Human gloss perception reproduced by tiny neural networks

Takuma Morimoto[1, 2*], Arash Akbarinia[1], Katherine R. Storrs[3], Jacob R. Cheeseman[1], Hannah E. Smithson[2], Karl R. Gegenfurtner[1], and Roland W. Fleming[1]

1: Department of Psychology, Justus-Liebig-Universität Gießen, Germany

2: Department of Experimental Psychology, University of Oxford, Oxford, UK

3: School of Psychology, University of Auckland, New Zealand

*Corresponding author: Takuma Morimoto

Email: takuma.morimoto@psy.ox.ac.uk

Keywords: human vision, material perception, convolutional neural networks, online experiment

Abstract

A key goal of vision science is to uncover the computations involved in perceiving visual properties like colour, curvature, or glossiness. Here, we used machine learning as a data-driven tool to identify potential computations of gloss perception. We generated thousands of object images using computer graphics—varying lighting, shapes, and viewpoints—and experimentally measured perceived glossiness for each image. Observers showed curious patterns of agreement and disagreement with the physical reflectance in their gloss estimates, yet their judgments remained highly consistent both within and across individuals. We then compared two sets of neural networks: one set trained to mimic the human responses ('human-like networks') and another trained on physical labels to approximate physical reality ('ground-truth networks'). We progressively reduced the size of the networks to identify the minimum computations capable of meeting the two objective functions. While quite deep networks were required to estimate physical reflectance, we found that surprisingly shallow networks—with as few as three convolutional layers—could accurately replicate human gloss judgments. Indeed, even a miniscule 'network' with just a single filter could predict human judgments better than the best ground-truth network. The human-like networks also successfully predicted some known perceptual gloss effects beyond the training range. Our findings suggest that humans do not judge material properties through complex computations such as inverse optics; instead, gloss perception arises from simpler computations useful for other visual tasks.

Introduction

In daily life, we encounter materials with a wide range of visual qualities, such as those generated by metal, cloth, or plastic [1]. In a brief glance, we can judge their colour, lightness, shape, glossiness, and translucency. We can infer their associated physical properties, such as light/heavy, soft/hard,

rough/smooth, and even their states, such as wet/dry, clean/dirty, solid/melted [2]. Humans use visually sampled information to guide our behaviors, e.g. lifting heavy objects slowly.

Among the various material properties that we can perceive visually, gloss is the most widely studied [3]. Past studies have proposed various cues the visual system may use to determine the glossiness of an object's surface, including the skewness of the luminance histogram [4-7], the luminance gradient [8], the standard deviation of luminance over a surface [9], and various image metrics derived from specular reflection patterns [10,11]. One significant challenge is that an object with specific material properties can differ substantially in visual appearance, depending on factors such as shape, lighting, and viewpoint. In response to this challenge, many behavioral studies have measured gloss perception across a wide variety of experimental conditions [12-27]. Yet, no single model fully captures human gloss perception across these diverse factors. Here, we sought to identify biologically plausible, image-computable models that account for human gloss judgments across diverse images.

One persistent challenge for such traditional cue-based approaches is that researchers must somehow invent candidate visual cues in advance to evaluate their efficacy. This can be a hurdle when the cues the visual system relies on may be too complex or abstract to formulate. Recent studies have taken a data-driven approach, seeking deep learning models that mirror the pattern of human gloss judgments [28,29]. Our approach differs fundamentally by aiming to create human-like convolutional neural networks (CNNs) by directly training them on human perceptual data. We gathered gloss judgments from hundreds of human observers for thousands of object images in an online experiment. Through optimization, the CNNs spontaneously acquired a set of filters useful for reproducing human-like gloss judgments, without manual feature selection. This allows exploration of a much larger search space, not limited by the experimenter's imagination. We then systematically adjusted the depth of the CNNs to understand the computational power required to predict human gloss judgments, aiming to make them as shallow as possible to increase interpretability [30-32]. We then interrogated the mechanisms that emerged within the networks. This allows for a data-driven, hypothesis-free approach to discovering the computations underlying human perception across diverse viewing conditions [33].

Figure 1 summarizes a schematic overview of our approach. We generated 3,888 images from the full combination of 36 lighting environments, 36 object geometries, and 3 viewpoints. Using an online crowd-sourced experiment, each image was labeled based on human gloss judgments, first averaged across two repetitions per observer and then averaged across at least three observers. The observer's task was to adjust the specular reflectance parameter of a reference object until the test and reference objects appeared to have the same glossiness. The quality of the data was confirmed through laboratory-based validation experiments (see **Supporting Information, SI**). We then trained two sets of CNNs with various depths: one on images with human labels ('human-like networks') and another on images with physical ground-truth labels ('ground-truth networks'). To ensure that the networks learned only the gloss computation—rather than tasks such as image segmentation—we trained them on images where the object was shown against a uniform gray background (R = G = B = 127). Finally, we

compared the performance of the two sets of CNNs and examined the internal computations of the human-like networks to gain insights into mechanisms that reproduce human gloss perception.

Our findings suggest that relatively shallow CNN architectures with as few as three convolutional layers are sufficient to predict human gloss judgments, while inferring the physical ground-truth requires more complex computations for equivalent accuracy. Analyzing the emerging filters in the human-like networks revealed that a filter with a center-surround receptive field plays a critical role in human gloss judgments. This insight indicates that what seems to be a complex visual task, such as judging material properties, may be resolved through a set of simple computations that are also used for other visual tasks.

