement related to coverage does play a role in human decisions, it seems likely that highlights weigh more than their actual coverage given that highlights propagate the impression of glossiness to a larger image area (Berzhanskaya et al., 2005), but this needs further investigation.

Previous studies have emphasized the role of so-called 'photo-geometric' constraints in identifying and interpreting highlights (Anderson & Kim, 2009; Beck & Prazdny, 1981; Kim et al., 2011; Marlow et al., 2011; Todd et al., 2004). In order for a bright image patch to be a specular reflection, it must align in orientation and position with the underling surface geometry and/or shading patterns. We do find that global rotation of the highlight component in our stimuli leads to a slight increase in prediction error (RMSE) by our pruned network compared to the manipulated specular map. This indicates that the network has learned to a limited degree to use orientations and positions of patches as a cue to identify highlights. However, the magnitude of the effect is small, suggesting that these constraints do not weigh heavily. It is interesting to note that a model can predict a large amount of variance in the human data without being very sensitive to violations of these constraints. This is similar to the results of Prokott, Tamura and Fleming (2021), where we found that neural networks trained to discriminate high- from low-gloss materials show little difference in their response when presented with images with displaced specular components.

Similar to these globally rotated highlights, our training and test images included conditions with false highlights, where we applied highlights from a different scene as texture. We included these as a challenging condition similar to stimuli used by (Anderson & Kim, (2009), but with both correct and incorrectly-placed highlights in one image. As with rotated highlights, our network erroneously identified most of these false highlights as highlights. While this is underperformance in terms of the training objective, our network is a better predictor of human responses to these images than the threshold model or ground truth, and the pruned network predicts human responses even better. Correlation to humans on pixels from these images in the target set: $r = 0.66$, $r = 0.27$, $r = 0.76$, $r = 0.78$ for threshold model, ground truth, unpruned network and pruned network respectively (validation set: $r = 0.65$, $r = 0.24$, $r =$

0.74, r = 0.73 respectively). While Anderson and Kim, 2009; Kim et al., 2011) showed that surfaces with displaced highlights are less likely to be perceived as glossy, we find some evidence that false highlights are rather equally perceived as highlights. We also find that the threshold model relying only on image intensity is a better predictor for human responses to these stimuli than for other texture conditions. However, it should be noted that this is not the focus of this study and that this subset of stimuli only consists of 96 individual pixels.

In an RSA we find similarities between the representation at single neurons in the network to various candidate predictors. Generally, we find that similarity to low complexity predictors that are computable directly from the image occurs throughout the network, but also earlier than more complex predictors. Complex geometric and intrinsic predictors show similarities only to very late neurons. We find similarities to summary predictors in spatially summarized units of the network. An alternative way of looking at the data is to classify each neuron according to the category of its most similar predictor (**Figure 15b**). This reveals that a very large proportion of neurons is most similar to either the input image or an image computable predictor describing either edges and pixel contrast or the image gradients and anisotropy. However, in a lesion analysis a neuron's similarity to any category of predictors was not predictive of its impact – when lesioned – on the model loss or its similarity to humans. Our data do not attest particular importance to any predictor category. This suggests that neurons similar to various predictors are important for the network to perform well and to predict human highlight perception.

A possible way of interpreting these results is that humans use different strategies. Where there are conflicting cues such as congruent and incongruent highlight-like patches, it might be that humans resort to simpler, less conflicting cues, such as luminance. Human responses to stimuli containing 'false highlights' as textures are more similar to predictions by the threshold model than human responses are overall. This represents one case in which humans responded very much like a simple intensity-based model. Our results suggest that photo-geometrical constraints are not the single most important cue to human gloss and highlight perception and will not prevail over simpler factors like overall relative brightness under all conditions.

3.5 Conclusion

We investigated human perception of highlights on glossy surfaces that also contain different types of bright texture patches. We developed, to our knowledge, the first image- computable highlight-detection algorithm that accurately reproduces human judgments. We demonstrated

an application of pruning using a genetic algorithm as a method for fine tuning a neural network trained on simulated physical data to a sparse dataset of human responses. Improvements due to pruning in network similarity to human judgements on a target dataset transfer well to a parallel dataset and are consistent over different stimulus conditions. On both datasets the pruned networks correlate with humans better than the full network and as high as the maximum human-to-human correlation we observed. Compared to the unpruned network, a pruned example network shows a criterion shift, which makes false positive judgements slightly more likely, while at the same time increasing the average difference in responses between pixels that contain a highlight and pixels that do not. We see modest evidence that our network has learned to use photo-geometric cues to identify whether bright patches are highlights, but these effects are very small. In an RSA and subsequent lesion analysis we find no evidence that neurons that are similar to geometric predictors (or any other class of predictors) are especially important for the network to achieve low objective loss or high similarity to humans. The lesion analysis provides no evidence that photo-geometric cues are particularly important for the network to respond similarly to human observers, suggesting that not only these relatively complex computations are being used by the human visual system in perceiving highlights.

Chapter 4:

DISCUSSION AND CONCLUSIONS

In this thesis I trained feed forward CNNs as image computable models of human gloss perception. The aims of the projects were to investigate the extent to which models using image features can imitate human responses and what higher-level representations neural networks learn to perform such a task. I investigated network behavior compared to humans on two central tasks of gloss perception – gloss classification and localization of highlights.

In **chapter 2** we found that networks of shallow to intermediate depth (2 to 5 layers) most typically show human-like responses in gloss classification. We also find that models of a similar depth become most sensitive to manipulations of image features of highlights, reacting similarly to humans. We find that DCGANs of 2 to 3 layers or more produce images that humans can reliably identify as high or low gloss. In **chapter 3** we investigated human perception of highlights on glossy textured surfaces. We trained one four-layer CNN to identify highlights in such images and pruned the network to a configuration that responds more similarly to human observers than the full network. While the network learned some individual internal representations that were most similar to geometrical predictors, we found these not to be particularly important for the network to perform the task well or to respond similarly to humans. The networks in both projects were only very weakly sensitive to violations of photo-geometric constraints.

4.1 Gloss perception

The results from both projects underpin the perspective of gloss perception as a function of statistical appearances of mid-level features. The summary image statistics used in **chapter 2** were not a good model of human responses, and CNNs from both projects explained human data well while only showing a very weak sensitivity to photeogeometric constraints. In **chapter 2** we found that neural networks with 3-5 layers most typically correlate highly with humans. We also showed images with manipulated highlights to networks and found that reduced highlight size, contrast and displacement on average cause predictions to shift towards the low gloss category, while increased highlight size causes predictions to shift towards high

gloss. These effects are smaller for shallow networks and found a maximum for networks with 3-6 or more layers. These effects on network predictions are similar to results for humans observed by Marlow et al. (2012). Interestingly, our networks have learned these variations in predicted gloss from only two material categories. According to Fleming (2014), under a statistical appearance view, the visual system would not aim to infer a complex reflectance function, but rather to capture the typical look of glossy surfaces. This is what our CNNs did when they learned to recognize a material under various viewing conditions. Our results demonstrate how this perceptual goal, given even a small number of materials and their characteristic appearances can lead to a continuous scale of predicted gloss and that variations along this scale are associated with image features that have been shown to influence human gloss perception. This is similar to what has been shown by Storrs et al. (2021) using unsupervised learning and generative networks.

We also found that observers could discriminate images from DCGANs into low and high gloss when the DCGANs had 2 or more layers (reaching similar levels to rendered images at 3 layers depth). The results of the DCGAN experiment also demonstrate that gloss perception in humans does not necessarily require a perceivable shape, as many images that humans perceived to be unrealistic or not to contain objects could be discriminated well into material categories.

In **chapter 3** we saw that a network trained to detect pixels and pruned to increase response similarity to humans, shows representational similarity mainly to simple image computable predictors with the activations of only a few units best explained by geometric or intrinsic predictors at later network stages. We also find that the network is only weakly sensitive to violations to photo-geometric constraints, as seen by network responses to images with rotated highlight components.

Together these results indicate that neural networks that respond to highlights or gloss are also sensitive to similar image cues as humans. Our highlight detection network contained some representations that are better explained by geometric or intrinsic factors than by simple image computable predictors, but a lesion analysis showed these to be no more important than neurons similar to image computable predictors. Networks from both chapters are only weakly sensitive to photo-geometric violations. In the tasks we investigated this suggests that neural networks can predict glossiness and detect highlights similarly to humans without in depth representation of other distal scene parameters such as 3D shape. In **chapter 3** such a network predicted human responses better than ground truth based on a physical simulation. This indicates that humans could be using higher order image statistics to recognize the typical

appearance of gloss without entirely decomposing a scene into all its veridical intrinsic components (or approximations thereof). The neural networks in this thesis lend some support to a *typical appearance* view of gloss perception (Fleming, 2012, 2014). They demonstrate a perceptual mechanism without deeper understanding of the underlying physics and that such models show a high similarity to human perceptual responses, even outperforming veridical labels in this respect (**chapter 3**).