Results

Perceptual experiments. Each gray circle in **Figure 2A** represents one of the 3,888 test images. The axes display ground-truth specular reflectance values and observer settings (averaged over observers). Notably, the average observer settings significantly deviate from the ground-truth reflectance of the objects. The histogram of Pearson's correlation coefficients for each of the 295 observers shows a median correlation value of 0.46, with no observer exceeding 0.75 (**Figure 2B**). This indicates that gloss constancy—where observers consistently perceive the glossiness of a surface despite changes in shape, lighting, and viewpoint—was surprisingly poor for our image set.

However, crucially these gloss judgments were highly consistent both within and across observers (**Figures 2C** and **2D)**. Intra-observer correlations were calculated between sessions 1 and 2 for each observer over 84 images (Figure 2C). Each inter-observer correlation was computed over 84 images between one observer and the average of the rest who judged the same set (left plot in Figure 2D), or between one observer and another for all possible observer pairs (295*294*0.5 = 43,365 pairs), over 12 common images (right plot). This confirms that the large deviation from the physical ground-truth was not due to random variation in observer judgments. Instead, observers made systematic and consistent patterns of agreement and disagreement with physical reflectance in gloss estimation. A separate, lab-based experiment validated that the online data are of comparable quality (SI, Figure S4).

Each grid in **Figure 2E** shows eight example images in each pair of physical ground-truth and observer settings (low, mid and high). Images in the diagonal arrays (low-low, mid-mid, high-high) display objects where ground-truth and human judgments agree. Images in the top left and bottom right grids (low-high, high-low) highlight interesting cases where human judgments and physical ground-truth disagree substantially. Such images effectively provide different objectives for networks trained on physical ground-truth versus human judgments. Although human judgments do correlate somewhat with ground-truth, there is also sufficient consistent deviation from physical ground-truth to allow us to capture the key idiosyncratic computational signatures of human gloss perception.

Computational Models. Figure 3 compares observers and candidate models based on their correlation to the physical ground-truth (x-axis) and human judgments (y-axis). The upper left half of the plot (pale blue region) indicates human-like models. The other half (pale pink region) indicates models that correlate more closely with the physical ground-truth. Each dark green circle represents an individual observer, where the y-axis shows the correlation to the average of the rest of the observers who judged the same 72 images. Thus, in this plot, we are seeking models positioned close to this green circle distribution.

Gray diamonds show models based on luminance statistics models (mean, median, first quartile, third quartile, minimum, maximum, standard deviation, skewness, kurtosis), which are positioned far from the human distribution. When these cues are combined using multiple linear regression fit to predict human responses, the correlation with human responses increases (indicated by the gray star symbol). Light green downward triangles show models that use sharpness, coverage and sub-band contrast computed from specular reflection images [10]. These models correlate more with physical ground-truth than with human observers and therefore do not effectively capture human perception in our task. Combining these cues yielded no improvement (the light-green star symbol).

The blue triangles and squares show CNNs trained on human responses with one and three convolutional layers, respectively. The number next to each symbol shows the number of kernels in each convolutional layer. As expected, the deeper networks, such as three-layer with 64 kernels, came closer to approximating the mean glossiness judgments across all observers (black cross). Yet it is extremely intriguing and noteworthy that a CNN with only one kernel (single-kernel model) comes close to the edge of the human distribution, with a correlation coefficient of 0.64. Given the correlation coefficient across observers was around 0.85, this very light model reaches 75.3% of this performance ceiling and indeed approximates human perception better than any of the models trained on physical ground-truth labels that we tested.

In contrast to the CNNs trained on human labels, the networks trained on physical ground-truth labels, shown by magenta symbols, struggle to achieve such impressive performance. Although trained to recover physical labels, some physical ground-truth networks are positioned near human-like networks, and even the three-layer models fall in the blue half of the plot. To predict physical labels effectively, more complex architectures were required. A ResNet18 (magenta rightward triangle) was needed to cross the diagonal unity line, highlighting the computational challenges involved in estimating physical ground-truth. Using additional training images rendered using novel lighting environments and object shapes, the network's ability to predict ground-truth substantially improved, as shown by orange symbols. Three-layer models with 64 kernels (orange square) still show a limited correlation below 0.70, even with additional 3.8×10^5 images. A ResNet18 trained with these additional images shows a correlation around 0.9 to ground-truth, but importantly, the correlation to human responses is below 0.5. The intriguing aspect of this observation is that human strategies for estimating material properties such as specular reflectance are unlikely to rely on complex computations that aim to recover physical

parameters, as suggested by the inverse optics approach. Instead, it appears that humans use relatively simple computations to intuit complex material properties.

Investigating internal representations of the neural network models. Some examples of kernels that emerged internally in the human-like CNNs are illustrated in **Figures 4A** and **4B**. The kernel in the single-kernel model displays a bright blob surrounded by a dark region, which seems well suited for identifying highlights given their typical spatial intensity profile. It is interesting that the bright portion has a yellowish tint, while the surround is blueish, reflecting typical colour relationships between direct (e.g., sun) and indirect (e.g., sky) illumination sources in the natural environment. We found that the bright blob region can be effectively fitted with a difference of Gaussian (DoG) function (**Figure S5**), a typical receptive field observed as early as retinal ganglion cells [34]. Although such a simple filter was far from the best of the models we tested for predicting human gloss judgments, it is striking just how large a proportion of the variance in human judgments it does account for (75.3%). Relatedly, in the three-layer model, 9 example kernels in the first convolutional layer show somewhat similar features, but with different spatial scales, contrasts, color tunings, and polarities. Taken together, these observations suggest, our gloss judgments can be predicted surprisingly well by a combination of well-known, low-level filtering mechanisms within the visual system.

To further visualize the networks' behaviors, we plotted all 3,888 images in relation to the human responses and the model predictions, in **Figure 4C** and **D**. From the predictions of the one-layer model, we notice two points. First, the range of model responses is narrower than for the three-layer model, meaning very high gloss objects and matte objects were not well predicted. Second, images with a low mean luminance level were consistently judged as low gloss. In contrast, the three-layer model overcomes both points. Thus, one of the functions that the three-layer model acquired in the second and third convolutional layers may be to separate the dimensions of overall luminance level and glossiness (**Figure S6**).

Surprisingly, given the simplicity of the internal representations, we also found that the human-like networks predicted a number of known gloss effects [6] as shown in **Figure 5**. For instance, the networks accurately anticipated the decrease in perceived glossiness when the specular highlight is rotated (panel **A**), moved horizontally (panel **B**), when surface roughness increases (panel **C**), and when the underlying specular reflectance weakens (panel **D**). The decrease was steeper for the three-layer models compared to the one-layer model. Notably, these modified images were not included in the training dataset, demonstrating the models' ability to generalize beyond the training range. It is straightforward that perceived gloss decreases when surface roughness or specular contrast decreases, as this correlates with a decrease in local intensity of the specular component. It is a bit surprising, however, that even a single-kernel model can predict the decrease in gloss level for rotation and translations. The spatial pattern of the receptive field suggests the filter utilizes the local spatial intensity profile of the specular highlight region. For a typical glossy object, there is likely a certain local intensity profile that correctly aligned specular highlights create. When only the specular component

shifts from the original position independently of the diffuse component, this geometrical regularity collapses, leading to a decrease in the filter output. However, we recognize this is not always the case. Out of 3,888 test images, 790 images (about 20%) actually showed an overall increase in gloss levels due to highlight rotation. Nevertheless, it is intriguing that simple models can detect seemingly complex inconsistencies of highlight position. It suggests that while there is surely some degree of coupling between mid-level representations of shape, material and lighting [7,10,11, 35-37], a significant portion of gloss perception may actually be accounted for by more basic mechanisms.

Discussion

This study is inspired by a fundamental question in behavioral neuroscience: what neural computations enable material perception from highly variable sensory inputs across different contexts? We address this by leveraging CNNs to replicate human-like gloss judgments and explore the internal computational strategies of these networks, in a data-driven, hypothesis-neutral approach that contrasts with traditional cue-based methods. The motivation is that tasks such as material perception might be too complex for intuitive identification of underlying mechanisms, whereas CNNs excel at finding statistical regularities in training datasets. We used this tool to guide our search for biologically plausible computations that underlie human gloss perception.

This approach identified a clear contrast between the computational demands required to develop human-like CNNs and physical ground-truth CNNs, suggesting that human gloss perception likely does not involve sophisticated inverse optics computations. This notion has been suggested by others [1], but what is new here is the identification of specific filters and the depth of computation required to predict human gloss perception and physical reflectance parameters. In particular, the effectiveness of a DoG-like filter is intriguing, directly supporting the idea that gloss perception relies, to a nontrivial extent, on relatively simple image features, as suggested by Motoyoshi et al. [4]. Much as cone photoreceptors provide foundational signals for color perception and other visual functions, bright blob detectors may serve as a key component of gloss perception, and perhaps other distal perceptual attributes too [38-40]. While we certainly are not suggesting that the human gloss perception is based *exclusively* on low-level filters, it is intriguing that when a large population is asked to judge gloss under diverse conditions, they resort to strategies that can be well approximated by such measurements. In addition to this computational perspective, a key behavioral finding is the significant deviation of observer responses from the underlying physical reflectance parameters. Clearly, gloss perception is not merely a direct and accurate estimate of physical properties. These findings collectively indicate that the goal of the visual system is not to accurately estimate underlying physical properties but to create perceptual representations of objects that allow us to distinguish stimuli as efficiently as possible.

One key aspect of our approach is the use of large-N psychophysical experiments. Since numerous distal factors such as shape, lighting, and reflectance properties are involved in generating a single object image, covering a wide range of stimulus space—ideally, comparable to the diversity of natural images—is critical to capture visual behaviors with sufficient resolution. Conversely, when we used only

a small number of shapes or lighting environments, we found that perceptual gloss judgments correlated well with physical ground-truth in many cases (**Figures S2 and S3**). This correlation is not inherent to our perceptual judgment but is influenced by the stimulus set. Recent studies have made significant efforts to collect extensive human judgment data across a large number of images [41-43]. To understand visual mechanisms, we must account for both successes and errors in perception, and large-scale measurement helps capture these diagnostic behavioral patterns. Although careful data curation and validation are essential to this approach, it offers a powerful data-driven framework for exploring the complex mechanisms underlying human perception and behavior. Combining our approach with unsupervised learning can help reduce the cost of collecting labeled images [44-46].