4.2 Neural networks

The studies in this thesis represent two different routes into investigating convolutional neural networks as candidate models for human gloss perception. The large scale hyperparameter search in **chapter 2** showed that even large architectural changes do not necessarily reveal a clear trend in matching human responses. However, the search revealed typical behavior of different network depths in terms of similarity to human responses as well as sensitivity to image manipulations. We proposed that those network depths where similarity to humans is decorrelated from network performance most typically react like humans.

In **chapter 3** we investigated pruning as a means to fine tune a trained network to a small dataset of human responses. We found that the increase in similarity to humans transferred well to a parallel set of observer response data. This use of pruning as a second stage of fitting has possible applications in other research areas where target data is sparse or difficult to obtain, while approximate data (such as physical ground truth in this case) is more readily available.

In both chapters we faced the challenge of assembling a test set that is characteristic of human responses while being sufficiently different from veridical responses that a well performing model is not confused with a human-like model. In **chapter 2** we addressed this issue by carefully assembling a diagnostic set of human data over the course of several experiments, in which ground truth and human responses were decorrelated. In **chapter 3** we identified informative pixels based on two candidate models – ground truth and an intensity threshold model. This much less laborious route turned out to be a valid solution to the same problem, capturing a component of typical human behavior as indicated by the pruned model's transfer performance on the human validation set.

4.3 Limitations and Outlook

One of the biggest limitations of the projects in this thesis is the generalizability of the trained networks and the results. This means that the conclusions we draw are specific for the range of

images we use and the perceptual task we train our networks to perform. To be able to make some general claims over scene variations, we aimed to randomize factors in our training data like viewing angle, geometry, illumination (direction in **chapter 3** and lightmaps in **chapter 2**), at the same time keeping other factors the same, such as the general scene composition or viewing distance. Yet our training data may include other constant factors, due to image generation or processing as for example the image quality, the sampling quality used during rendering, color balancing or distortions due to the camera lens simulation. To investigate our networks we use test datasets within the boundaries of our training datasets by picking both sets from a larger base set at random to investigate specific properties of our networks. Yet conclusions that we draw about the specific behavior of our networks are valid only within the *style* and parameter range of our image sets. Direct comparisons to previous research from other authors using stimuli that were already tested on human observers would mean leaving the boundaries of our training data. To accommodate this limitation we recreated some of the test conditions from previous studies using our present stimuli (such as the rotated highlight test conditions in both chapters). Such limitations imposed by datasets are a known problem (Torralba & Efros, 2011). Current approaches to improve generalization often include image augmentation – different filters and manipulations or additive noise applied to the training images. However, Geirhos et al. (2018) found that training a CNN on a specific noise condition only improves performance for this specific type of noise. To at least partially address this limitation, it could be useful to compile training data from a variety of sources and styles (eg. renderings and photographs), to accommodate the styles of potential test stimuli in the range of training images.

Related to this, another issue that will need to be addressed to be able to better use neural networks for investigating human mid-level vision, is to what extent training data and task objective influence which specific representations *can* or *need* to be learned. Whether humans are able to use and perceive certain features in a dataset does not necessarily mean that these features are also the most salient or accessible features to a CNN trained only on this dataset. Recently several studies have focused on effects on network strategies and behavior due to augmented training images (Geirhos, Rubisch, et al., 2018; Singer et al., 2021), ecologically more valid training images (Mehrer, Spoerer, Jones, Kriegeskorte, & Kietzmann, 2021) and different loss functions (Hermann, Chen, & Kornblith, 2020). The networks trained in this thesis demonstrate certain mechanisms that networks learned to solve the given perceptual tasks, and that these go a long way in imitating human behavior. Better understanding the availability of cues and how a certain task or dataset influence neural networks to use different

features will help us make better use of neural networks as a tool for investigating human perception.

A related limitation of this thesis is that the projects and the learning shown here are confined to a very narrow area of perception. We train networks only on gloss or highlight perception and compare them to human observers who are well capable of highlight perception and responding to specific questions, but who don't naturally perceive gloss or highlights in isolation. This is important, because it means that gloss perception for humans is embedded in the context of other factors of materials, objects or scenes. For humans, gloss perception also contributes to other percepts, for example recognizing materials or identifying objects. To address this limitation, it would be interesting to investigate whether and how simultaneous perception of several scene factors influence each other or how a model performs compared to humans if given explicit scene information such as surface geometry.

One of the largest limitations of the networks discussed in this thesis is that they show only very weak sensitivity to violations of photo-geometric constraints. This demonstrates that the tasks we posed here can be solved to give human-like responses without a marked sensitivity to these constraints. However, research suggests that humans at least partly rely on photo-geometric constraints (Anderson & Kim, 2009; Beck & Prazdny, 1981; Kim et al., 2011; Kim & Anderson, 2010; Marlow et al., 2011; Marlow et al., 2015; Marlow & Anderson, 2015; Todd et al., 2004). A possible and obvious explanation for our CNNs' weak sensitivity to such regularities is that the training images used here caused little necessity to learn them. While photo-geometric constraints are contained in the training images, violations of them are not, and learning them is not useful for the visual tasks we posed.

In the context of possible extensions of this work to address effects of tasks and datasets discussed above, it is interesting to speculate what might compel a CNN to learn more about photo-geometric constraints. A straightforward way would be to train a network on a large dataset of images with correctly or incorrectly aligned specular components. This might work as a proof-of-concept, that networks can indeed learn to identify photo-geometric violations, but as a visual diet, such a dataset is very different from the images humans encounter every day. A more elegant way might be to address the issue using images of correctly aligned highlights and different visual tasks, or even parallel tasks. Possibly, photo-geometric constraints are costly to learn for a network and bring little advantage in the tasks shown here (for example resulting in a shallow learning gradient). A network that is trained to extract both the gloss of a material and the surface shape, and learns shape-like representations for their own sake, might find these more convenient to use as an additional cue to gloss. Or possibly the

misalignment of highlights and shading causes an interference from other mechanisms, such as foreground – background segmentation, in which case these highlights would not be considered as gloss cues anymore because they are not perceived as being part of the surface. CNNs could be used to test such hypotheses, using the targeted learning of isolated tasks to our advantage. These conjectures are highly speculative and are only meant as examples of possible questions of multiple parallel visual tasks that could be addressed with neural networks. Investigating such parallel tasks could help us generate hypothesis about the perceptual goals of the human visual system and to better understand how features are shared for multiple perceptual tasks.

A useful continuation of the work in this thesis would be the closer investigation of intermediate representations learned by CNNs for gloss or highlight perception. We showed that our network for highlight detection learned representations that are similar to geometric and intrinsic predictors, and this net as well as the gloss discrimination networks show a weak sensitivity to photo-geometric violations. It is possible that these effects reflect an explicit representation of geometry and relative highlight placement, but also conceivable that there are higher level image statistics that correlate with geometric predictors without explicitly calculating geometry as a prerequisite for gloss predictions. Alternatively, it could be speculated that these representations are examples of a shared intermediate processing stage, prior to isolated representations of gloss, but also prior to explicitly representing shape. In a retinal image shape, material and illumination are confounded, and after perceptual processing they are still not perfectly isolated and can still influence human perception of each other. Given this it is conceivable that there are intermediate stages to gloss perception in which other qualities are partially, if not explicitly represented. Future research investigating mid-level vision with neural networks could address the usefulness of such intermediate stages for predicting other factors, such as shape. A study investigating this could for example train a series of networks to extract different intrinsic components from the same database of images, and then use transfer training to evaluate the usefulness of representations within these networks for predicting the other components. Such a study could potentially provide grounds for hypotheses about a hierarchy of different mid-level representations, for example whether representations of shape are a prerequisite for perceiving gloss or vice versa.

In this context a useful extension of our work with DCGANs in **chapter 2** would be to use generative networks to create datasets with deliberately manipulated or excluded higher order statistics. For example, the coherence of surface cues could be suppressed to create images of gloss without shape or surface. This would mean to systematize effects like those we have seen with shallower DCGANs in **chapter 2**, for example through the use of different

architectures or possibly through post-training manipulations such as pruning. This way it could be possible to create datasets that retain some higher-level perceptual information (for example gloss), while specifically excluding others (for example shape or depth). Research into creating such datasets could address some of the limitations related to training data raised earlier. Transfer performance for networks trained and tested on different datasets would be more comparable if these different datasets could be generated based on one shared base dataset. Such manipulations of stimuli could also have applications in human perceptual studies without comparisons to CNNs.