Two closely related works utilize perceptual data to train deep learning models. Lagunas et al. [42] applied transfer learning with the deep ResNet architecture, transitioning from an object recognition task to a material similarity judgment task, successfully replicating human similarity judgments. More recently, Serrano et al. [43] generated 42,120 images using a full combination of 9 lighting environments, 9 object geometries, and 520 bidirectional reflectance distribution functions (BRDFs), measuring perceptual ratings of various material properties including glossiness. They trained VGG16 and ResNet52 to recover perceptual responses with high accuracy. Our study differs significantly in motivation, as neither of the previous studies aims to understand the mechanisms of human material perception. Additionally, our approach focuses on identifying the *shallowest* possible neural networks to maximize the interpretability of the computations that emerge within the network [30-32]. Interestingly, although our model was tasked with predicting gloss levels only, the required computational architecture was much lighter in our case.

One outcome of this study is the creation of a publicly available database containing 3,888 images. While many databases exist for labeled objects, those focusing on material properties are less common. Each image is annotated with full stimulus parameters, including physical ground-truth labels and perceptual labels from at least three different observers. We expect this dataset to be useful in multiple disciplines such as vision science, cognitive neuroscience, and computer vision. In addition, the availability of perceptual labels is useful for industrial applications, such as predicting the perceived glossiness of car paint, enhancing the presentation of jewelry, or achieving specific visual impressions in artworks.

To conclude, our study has provided a novel data-driven perspective on the computational mechanisms in human gloss perception using large-scale perceptual measurement and CNN-based modeling. We found a relatively simple and biologically plausible model that predicts the idiosyncrasies of human gloss constancy, as well as several known gloss effects outside the training range.

Materials and Methods

Stimuli

The detailed procedure for stimulus generation is explained in our previous study [47]. All reference and test images were generated using the physically-based rendering software Mitsuba v 0.6 [48]. All images contained a single object at the center of the scene, to which we applied diffuse and specular reflectance using the Ward reflectance model [49, 50] with a fixed surface roughness level of 0.05. The specular reflectance value was randomly sampled from zero (matte) to 0.0999 (highly glossy). We applied image-based environmental illumination to the scene, with the camera set at the same height as the object and pointing directly at it. The asymmetric gloss matching task required a matching object whose physical specular reflectance parameter could be continuously adjusted by observers. For this, we used Pellacini's c [51], a perceptually-linear characterization of Ward specular reflectance, instead of making changes in image space. A total of 3,888 test images were generated by combining 36 lighting environments, 36 object shapes, and 3 random viewpoints. For each test image, a random body color was used, with fixed lightness at 50, and random chroma and hue between 8-26 and 0-360 degrees in the L*a*b* perceptually-uniform color space [52], respectively. A full list of object shapes and environmental illuminations is shown in **Figure S1**.

All images were generated at 512×512 pixels in XYZ format. After converting the raw XYZ image to a linear monitor RGB format, images were normalized separately for each lighting environment, using the 97th percentile pixel value computed across the 108 images within that environment. This operation ensured consistency in overall light levels across images within each lighting environment. These scaled linear RGB images were then gamma-corrected using the standard sRGB formula. This spatial resolution of 512×512 pixels was used for the psychophysical experiment while this was reduced to 224×224 pixels for the network's inputs to shorten the training time.

For the online experiment, a total of 3,888 images were divided into 54 independent sets. Twelve common images were added to each set to evaluate the consistency of observer responses. For two of these sets, we recruited 21 and 23 observers to validate the online data with a lab-based experiment (see **SI**). At least three observers were recruited for each set.

Experiment

The main experiment was conducted online using PsychoPy (version 2022.1.4) [53] and hosted on an online platform Pavlovia. Observers were recruited via Prolific and compensated 12 euros per hour. Before participating, they downloaded and read an information sheet explaining the procedure and signed a digital consent form. A validation experiment was conducted in the laboratory (detailed in SI, Figure S4).

Observers and Ethics

A total of 467 observers participated in the online experiment, with data from 295 analyzed (see exclusion criteria in a later section). Observers had normal or corrected-to-normal visual acuity and

normal color vision (self-report). Their age ranged from 19 to 65 (mean 41.4; S.D. 12.4). The observer group comprised 159 females and 136 males. Informed consent forms were in English, so only English-speaking observers were recruited for the online experiment. No language criteria were applied for the laboratory experiment. No observer was informed of the experiment's purpose. The study was approved by the local ethics committee at Justus Liebig University Giessen, following the Helsinki Declaration (sixth revision, 2008).

Procedure and task

Observers were instructed to turn off the room lights and sit at arm's length from the display during the experiment. Observers viewed the screen binocularly. Before the experiment, observers completed a size calibration using a bank card [54] to ensure the stimulus image appeared at the correct size. In each trial, a test image was shown on the left and a reference image on the right. The task of observers was to move a slider under the reference image to match its perceived glossiness with the test image. The reference image stayed the same in all parameters except Pellacini's c, while the test image varied between trials in terms of shape, color, viewpoint and lighting environment (asymmetric matching [55,56]). After completing the size calibration, the observers received on-screen instructions: "Adjust the gloss level of the right 'reference' object using the slider until it matches the gloss level of the left 'target' object." The slider's starting position was randomized for each trial, excluding edge values, which were reserved for catch trials used to detect autoresponders. The first three trials were practice rounds to help observers get used to the task. After practice, they adapted for one minute to a uniform white screen (R=G=B=127). Each observer was randomly assigned to one of 54 image sets, each containing a set comprising 84 images (72 unique to the set and 12 common images). Each session consisted of 84 trials, and all observers completed 2 sessions, totaling 168 trials. To detect autoresponders, each session included a catch trial in which no stimuli were shown; instead, a text screen instructed observers to move the slider to the left edge before proceeding to the next trial. The order of stimulus presentation was randomized. Median time across 295 observers to complete the experiment was 13.7 mins (4.89 sec per trial).

Exclusion criteria

For quality control, we set two exclusion criteria. First, observers who failed either of the catch trials were excluded. Second, observers who completed each trial too quickly were excluded; this was indicated by a mean response time shorter than the mean minus two standard deviations of laboratory-based experiments. We continued our experiments until at least 3 observers passed these criteria for each image set. This exclusion process resulted in 295 out of 467 observers (63.2%) being included in the final analysis.

Computational Modeling

Luminance statistics models.

The Y (luminance) channel from raw XYZ images was extracted. We implemented nine models to determine glossiness using the mean, median, standard deviation (std.), skewness, kurtosis, first

quartile (Q1), third quartile (Q3), and the minimum and maximum pixel luminance values within the object region of each test image, excluding the surrounding region from these calculations.

Specular reflection models.

We computed three metrics—coverage, sharpness, and contrast [10]—derived from the specular reflection component of each image. Each model involved free parameters, which were optimized to align most closely with human gloss judgments. For each test image, we rendered a version of the object containing only the specular component by setting the diffuse reflectance to zero, and converted the resulting image to a luminance image. To extract direct specular reflections and exclude secondary and higher-order interreflections, we selected only those pixels whose intensities exceeded a certain percentage (k%) of the maximum intensity in the specular image. The value of k was selected from the set {0, 1, 3, 5, 10, 20, 40}. Based on this thresholded highlight image, we computed the three metrics. Coverage was defined as the proportion of the object's region occupied by highlights. Sharpness was calculated using spatial convolution to detect regions with rapid luminance changes [57]. For contrast, rather than using the raw highlight image, we applied a Gaussian band-pass filter to decompose the image into seven spatial frequency sub-bands, with cutoff ranges of 1.5–3.0, 3.0–6.0, 6.0–12.0, 12.0–24.0, 24.0–48.0, 48.0–96.0, and 96.0–192.0 cycles per image. This approach follows previous findings suggesting that certain frequency channels contribute more significantly to gloss perception [58]. We then computed the root-mean-square contrast—equivalent to the standard deviation of pixel intensities—for each sub-band image, as well as for an aggregated image across all frequency bands. Unlike coverage and sharpness, which each depend on a single parameter, the sub-band contrast metric includes two free parameters: the pixel intensity threshold k and the spatial frequency band. These parameters were optimized separately for each metric to determine the values that best correlated with human gloss judgments.

Manipulation of the pattern of specular reflection

First, for each of the 3,888 test images, the diffuse and specular components were rendered separately. We then manipulated the specular reflection in three different ways (rotation, translation and roughness) before combining it with the diffuse component. The strength of the specular reflectance was set to a maximum value for all images (0.0999 in the Ward reflectance model, equivalent to 0.1487 in Pellacini's c) to clearly observe the effect of highlight manipulation.

For rotated specular reflection, we rotated the specular component from -90 to +90 degrees in 30-degree steps, using the image center as the rotation axis. For translation, we shifted the specular component horizontally from -30 to 30 pixels in 10-pixel steps. After these manipulations, a mask was applied to remove any specular reflection extending beyond the object region. For surface roughness, we rendered the specular component with varying roughness values of 0.001, 0.05, 0.10, 0.15, and 0.20. For surface contrast, we adjusted the strength of the specular component between 0, 0.0372, 0.0743, 0.1115, and 0.1487 in Pellacini's c unit. Then, we combined these manipulated specular images with the diffuse image.

All manipulated and original images were input to CNNs, and the proportional change in predicted gloss level relative to the non-manipulated image was computed.

Network architectures

We used a ResNet architecture for two reasons. First, it has proven effective in solving object recognition tasks [59], which are highly relevant to our visual task. Second, it consists of several stages, each with a flexible number of convolutional layers, allowing us to systematically adjust architectural complexity.

The one-layer model consists of a single convolutional layer, followed by max pooling and a regression layer to predict continuous gloss values. It does not include non-linear operations such as ReLU or batch normalization. The three-layer model begins with a convolutional layer followed by ReLU, batch normalization and max pooling, then two additional convolutional layers with a skip connection that allows the input to bypass them if beneficial. This is followed by average pooling, and finally, a regression layer.