Another limitation of this work is that we do not investigate whether CNNs show interactions of shape and illumination on gloss perception as have been described for humans (Olkkonen & Brainard, 2011; Wijntjes & Pont, 2010). We demonstrated that CNNs, like humans, show failures of gloss constancy which can be very similar to humans, but we do not investigate these failures in such detail as to compare the contributions and effects of individual physical factors as for example Doerschner et al. (2010) have done. CNNs could provide the means to test transfer functions of gloss perception between illuminations and geometries for large datasets and to create hypotheses for testing humans on specific examples.

Future research into gloss perception should also investigate the possibility that gloss perception may not only involve image-based features *or* always incorporates calculations of geometry, but rather both. The results of our lesion analysis in **chapter 3** hint at the possibility that features of different mechanisms are involved, showing CNN units similar to simple image features to be equally important to units that are similar to geometric predictors. It has been observed that different features contribute to gloss perception (Hunter, 1937; Leloup et al., 2012; Marlow et al., 2012; Motoyoshi & Matoba, 2012) and that these may be weighted differently for different stimuli (Marlow & Anderson, 2013) or that different observers may use different strategies (Leloup et al., 2012). Since image properties of highlights cannot explain all of human gloss perception, as highlights need to be identified as such first, other cues are involved in gloss perception. These could be other image features that identify highlights as such, but they could also include more complex geometrical computations. It is conceivable that different cues are involved that do not share the same mechanisms, such as image features *and* features involving knowledge of geometry. The integration of such cues of different mechanisms may well vary depending on the availability of information. A better understanding of this could reveal hierarchies in mid-level processing that apply beyond gloss perception.

This thesis can be seen as one of many steps in investigating human gloss perception or other aspects of mid-level vision using artificial neural networks. CNNs are a powerful tool for investigating higher level image statistics and how these are organized by a learning system. We trained CNNs as image computable models of two tasks of gloss perception – distinguishing high- from low-gloss images and localizing highlights. We showed how a large-scale network architecture search and hyperparameter optimization can be used to address questions about the computational complexity or demands of visual processes. We also demonstrated a use of network pruning for fine-tuning a neural network. In **chapter 2** we found that effects of image features on gloss perception previously observed in human participants showed in CNNs, providing some support for statistical appearance models of gloss perception. In both projects we found CNNs to be very weakly sensitive to photo-geometric constraints in glossy surfaces, suggesting that a marked sensitivity to such constraints is not necessary for a model to perceive gloss like human observers in the tasks we investigated.

REFERENCES

Adams, W. J., Elder, J. H., Graf, E. W., Leyland, J., Lugtigheid, A. J., & Muryy, A. (2016). The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. *Scientific Reports*, *6*, 35805. https://doi.org/10.1038/srep35805

Adelson, E. H. (2000). Lightness Perception and Lightness Illusions. In M. Gazzaniga (Ed.), *The New Cognitive Neurosciences* (2nd ed., pp. 339–351). MIT Press.

Adelson, E. H. (2001). On Seeing Stuff: The Perception of Materials by Humans and Machines. In B. E. Rogowitz & T. N. Pappas (Eds.), *Proceedings of the SPIE 4299, Human Vision and Electronic Imaging VI* (pp. 1–12). https://doi.org/10.1117/12.429489

Anderson, B. L. (2011). Visual perception of materials and surfaces. *Current Biology*, *21*(24), R978–R983. https://doi.org/10.1016/j.cub.2011.11.022

Anderson, B. L. (2020). Mid-level vision. *Current Biology*. https://doi.org/10.1016/j.cub.2019.11.088

Anderson, B. L., & Kim, J. (2009). Image statistics do not explain the perception of gloss and lightness. *Journal of Vision*, *9*(11), 1–17. https://doi.org/https://doi.org/10.1167/9.11.10.

Attard, L., Debono, C. J., Valentino, G., & Castro, M. di. (2020). Specular Highlights Detection Using a U-Net Based Deep Learning Architecture. *2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA)*, 4–9. https://doi.org/10.1109/MCNA50957.2020.9264278.

Barrow, H., & Tenenbaum, J. (1978). Recovering intrinsic scene characteristics from images. In A. Hanson & E. Riseman (Eds.), *Computer vision systems* (pp. 2–25). Academic Press.

Bartoldson, B. R., Morcos, A. S., Barbu, A., & Erlebacher, G. (2020). The generalization-stability tradeoff in neural network pruning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33* (pp. 1–13). https://proceedings.neurips.cc/paper/2020/file/ef2ee09ea9551de88bc11fd7eeea93b0-Paper.pdf

Baslamisli, A. S., Le, H.-A., & Gevers, T. (2018). CNN Based Learning Using Reflection and Retinex Models for Intrinsic Image Decomposition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 6674–6683. https://doi.org/10.1109/CVPR.2018.00698

Beck, J., & Prazdny, S. (1981). Highlights and the perception of glossiness. *Perception & Psychophysics*, *30*(4), 407–410. https://doi.org/10.3758/BF03206160

Bell, S., Upchurch, P., Snavely, N., & Bala, K. (2013). OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics*, *32*(4), 1–17. https://doi.org/10.1145/2461912.2462002

Bell, S., Upchurch, P., Snavely, N., & Bala, K. (2015). Material recognition in the wild with the Materials in Context Database. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June*, 3479–3487. https://doi.org/10.1109/CVPR.2015.7298970

Berzhanskaya, J., Swaminathan, G., Beck, J., & Mingolla, E. (2005). Remote effects of highlights on gloss perception. *Perception*, *34*(5), 565–575. https://doi.org/10.1068/p5401

Billmeyer, F. W., & O'Donnell, F. X. D. (1987). Visual gloss scaling and multidimensional scaling analysis of painted specimens. *Color Research & Application*, *12*(6), 315–326.

Blalock, D., Ortiz, J. J. G., Frankle, J., & Guttag, J. (2020, March 6). What is the State of Neural Network Pruning? *Proceedings of Machine Learning and Systems 2020*. http://arxiv.org/abs/2003.03033

Bonneel, N., Kovacs, B., Paris, S., & Bala, K. (2017). Intrinsic Decompositions for Image Editing. *Computer Graphics Forum*, *36*(2), 593–609. https://doi.org/10.1111/cgf.13149

Boyadzhiev, I., Bala, K., Paris, S., & Adelson, E. (2015). Band-Sifting Decomposition for Image-Based Material Editing. *ACM Transactions on Graphics, 34*(5). https://doi.org/10.1145/2809796

Chadwick, A. C., & Kentridge, R. W. (2015). The perception of gloss: A review. *Vision Research*, *109*(PB), 221–235. https://doi.org/10.1016/j.visres.2014.10.026

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(27755). https://doi.org/10.1038/srep27755

Cox, D. D., & Dean, T. (2014). Neural networks and neuroscience-inspired computer vision. *Current Biology*, *24*(18), R921–R929. https://doi.org/10.1016/j.cub.2014.08.026

Debevec, P. (1998). Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1998*, 189–198. https://doi.org/10.1145/280814.280864

Doerschner, K., Boyaci, H., & Maloney, L. T. (2010). Estimating the glossiness transfer function induced by illumination change and testing its transitivity. *Journal of Vision*, *10*(4), 1–9. https://doi.org/10.1167/10.4.8

Fleming, R. W. (2012). Human perception: Visual heuristics in the perception of glossiness. *Current Biology*, *22*(20), R865–R866. https://doi.org/10.1016/j.cub.2012.08.030

Fleming, R. W. (2014). Visual perception of materials and their properties. *Vision Research*, *94*(24), 62–75. https://doi.org/10.1016/j.cub.2011.11.022

Fleming, R. W. (2017). Material Perception. *Annual Review of Vision Science*, *3*(1), 365–388. https://doi.org/10.1146/annurev-vision-102016-061429

Fleming, R. W., Dror, R. O., & Adelson, E. H. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, *3*(5), 347–368. https://doi.org/10:1167/3.5.3

Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, *13*(8), 1–20. https://doi.org/10.1167/13.8.9

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201. https://doi.org/10.1038/nn.2889

Fu, G., Zhang, Q., Lin, Q., Zhu, L., & Xiao, C. (2020). Learning to Detect Specular Highlights from Real-world Images. *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, 1873–1881. https://doi.org/https://doi.org/10.1145/3394171.3413586

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

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv*, *1811.12231*.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*, 7549–7561.