In the first convolutional layer, the kernel size was 15 × 15 pixels for the one-layer model and 7 × 7 pixels for the three-layer model, determined through preliminary exploration of suitable architectures. The number of kernels varied between 1, 2, 4, and 9 for the one-layer model, and 9, 16, 32, and 64 for the three-layer model. Networks were trained for 90 epochs until reaching a performance plateau, starting with an initial learning rate of 0.01, which decreased by a factor of 10 every 30 epochs. We used the Adam optimizer and mean absolute error as the evaluation function. Model training was conducted on a DGX A100 system (four NVIDIA A100 GPUs each with 80GB video memory, Dual AMD Rome 7742 CPU, 128 cores total, 2.25 GHz base) running Ubuntu 22.04 LTS.

Training and validation procedure

Our evaluation procedure consists of two independent sets of cross-validation. In the first set, networks were trained on 33 lighting environments (3,564 images) and tested on 3 novel lighting environments (324 images). This process was repeated 12 times until all lighting environments were used as test sets. In the second set, networks were trained on 33 shapes (3,564 images) and tested on the 3 novel lighting environments (324 images). Again, this was repeated 12 times until all shapes were used as test sets. Thus, each network architecture underwent 24 unique training and testing pairs, resulting in 24 trained models. The correlation coefficient reported in Figure 3 is the average value across these 24 networks, each computed from 324 test images. Luminance statistics models and specular reflection models were evaluated in the same way for consistency, although no fitting was involved for the luminance statistics models.

For each training session, the 3,564 images were augmented to 320,760 images through horizontal flipping and random size cropping, ensuring that 80% of the object region was preserved. This

conservative augmentation approach was chosen because certain manipulations, such as vertical flipping, may disrupt the geometrical regularity in natural environments, potentially affecting the representativeness of gloss judgments.

Acknowledgments

The authors thank Annika Zentel for assisting with data collection for the laboratory-based experiment. TM was supported by a Sir Henry Wellcome Postdoctoral Fellowship from Wellcome Trust (218657/Z/19/Z). This research was also supported by “The Adaptive Mind,” funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research, by the DFG SFB-TRR-135 “Cardinal Mechanisms of Perception” (project No. 222641018, TPs C1 and C2), and by a Marsden Fast Start Grant to KS from the Royal Society of New Zealand (project MFP-UOA2109), and European Research Council Advanced Grant Color3.0 (project number 884116) to KRG and European Research Council Advanced Grant STUFF (project number 101098225) to RWF. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

1. Fleming RW. Material perception. Annu Rev Vis Sci. 2017;3:365-388.
2. Nishida S. Image statistics for material perception. Curr Opin Behav Sci. 2019;30:94-99.
3. Chadwick A, Kentridge R. The perception of gloss: A review. Vision Res. 2015;109(Part B):221-235.
4. Motoyoshi I, et al. Image statistics and the perception of surface qualities. Nature. 2007;447:206-209.
5. Sharan L, et al. Image statistics for surface reflectance perception. J Opt Soc Am A. 2008;25:846-865.
6. Anderson BL, Kim J. Image statistics do not explain the perception of gloss and lightness. J Vis. 2009;9:10.
7. Kim J, Anderson BL. Image statistics and the perception of surface gloss and lightness. J Vis. 2010;10:3.
8. Sawayama M, Nishida SY. Material and shape perception based on two types of intensity gradient information. PLoS Comput Biol. 2018;14:e1006061.
9. Wiebel C, Toscani M, Gegenfurtner KR. Statistical correlates of perceived gloss in natural images. Vis Res. 2015;115B:175-187.
10. Marlow PJ, Kim J, Anderson BL. The perception and misperception of specular surface reflectance. Curr Biol. 2012;22:1909-1913.
11. Marlow PJ, Anderson BL. Generative constraints on image cues for perceived gloss. J Vis. 2013;13:2.
12. Fleming RW, Dror RO, Adelson EH. Real-world illumination and the perception of surface reflectance properties. J Vis. 2003;3:347-368.