Georgoulis, S., Rematas, K., Ritschel, T., Gavves, E., Fritz, M., van Gool, L., & Tuytelaars, T. (2018). Reflectance and Natural Illumination from Single-Material Specular Objects Using Deep Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(8), 1932–1947. https://doi.org/10.1109/TPAMI.2017.2742999

Gilchrist, A. L. (1977). Perceived Lightness Depends on Perceived Spatial Arrangement. *Science*, *195*(4274), 185–187. https://doi.org/10.1126/science.831266

Gomez-Villa, A., Martin, A., Vazquez-Corral, J., & Bertalmio, M. (2018). Convolutional neural networks deceived by visual illusions. *ArXiv*, *1811.10565*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. https://papers.nips.cc/paper/5423-generative-adversarial-nets

Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *352*(1358), 1121–1127. https://doi.org/10.1098/rstb.1997.0095

Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. *Advances in Neural Information Processing Systems 28*, 1135–1143.

Harrison, V. G. W. (1949). Gloss measurement of papers: A comparative study. *Journal of Scientific Instruments*, *26*(3), 84–90. https://doi.org/10.1088/0950-7671/26/3/307

Hassibi, B., & Stork, D. G. (1993). Second Order Derivatives for Network Pruning: Optimal Brain Surgeon. *Advances in Neural Information Processing Systems 5 (NIPS 1992)*, 164–171. https://resolver.caltech.edu/CaltechAUTHORS:20150219-075206704

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hermann, K. L., Chen, T., & Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1016/B978-0-08-044894-7.00081-6

Hiramatsu, C., Goda, N., & Komatsu, H. (2011). Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. *NeuroImage*, *57*(2), 482–494. https://doi.org/10.1016/j.neuroimage.2011.04.056

Ho, Y.-X., Landy, M. S., & Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture: Research article. *Psychological Science*, *19*(2), 196–204. https://doi.org/10.1111/j.1467-9280.2008.02067.x

Hunter, R. S. (1937). Methods of determining gloss. *Journal of Research of the National Bureau of Standards*, *18*(1), 19. https://doi.org/10.6028/jres.018.006

Ingersoll, L. R. (1921). The Glarimeter an Instrument for Measuring the Gloss of Paper. *Journal of the Optical Society of America*, *5*(3), 213–217.

Jähne, B. (1993). *Spatio-Temporal Image Processing: Theory and Scientific Applications* (1st ed.). Springer-Verlag. https://doi.org/https://doi.org/10.1007/3-540-57418-2

Janowsky, S. A. (1989). Pruning versus clipping in neural networks. *Physical Review A*, *39*(12), 6600. https://doi.org/10.1103/PhysRevA.39.6600

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2018). Deep Neural Networks In Computational Neuroscience. *BioRxiv*, *133504*. https://doi.org/10.1101/133504

Kim, J., & Anderson, B. L. (2010). Image statistics and the perception of surface gloss and lightness. *Journal of Vision*, *10*(9), 1–17. https://doi.org/10.1167/10.9.3

Kim, J., Marlow, P., & Anderson, B. L. (2011). The perception of gloss depends on highlight congruence with surface shading. *Journal of Vision*, *11*(9), 1–19. https://doi.org/10.1167/11.9.4

Kim, J., Marlow, P. J., & Anderson, B. L. (2012). The dark side of gloss. *Nature Neuroscience*, *15*(11), 1590–1595. https://doi.org/10.1038/nn.3221

Knutsson, H. (1989). Representing local structure using tensors. *Proceedings of 6th Scandinavian Conference on Image Analysis*. https://doi.org/10.1007/978-3-642-21227-7_51

Komatsu, H., & Goda, N. (2018). Neural Mechanisms of Material Perception: Quest on Shitsukan. *Neuroscience*, *392*, 329–347. https://doi.org/10.1016/j.neuroscience.2018.09.001

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. *Proceedings of the 36th International Conference on Machine Learning*. https://proceedings.mlr.press/v97/kornblith19a.html.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25*, 1097–1105. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N. J., Issa, E. B., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.

Kubilius, J., Wagemans, J., & op de Beeck, H. P. (2014). A conceptual framework of computations in mid-level vision. *Frontiers in Computational Neuroscience*, *8*, 158. https://doi.org/10.3389/fncom.2014.00158

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2018). Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics. *Proceedings of the European Conference on Computer Vision (ECCV)*, 770–787. https://doi.org/10.1007/978-3-030-01270-0_47

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2019). Adversarial examples in the physical world. *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, *c*, 1–14.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Handwritten Digit Recognition with a Back-Propagation Network. *Advances in Neural Information Processing Systems 2*, 396–404. https://doi.org/10.1111/dsu.12130

LeCun, Y., Denker, J. S., & Solla, S. A. (1990). Optimal Brain Damage (Pruning). *Advances in Neural Information Processing Systems*, 598–605.

Leloup, F. B., Hanselaer, P., Pointer, M. R., & Dutré, P. (2012). Integration of Multiple Cues for Visual Gloss Evaluation. *Predicting Perceptions: Proceedings of the 3rd International Conference on Appearance*, *29*(April), 52–55. http://opendepot.org/1052/

Li, Y., Yosinski, J., Clune, J., Lipson, H., & Hopcroft, J. (2015). Convergent learning: Do different neural networks learn the same representations? *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, *PMLR 44*, 196–212.

Lin, J., el Amine Seddik, M., Tamaazousti, M., Tamaazousti, Y., & Bartoli, A. (2019). Deep Multi-class Adversarial Specularity Removal. *SCIA 2019: Image Analysis*, *LNCS, 1148*, 3–15. https://doi.org/10.1007/978-3-030-20205-7_1

Liu, C., Sharan, L., Adelson, E. H., & Rosenholtz, R. (2010). Exploring features in a Bayesian framework for material recognition. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 239–246.

Lotter, W., Kreiman, G., & Cox, D. (2017). Deep predictive coding networks for video prediction and unsupervised learning. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 1–18.

Madessa, A. H., Dong, J., Gan, Y., & Gao, F. (2020). A deep learning approach for specular highlight removal from transmissive materials. *Expert Systems*, e12598. https://doi.org/https://doi.org/10.1111/exsy.12598

Marlow, P. J., & Anderson, B. L. (2013). Generative constraints on image cues for perceived gloss. *Journal of Vision*, *13*(14), 1–23. https://doi.org/10.1167/13.14.2

Marlow, P. J., & Anderson, B. L. (2015). Material properties derived from three-dimensional shape representations. *Vision Research*, *115*, 199–208. https://doi.org/10.1016/j.visres.2015.05.003

Marlow, P. J., Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, *22*(20), 1909–1913. https://doi.org/10.1016/j.cub.2012.08.009

Marlow, P. J., Todorović, D., & Anderson, B. L. (2015). Coupled computations of three-dimensional shape and material. *Current Biology*, *25*(6), R221–R222. https://doi.org/10.1016/j.cub.2015.01.062

Marlow, P., Kim, J., & Anderson, B. L. (2011). The role of brightness and orientation congruence in the perception of surface gloss. *Journal of Vision*, *11*(9), 1–12. https://doi.org/10.1167/11.9.16

Matusik, W., Pfister, H., Brand, M., & McMillan, L. (2003). A data-driven reflectance model. *ACM SIGGRAPH 2003 Papers, SIGGRAPH '03*, 759–769. https://doi.org/10.1145/1201775.882343

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(8), 1–9. https://doi.org/10.1073/pnas.2011417118

Mordvintsev, A., Olah, C., & Tyka, M. (2015). *DeepDream - a code example for visualizing Neural Networks*. Google AI Blog. https://ai.googleblog.com/2015/07/deepdream-code-example-for-visualizing.html

Motoyoshi, I., & Matoba, H. (2012). Variability in constancy of the perceived surface reflectance across different illumination statistics. *Vision Research*, *53*(1), 30–39. https://doi.org/10.1016/j.visres.2011.11.010

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature*, *447*(7141), 206–209. https://doi.org/10.1038/nature05724

Mozer, M. C., & Smolensky, P. (1989). Skeletonization: A Technique for Trimming the Fat from a Network via Relevance Assessment. *Advances in Neural Information Processing Systems 1*, 107–115.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 427–436. https://doi.org/10.1109/CVPR.2015.7298640

Nicodemus, F. E. (1965). Directional Reflectance and Emissivity of an Opaque Surface. *Applied Optics*, *4*(7), 767–775.