13. Zhang F, de Ridder H, Barla P, Pont S. A systematic approach to testing and predicting light-material interactions. J Vis. 2019;19:11.
14. Zhang F, de Ridder H, Barla P, Pont S. Effects of light map orientation and shape on the visual perception of canonical materials. J Vis. 2020;20:13.
15. Gigilashvili D, Islam AJ. The role of shape in modeling gloss. Proc 30th Color Imaging Conf. 2022;271-276.
16. Gigilashvili D, Thomas J, Pedersen M, Hardeberg JY. On the appearance of objects and materials: Qualitative analysis of experimental observations. J Int Colour Assoc. (JAIC) 2021;27:26-55.
17. Vangorp P, Laurijssen J, Dutre P. The influence of shape on the perception of material reflectance. ACM Trans Graph. 2007;26:77.
18. Chen B, Wang C, Piovarči M, Seidel HP, Didyk P, Myszkowski K, Serrano A. The effect of geometry and illumination on appearance perception of different material categories. Vis Comput. 2021;37:2975–2987.
19. Wendt G, Faul F, Ekroll V, Mausfeld R. Disparity, motion, and color information improve gloss constancy performance. J Vis. 2010;10:7.
20. Adams WJ, Kucukoglu G, Landy MS, Mantiuk RK. Naturally glossy: Gloss perception, illumination statistics, and tone mapping. J Vis. 2018;18:4.
21. Honson V, et al. Effects of shape, roughness and gloss on the perceived reflectance of colored surfaces. Front Psychol. 2020;11:485.
22. Cheeseman JR, Ferwerda J, Morimoto T, Fleming R. Gloss discrimination: Towards an image-based perceptual model. PsyArXiv. 2024. https://doi.org/10.31234/osf.io/anx4q
23. Wijntjes MWA, Pont SC. Illusory gloss on Lambertian surfaces. J Vis. 2010;10(9):13.
24. Radonjić A, Cottaris NP, Brainard DH. The relative contribution of color and material in object selection. PLoS Comput Biol. 2019;15(4):e1006950.
25. Norman JF, Todd JT, Phillips F. Effects of illumination on the categorization of shiny materials. J Vis. 2020;20(5):2.
26. Ho YX, Landy MS, Maloney LT. Conjoint measurement of gloss and surface texture. Psychol Sci. 2008;19(2):196–204.
27. Doerschner K, Maloney LT, Boyaci H. Perceived glossiness in high dynamic range scenes. J Vis. 2010;10(9):11.
28. Prokott KE, Tamura H, Fleming RW. Gloss perception: Searching for a deep neural network that behaves like humans. J Vis. 2021;21:14.
29. Tamura H, Prokott KE, Fleming RW. Distinguishing mirror from glass: A "big data" approach to material perception. J Vis. 2022;22:4.
30. Goncalves NR, Welchman AE. "What Not" detectors help the brain see in depth. Curr Biol. 2017;27:1403-1412.e8.
31. Rideaux R, Welchman AE. But still it moves: Static image statistics underlie how we see motion. J Neurosci. 2020;40:2538-2552.

32. Rideaux R, Welchman AE. Exploring and explaining properties of motion processing in biological brains using a neural network. J Vis. 2021;21:11.
33. Anderson BL, Storrs KR, Fleming RW (2023). Where do the hypotheses come from? Data-driven learning in science and the brain. Behavioral and Brain Sciences. 46: e386.
34. Young RA. The Gaussian derivative model for spatial vision: I. Retinal mechanisms. Spatial Vis. 1987;2(4):273–293.
35. Marlow PJ, Todorović D, Anderson BL. Coupled computations of three-dimensional shape and material. Curr Biol. 2015;25(6):R221–R222.
36. Anderson BL. Mid-level vision. Curr Biol. 2020;30(3):R105–R109.
37. Anderson BL, Marlow PJ. Perceiving the shape and material properties of 3D surfaces. Trends Cogn Sci. 2023;27(1):98–110.
38. Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis. 2004;60(2):91–110.
39. Akbarinia A, Parraga CA. Colour constancy beyond the classical receptive field. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):830–845.
40. Akbarinia A, Parraga CA. Feedback and surround modulated boundary detection. Int J Comput Vis. 2018;126(6):583–603.
41. Sawayama M, et al. Visual discrimination of optical material properties: A large-scale study. J Vis. 2022;22:17.
42. Lagunas M, et al. A similarity measure for material appearance. ACM Trans Graph. 2019;38:Article 135.
43. Serrano A, et al. The effect of shape and illumination on material perception: Model and applications. ACM Trans Graph. 2021;40:4.
44. Storrs KR, Anderson BL, Fleming RW. Unsupervised learning predicts human perception and misperception of gloss. Nat Hum Behav. 2021;5:1402-1417.
45. Guerrero-Viu J, et al. Predicting Perceived Gloss: Do Weak Labels Suffice?. Comput Graph Forum. 2024;43:e15037.
46. Liao C, Sawayama M, Xiao B. Unsupervised learning reveals interpretable latent representations for translucency perception. PLoS Comput Biol. 2023;19(2):e1010878.
47. Morimoto T, Akbarinia A, Storrs K, et al. Color and gloss constancy under diverse lighting environments. J Vis. 2023;23(7):8, 1–25.
48. Jakob W. Mitsuba: Physically based renderer. Retrieved from https://www.mitsuba-renderer.org/ (2010).
49. Ward GJ. Measuring and modeling anisotropic reflection. Proc 19th Annu Conf Comput Graph Interact Tech. 1992;265-272.
50. Geisler-Moroder D, Dür A. A new ward BRDF model with bounded albedo. Comput Graph Forum. 2010;29(4):1391–1398.
51. Pellacini F, Ferwerda JA, Greenberg DP. Toward a psychophysically-based light reflection model for image synthesis. Proc 27th Annu Conf Comput Graph Interact Tech (SIGGRAPH '00). 2000;55-64.