Nishida, S. (2019). Image statistics for material perception. *Current Opinion in Behavioral Sciences*, *30*, 94–99. https://doi.org/10.1016/j.cobeha.2019.07.003

Nishida, S., & Shinya, M. (1998). Use of image-based information in judgments of surface-reflectance properties. *Journal of the Optical Society of America A*, *15*, 2951–2965. https://doi.org/10.1364/josaa.15.002951

Nishio, A., Goda, N., & Komatsu, H. (2012). Neural Selectivity and Representation of Gloss in the Monkey Inferior Temporal Cortex. *Journal of Neuroscience*, *32*(31), 10780–10793. https://doi.org/10.1523/JNEUROSCI.1095-12.2012

Okazawa, G., Goda, N., & Komatsu, H. (2012). Selective responses to specular surfaces in the macaque visual cortex revealed by fMRI. *NeuroImage*, *63*(3), 1321–1333. https://doi.org/10.1016/j.neuroimage.2012.07.052

Olah, C., Mordvintsev, A., & Schubert, L. (2017). *Feature Visualization*. Distill. https://doi.org/10.23915/distill.00007

Olkkonen, M., & Brainard, D. H. (2010). Perceived glossiness and lightness under real-world illumination. *Journal of Vision*, *10*(9), 1–19. https://doi.org/10.1167/10.9.5

Olkkonen, M., & Brainard, D. H. (2011). Joint effects of illumination geometry and object shape in the perception of surface reflectance. *I-Perception*, *2*(9), 1014–1034. https://doi.org/10.1068/i0480

Pellacini, F., Ferwerda, J. A., & Greenberg, D. P. (2000). Toward a psychophysically-based light reflection model for image synthesis. *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*, 55–64. https://doi.org/10.1145/344779.344812

Pelli, D. G., & Farell, B. (1999). Why use noise? *Journal of the Optical Society of America A*, *16*(3), 647–653. https://doi.org/10.1364/josaa.16.000647

Pont, S. C., & te Pas, S. F. (2006). Material - Illumination ambiguities and the perception of solid objects. *Perception*, *35*(10), 1331–1350. https://doi.org/10.1068/p5440

Portilla, J., & 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.

Prokott, K. E., Tamura, H., & Fleming, R. W. (2021). Gloss perception: Searching for a deep neural network that behaves like humans. *Journal of Vision*, *21*(12), 1–20. https://doi.org/10.1167/jov.21.12.14

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

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & 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. https://doi.org/10.1523/JNEUROSCI.0388-18.2018

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536. https://doi.org/10.7551/mitpress/1888.003.0013

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. *Advances in Neural Information Processing Systems 29*, 2234–2242.

Sawayama, M., Adelson, E. H., & Nishida, S. (2017). Visual wetness perception based on image color statistics. *Jornal of Vision*, *17*(5), 1–24. https://doi.org/10.1167/17.5.7

Sawayama, M., & Nishida, S. (2018). Material and shape perception based on two types of intensity gradient information. *PLOS Computational Biology*, *14*(4), e1006061. https://doi.org/10.1371/journal.pcbi.1006061

Sawayama, M., Nishida, S., & Shinya, M. (2017). Human perception of subresolution fineness of dense textures based on image intensity statistics. *Journal of Vision*, *17*(4), 1–18. https://doi.org/10.1167/17.4.8

Schwartz, G., & Nishino, K. (2020). Recognizing Material Properties from Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(8), 1981–1995. https://doi.org/10.1109/TPAMI.2019.2907850

Serrano, A., Gutierrez, D., Myszkowski, K., Seidel, H.-P., & Masia, B. (2016). An intuitive control space for material appearance. *ACM Transactions on Graphics*, *35*(6), 1–12. https://doi.org/10.1145/2980179.2980242

Sève, R. (1993). Problems connected with the concept of gloss. *Color Research & Application*, *18*(4), 241–252.

Sharan, L., Liu, C., Rosenholtz, R., & Adelson, E. H. (2013). Recognizing Materials Using Perceptually Inspired Features. *International Journal of Computer Vision*, *103*, 348–371. https://doi.org/10.1007/s11263-013-0609-0

Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of Vision*, *14*(9), 1–24. https://doi.org/10.1167/14.9.12

Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*, 1528–1540. https://doi.org/10.1145/2976749.2978392

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

Singer, J. J. D., Seeliger, K., Kietzmann, T. C., & Hebart, M. N. (2021). From photos to sketches - how humans and deep neural networks process objects across different levels of visual abstraction. *PsyArXiv*. https://doi.org/10.31234/osf.io/xg2uy

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*, 1–14. https://doi.org/10.3389/fpsyg.2017.01551

Stabinger, S., Rodríguez-Sánchez, A., & Piater, J. (2016). 25 years of CNNS: Can we compare to human abstraction capabilities? In A. Villa, P. Masulli, & A. P. Rivero (Eds.), *International Conference on Artificial Neural Networks (ICANN 2016)* (pp. 380–387). https://doi.org/10.1007/978-3-319-44781-0_45

Storrs, K. R., Anderson, B. L., & Fleming, R. W. (2021). Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, *5*, 1402–1417. https://doi.org/10.1038/s41562-021-01097-6

Sun, H.-C., di Luca, M., Ban, H., Muryy, A., Fleming, R. W., & Welchman, A. E. (2016). Differential processing of binocular and monocular gloss cues in human visual cortex. *Journal of Neurophysiology*, *115*(6), 2779–2790. https://doi.org/10.1152/jn.00829.2015

Suzuki, T., Abe, H., Murata, T., Horiuchi, S., Ito, K., Wachi, T., Hirai, S., Yukishima, M., & Nishimura, T. (2020). Spectral pruning: Compressing deep neural networks via spectral analysis and its generalization error. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, 2839–2846. https://doi.org/10.24963/ijcai.2020/393

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 1–10.

Todd, J. T., Norman, J. F., & Mingolla, E. (2004). Lightness Constancy in the Presence of Specular Highlights. *Psychological Science*, *15*(1), 33–39. https://doi.org/10.1111/j.0963-7214.2004.01501006.x

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347

Tripp, B. P. (2017). Similarities and differences between stimulus tuning in the inferotemporal visual cortex and convolutional networks. *Proceedings of the International Joint Conference on Neural Networks*, *2017-May*, 3551–3560. https://doi.org/10.1109/IJCNN.2017.7966303

van Assen, J. J. R., Nishida, S., & Fleming, R. W. (2020). Visual perception of liquids: Insights from deep neural networks. *PLoS Computational Biology*, *16*(8), 1–29. https://doi.org/10.1371/journal.pcbi.1008018

van Assen, J. J. R., Wijntjes, M. W. A., & Pont, S. C. (2016). Highlight shapes and perception of gloss for real and photographed objects. *Journal of Vision*, *16*(6), 1–14. https://doi.org/10.1167/16.6.6

Vangorp, P., Laurijssen, J., & Dutré, P. (2007). The influence of shape on the perception of material reflectance. *ACM Transactions on Graphics*, *26*(99), 77. https://doi.org/10.1145/1239451.1239528

VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology*, *8*, 142. https://doi.org/10.3389/fpsyg.2017.00142

Vedaldi, A., & Lenc, K. (2015). MatConvNet: Convolutional neural networks for MATLAB. *MM '15: Proceedings of the 23rd ACM International Conference on Multimedia*, 689–692. https://doi.org/10.1145/2733373.2807412

Wada, A., Sakano, Y., & Ando, H. (2014). Human cortical areas involved in perception of surface glossiness. *NeuroImage*, *98*, 243–257. https://doi.org/10.1016/j.neuroimage.2014.05.001

Wang, L., Hu, L., Gu, J., Wu, Y., Hu, Z., He, K., & Hopcroft, J. (2018). Towards understanding learning representations: To what extent do different neural networks learn the same representation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*.

Wang, T.-C., Zhu, J.-Y., Hiroaki, E., Chandraker, M., Efros, A. A., & Ramamoorthi, R. (2016). A 4D Light-Field Dataset and CNN Architectures for Material Recognition. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision -- ECCV 2016* (pp. 121–138). Springer International Publishing.

Ward, E. J. (2019). Exploring Perceptual Illusions in Deep Neural Networks. *BioRxiv*, *687905*. https://doi.org/10.1101/687905

Ward, G. J. (1992). Measuring and modeling anisotropic reflection. *Computer Graphics*, *26*(2), 265–272.

Ward, G. J. (1994). The RADIANCE lighting simulation and rendering system. *SIGGRAPH '94: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, 459–472. https://doi.org/10.1145/192161.192286

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., & Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in Psychology*, *9*, 345. https://doi.org/10.3389/fpsyg.2018.00345

Werbos, P. J. (1974). *Beyond regressions: New tools for prediction and analysis in the behavioral sciences*.

Werbos, P. J. (1981). Applications of advances in nonlinear sensitivity analysis. In R. F. Drenick & F. Kozin (Eds.), *System Modeling and Optimization* (pp. 762–770). Springer. https://doi.org/10.1007/BFb0006203

Wijntjes, M. W. A., & Pont, S. C. (2010). Illusory gloss on Lambertian surfaces. *Journal of Vision*, *10*(9), 1–12. https://doi.org/10.1167/10.9.13

Wills, J., Agarwal, S., Kriegman, D., & Belongie, S. (2009). Toward a perceptual space for gloss. *ACM Transactions on Graphics*, *28*(4), 103:1-103:15. https://doi.org/10.1145/1559755.1559760

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

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher

visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *ECCV 2014: Computer Vision – ECCV 2014* (pp. 818–833). Springer. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, Y., Ozay, M., Liu, X., & Okatani, T. (2016). Integrating deep features for material recognition. *Proceedings - International Conference on Pattern Recognition*, *0*, 3697–3702. https://doi.org/10.1109/ICPR.2016.7900209

Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2018). Interpreting Deep Visual Representations via Network Dissection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(9), 2131–2145. https://doi.org/10.1109/TPAMI.2018.2858759

Zhu, Z., Xie, L., & Yuille, A. (2017). Object recognition with and without objects. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 3609–3615. https://doi.org/10.24963/ijcai.2017/505

Zou, G. Y. (2007). Toward Using Confidence Intervals to Compare Correlations. *Psychological Methods*, *12*(4), 399–413. https://doi.org/10.1037/1082-989X.12.4.399

APPENDIX

A: Supplementary material for chapter 2

A similar version of this appendix was published as supplementary material for:
Prokott, K. E., Tamura, H., & Fleming, R. W. (2021). Gloss perception: Searching for a deep neural network that behaves like humans. *Journal of Vision, 21*(12), 1–20.

Diagnostic image set selection process

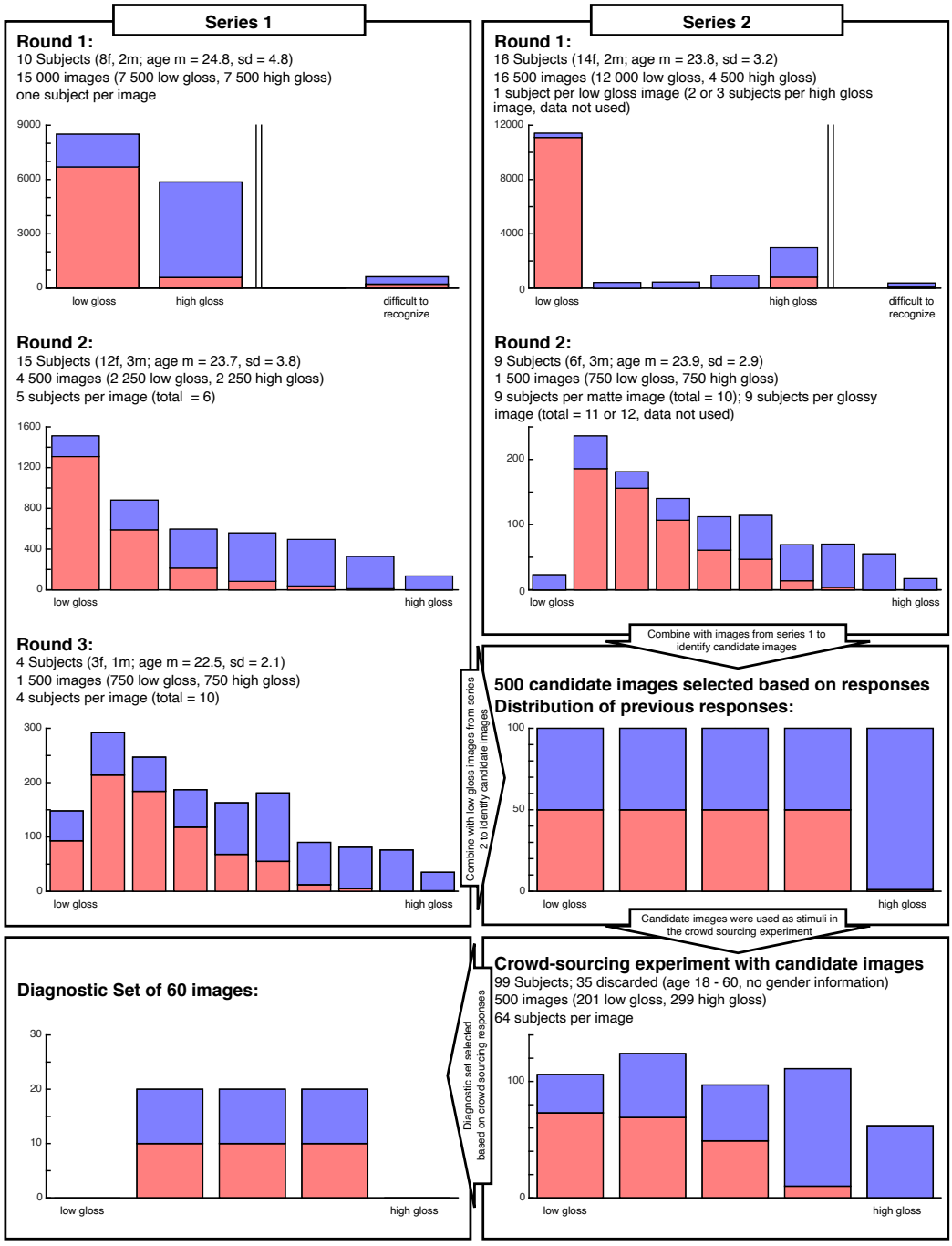

Figure S1: *The selection process of the diagnostic image set consisted of two series of experiments in our lab and one crowd sourcing experiment. See Methods for details. Blue indicates ground truth high gloss images, red indicates ground truth low gloss images, bars are stacked. Every graph shows the mean of subjects' binary high-gloss / low-gloss responses.* ***Top left box:*** *the first series of experiments started with 15 000 images, which we narrowed down to 1 500 images over the course of two experiments (top and middle graphs). The final 1 500 images were rated by 10 observers each (bottom graph).* ***Top right box:*** *The second series of experiments started with 16 500 images, of which only the 12 000 low gloss images were of interest to us, as the first series had already yielded a sufficient number of misperceived and ambiguous high-gloss images. Over two experiments these were narrowed down to 750 low-gloss images.* ***Center right box:*** *We selected 500 candidate images based on subjects' responses in the two series of screening experiments.* ***Bottom right box:*** *Responses to 500 candidate images from an online crowd-sourcing experiment.* ***Bottom left box:*** *The distribution of crowd sourcing responses for the 60 images in our diagnostic set*

DCGAN architectures

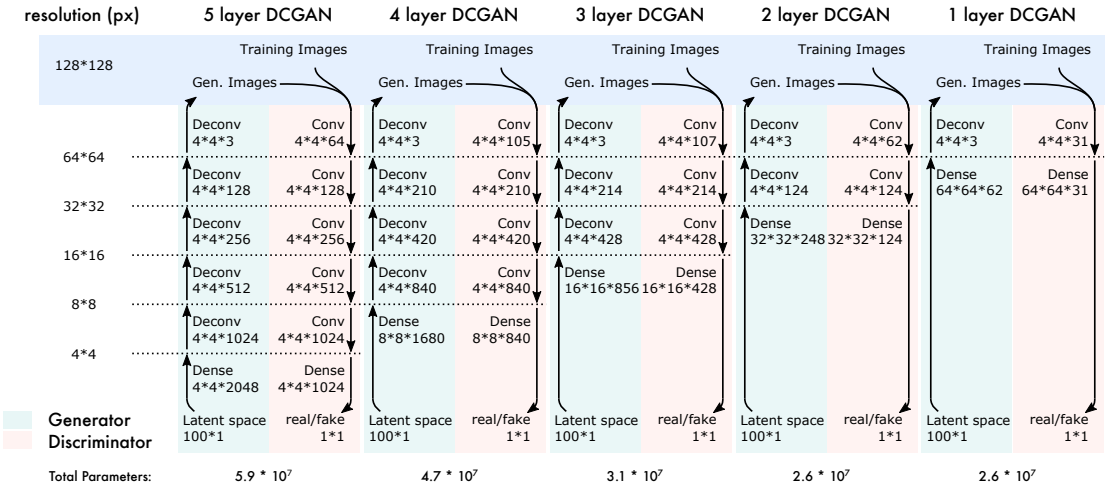

Figure S2: *DCGAN architectures. The parameters of convolutional and dense layers are shown for increasingly shallow DCGANs from left to right. Parameters were chosen so that image resolution would double between deconvolutional layers, processing depth (number of filters) would decrease by half for later deconvolutional layers and double for later convolutional layers. The latent space was always a vector of length 100.*

Use of color information in our CNNs

To investigate the role of color in our CNNs, we took a set of 750 low- and 750 high-gloss images from the image set used for training DCGANs (not used for training the CNNs) and passed these through the network both as RGB and as 3-channel grayscale inputs. Correlation between predictions for RGB and grayscale images per network were on average r = 0.90, 0.88, 0.91, 0.89, 0.89, 0.89, 0.89, 0.88, 0.87 for 1,2,3,4,5,6,7,8,12-layer networks respectively. Out of all 2639 networks we tested, only one had a correlation of less than 0.7 for colour and grayscale. See also **figure S3a**. This suggests a rather moderate role of colour relative to other cues. On average the networks have a lower accuracy with grayscale than with colour images.

Network accuracy (the proportion of images judged correctly) dropped on average by 0.044, 0.059, 0.051, 0.064, 0.062, 0.062, 0.064, 0.067, 0.072 for 1,2,3,4,5,6,7,8,12-layer networks respectively. See **Figure S3b**. Taken together, these results indicate that grayscale images are judged slightly less accurately but for most networks this effect is rather small.

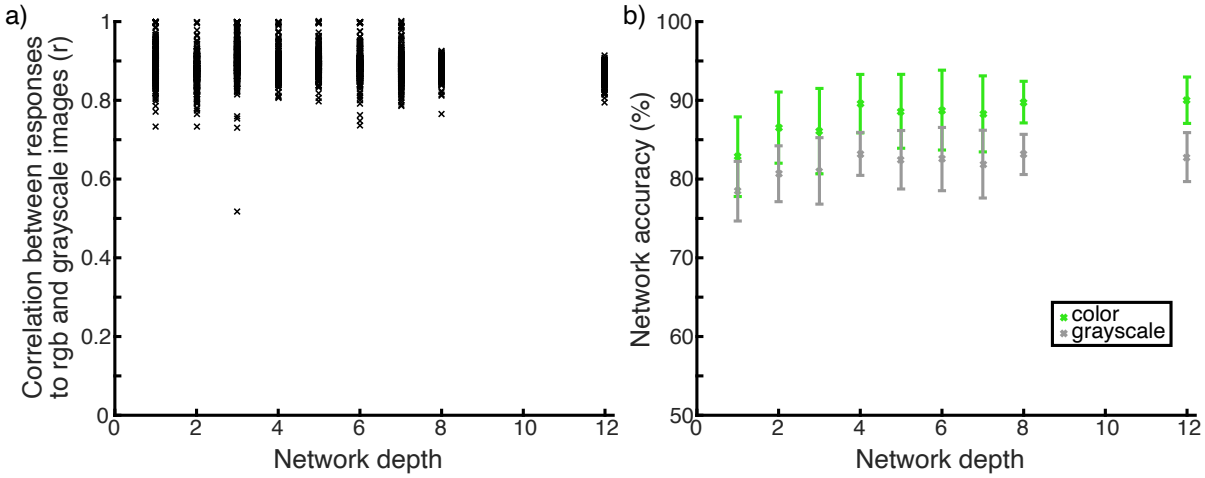

Figure S3: (a) *Correlation between individual CNNs' responses to RGB and grayscale versions of the same 1 500 images (750 high gloss, 750 low gloss).* ***(b)*** *Mean network accuracy for the same RGB and grayscale images used in* ***figure S3a*** *for different network depths. Error bars show sd.*

CNN accuracy outlier analysis

To address concerns that our analysis of CNN typicality (shown in **Figure 4c and 4d**) is driven by outliers we performed two analyses.

In the first analysis we left out the least accurate 1,2,5,10 or 15 % of networks from each depth group to repeat the analysis without outliers. **Figure S4a** shows how the distributions of networks of different depths in terms of accuracy and correlation to humans on the diagnostic image set change when the least accurate networks are left out of the analysis. **Figure S4b** shows the correlation coefficients of the individual depth groups in **figure S4a**, similar to **figure 2d**. We find similar results when we exclude the outliers. Deeper layers show a negative correlation between correlation to humans and accuracy. With increasing percentage removed from the dataset we find that shallower depth groups show a more negative correlation, moving the intersection with the x axis (the point of 0 correlation between accuracy and correlation to humans) towards shallower networks. Rejecting the least accurate 15% of networks results in an intersection with the x-axis between 1 and 2 layers.

In another analysis we took an alternative approach, looking within each depth group at the distance between the highest correlating network and the centroid of that depth group (**figure S4c**). Rather than a correlation as in the previous analysis, we look at the inverse gradient of these lines - the change of accuracy / the change in correlation to humans. The

inverse gradients are shown against network depth in **figure S4d**. In this case, the x axis intersection of a fitted curve happens around 2 layers. At this point the estimated gradient is 0, meaning that networks of this depth can be more similar to humans independently from their accuracy.

Our concern with analyses leaving out or mitigating the effect of outliers is that the accuracy distributions appear to be systematically skewed - more so for shallow networks than for deep ones (see also **figure S4e**). Especially for 1-layer networks there appears to be a second cluster of less accurate networks. (dark blue, around 70% accuracy). Leaving out even the lowest 1% already affects this cluster. Omitting an increasing percentage of the least accurately performing networks affects depth groups differently and systematically. The deepest groups are hardly affected while the correlation of accuracy to correlation with humans becomes increasingly lower for shallower networks.

We therefore find that all networks with non-random responses should be part of the analysis, as reported in the main text. However, it is interesting to see that leaving out the least accurate networks still shows a similar trend in correlations between accuracy and correlations to humans. The main difference is that the point of 0 correlation shifts towards shallower networks.

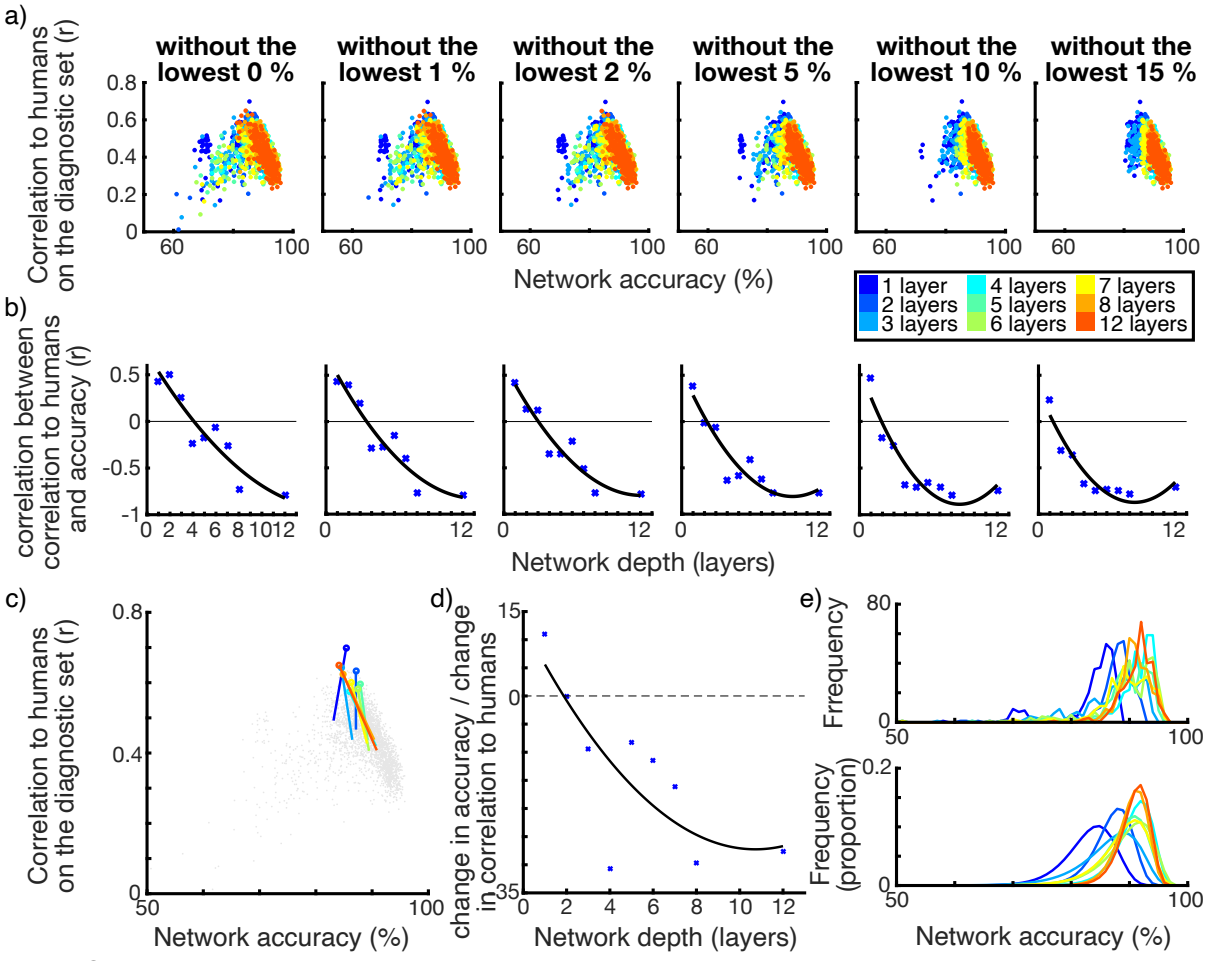

Figure S4: (a) *CNNs from the Bayesian search in terms of accuracy and correlation to humans on the diagnostic image set, omitting an increasing percentage of the least accurate networks per depth group. The left- most plot shows the same data as Figure 4c in the main text.* ***(b)*** *Correlation coefficients between accuracy and correlation to humans for each depth group, omitting an increasing percentage of least accurate networks, corresponding to the plots in Figure S4a* ***(c)*** *Lines from the centroid of each depth group to the network of that depth group that shows the highest correlation to humans (circles)* ***(d)*** *Inverse gradients of the lines shown in Figure S4c. The pattern resembles that in Figure 4d and Figure S4b. The intersection of the fitted line with the x axis has a similar meaning – independence between network accuracy and correlation to humans – and is situated around 2 layers network depth.* ***(e)*** *Distributions of network accuracies. The top graph shows raw data, the bottom graph shows fitted distributions.*

Sources of high-dynamic-range illumination maps

https://syns.soton.ac.uk/ (Adams et al., 2016)

http://www.pauldebevec.com/ (Debevec, 1998)

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/

Code and software used

Radiance (Ward, 1994)

Matlab

MathConvNet (Vedaldi & Lenc, 2015)

DCGAN toolbox: The 'dcgan-matconvnet' toolbox by Sung-Ho Bae on github; https://github.com/sunghbae/dcgan-matconvnet

Appendix A references

Adams, W. J., Elder, J. H., Graf, E. W., Leyland, J., Lugtigheid, A. J., & Muryy, A. (2016). The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. *Scientific Reports*, *6*, 35805. https://doi.org/10.1038/srep35805

Debevec, P. (1998). Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1998*, 189–198. https://doi.org/10.1145/280814.280864

Vedaldi, A., & Lenc, K. (2015). MatConvNet: Convolutional neural networks for MATLAB. *MM '15: Proceedings of the 23rd ACM International Conference on Multimedia*, 689–692. https://doi.org/10.1145/2733373.2807412

Ward, G. J. (1994). The RADIANCE lighting simulation and rendering system. *SIGGRAPH '94: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, 459–472. https://doi.org/10.1145/192161.192286

B: Supplementary material for chapter 3

A similar version of this appendix was submitted as supplementary material for:
Prokott, K. E. & Fleming, R. W. (*under review*). Identifying specular highlights: insights from deep learning

Detailed descriptions of predictors:

Input image:

*The **input image*** in one channel grayscale

Summary statistics:

mean intensity per image

***standard deviation** of intensity* per image

skewness of intensity distributions per image

kurtosis of intensity distributions per image

Edge detection / direction:

pixel gradients in x direction - using a sobel function

pixel gradients in y direction - using a sobel function

local contrast – the variance in each 3x3 pixel patch

locally normalized image – the difference between an individual pixel and the mean of the 3x3 pixel patch of which it is the center

Image gradients / anisotropy:

These predictors are based on the structure tensors of the images (Knutsson, 1989), which we calculated from image gradients smoothed using a gaussian with sd = 5. The structure tensor again was smoothed with a gaussian with sd = 5.

gradients of the smoothed image in x direction – the x component of local image orientation as described by the first eigenvector of the structure tensor

gradients of the smoothed image in y direction – the y component of local image orientation as described by the first eigenvector of the structure tensor

anisotropy of the smoothed image – the 'coherence' of the image calculated locally as $c = \left(\frac{\mu_1 - \mu_2}{\mu_1 + \mu_2}\right)^2$; where μ_1 and μ_2 are the first and second eigenvalues of the structure tensor respectively (Jähne, 1993)

Geometry information:

camera distance – distance between the camera and the surface at each pixel in the image. The units are arbitrary but standardized for all images

angle to camera – the angle between the surface normal at each pixel and the

light distance – the distance between the surface at each pixel and the center of the light source

angle to light source – the angle between the surface normal at each pixel and the direction to the center of the light source

convexity – the mesh curvature per vertex at each pixel in the image. This was obtained from the rendering scenes in Blender using the 'pointiness' property of the geometry node.

pointiness –the unsigned difference of the convexity data to zero curvature

x normal – the x component of the normal vector of the surface at each pixel

y normal – the y component of the normal vector of the surface at each pixel

z normal – the z component of the normal vector of the surface at each pixel

occluding edges – marking the location of edges that occlude other parts of the surface. These were rendered with Blender's 'Freestyle Line Set' functionality, using the 'Silhouette' edge type. These lines mark the edges between regions of a surface that face towards or away from the camera.

distance from occluding edges – distance in image space of each pixel from the nearest occluding edge

Intrinsic components:

texture – the (grayscale) texture component of each image (without shading information)

matte (shading) – the diffuse component of each image

specular – the ground truth specular reflections for each image

specular direct – only direct specular reflections of the light source

specular indirect – all indirect specular reflections (interreflections) for each image

specular coverage – the proportion of pixels covered by specular reflections in each image

texture coverage – the proportion of pixels covered by bright texture markings in each image

Scene information:

surface scale – the scale of the geometry of the surface. The surfaces for the scenes were created using Blender's 'ocean' simulation, the scale was determined by the

'smallest wave' parameter. We used 4 values spaced equally on a log2 scale. For the predictor RDM we used an ordinal scale – the RDM indicates how many steps apart the surface scale of two images is.

texture type – the category of texture of each image – marble, Voronoi, checkered, untextured or false highlights. This factor was used nominally so the RDM was binary, indicating whether the texture type for two images was the same or different

texture condition – the category and scale of texture of each image. For Voronoi, marble and checker textures there are 4 texture scales, plus two different false highlight and the untextured conditions making a total of 15 texture conditions for each scene. The RDM is binary, showing whether two images are of the same texture condition or not.

scene – a nominal factor resulting in a binary RDM that showed whether two images shared the same underlying scene / geometry.

Appendix B references:

Knutsson, H. (1989). Representing local structure using tensors. *Proceedings of 6th Scandinavian Conference on Image Analysis*. https://doi.org/10.1007/978-3-642-21227-7_51

Jähne, B. (1993). *Spatio-Temporal Image Processing: Theory and Scientific Applications* (1st ed.). Springer-Verlag. https://doi.org/https://doi.org/10.1007/3-540-57418-2

Liste der Veröffentlichungen

Prokott, K. E., Tamura, H., & Fleming, R. W. (2021). Gloss perception: Searching for a deep neural network that behaves like humans. *Journal of Vision*, *21*(12), 1–20. https://doi.org/10.1167/jov.21.12.14

Morgenstern, Y., Schmidt, F., Hartmann, F., Tiedemann, H., Prokott, E., Maiello, G., Fleming, R. W. (2021). An image-computable model of human visual shape similarity. *PLoS Computational Biology, 17*(6): e1008981. https://doi.org/10.1371/journal.pcbi.1008981

Under review:

Prokott, K. E. & Fleming, R. W. (*under review*). Identifying specular highlights: insights from deep learning

Tamura, H., Prokott, K. E., & Fleming, R. W. (*under review*). Distinguishing mirror from glass: A 'big data' approach to material perception. arXiv:1903.01671

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und ohne unzulässige Hilfe oder Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder nichtveröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht. Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten sowie ethische, datenschutzrechtliche und tierschutzrechtliche Grundsätze befolgt. Ich versichere, dass Dritte von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen, und dass die vorgelegte Arbeit weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde zum Zweck einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt wurde. Alles aus anderen Quellen und von anderen Personen übernommene Material, das in der Arbeit verwendet wurde oder auf das direkt Bezug genommen wird, wurde als solches kenntlich gemacht. Insbesondere wurden alle Personen genannt, die direkt und indirekt an der Entstehung der vorliegenden Arbeit beteiligt waren. Mit der Überprüfung meiner Arbeit durch eine Plagiatserkennungssoftware bzw. ein internetbasiertes Softwareprogramm erkläre ich mich einverstanden.