52. Carter EC, Schanda J, Hirschler R, Jost S, Luo MR, Melgosa M, Ohno Y, Pointer MR, Rich DC, Viénot F, Whitehead L, Wold JH. CIE 015:2018 Colorimetry. 4th ed. 2018.
53. Peirce JW, Gray JR, Simpson S, et al. PsychoPy2: experiments in behavior made easy. Behav Res Methods. 2019. https://doi.org/10.3758/s13428-018-01193-y
54. Li Q, Joo SJ, Yeatman JD, et al. Controlling for Participants' Viewing Distance in Large-Scale, Psychophysical Online Experiments Using a Virtual Chinrest. Sci Rep. 2020;10:904. https://doi.org/10.1038/s41598-019-57204-1
55. Arend L, Reeves A. Simultaneous color constancy. J Opt Soc Am A. 1986;3:1743-1751.
56. Brainard DH, Wandell BA. Asymmetric color matching: How color appearance depends on the illuminant. J Opt Soc Am A. 1992;9:1433-1448.
57. Vu CT, Phan TD, Chandler DM. S3: A spectral and spatial measure of local perceived sharpness in natural images. IEEE Trans Image Process. 2012;21:934-945.
58. Boyadzhiev I, Bala K, Paris S, Adelson E. Band-sifting decomposition for image-based material editing. ACM Trans Graph. 2015;34:1-16.
59. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Proc IEEE Conf Comput Vis Pattern Recognit (CVPR). 2016;770-778. doi: 10.1109/CVPR.2016.90.

Figures

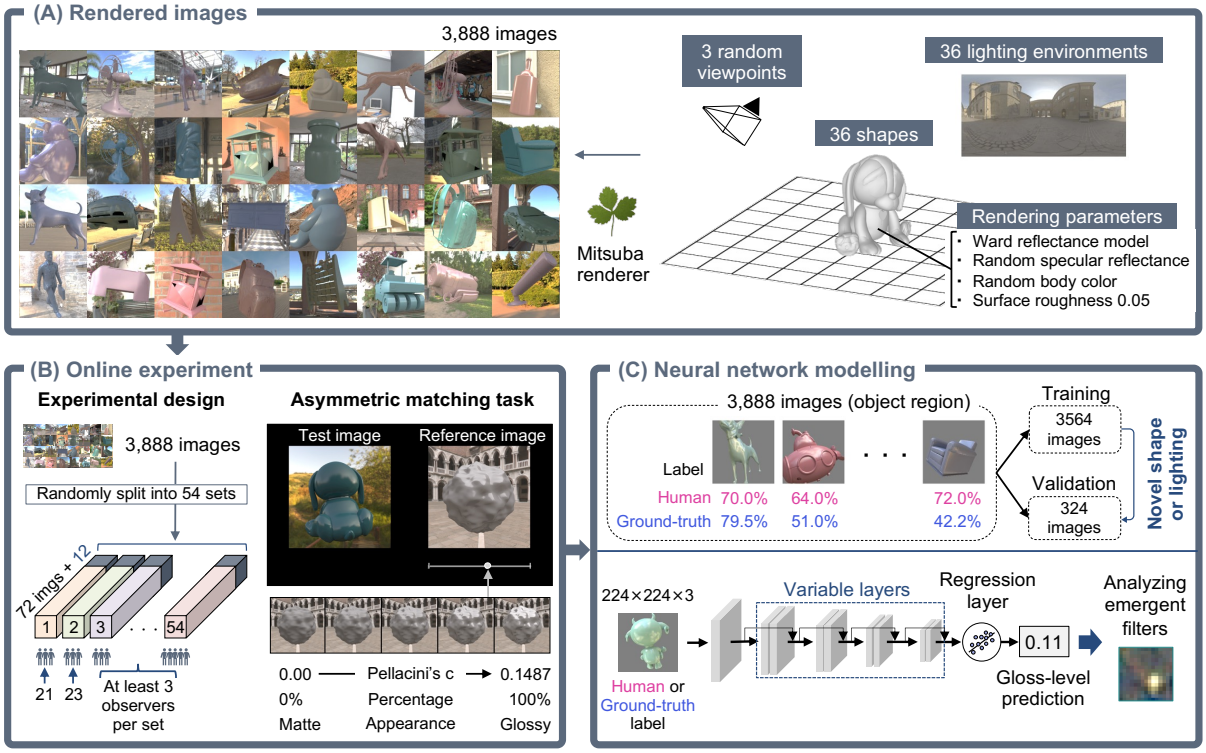

Figure 1. Schematic of our approach. (**A**) We rendered 3,888 images using computer graphics software. The scenes contained an object with a glossy surface (Ward reflectance model [37, 38], surface roughness fixed at 0.05). Random specular reflectance values and body colors were applied. One of 36 environmental lightings and one of 36 shapes were set, and three random viewpoints were selected per shape-lighting pair. (**B**) **Left**: Design of online experiment. The 3,888 images were randomly assigned to 54 sets of 72 images plus 12 additional images shared across sets to evaluate variation among observers. Two image sets were also tested in a lab-based experiment (see **Supplementary Material**), to validate the online data quality. At least three observers were recruited for each set. **Right:** stimulus configuration for the asymmetric gloss matching task. By moving a slider, observers adjusted the gloss (i.e., Pellacini's c [39]) of the reference object to match the perceived gloss level between the two objects. **(C)** The 3,888 images were divided into training and validation sets, ensuring the validation dataset consisted of three novel shape or lighting environments that were not included in the training set (producing 24 unique combinations of training and validation datasets). We trained CNNs with varying numbers of intermediate layers using images labeled by human gloss judgments ('human-like networks') or by physical ground-truth labels ('ground-truth networks'). Gloss levels predicted by the networks were compared against human responses. The trained networks were analyzed to understand the computational mechanisms that emerged within them.

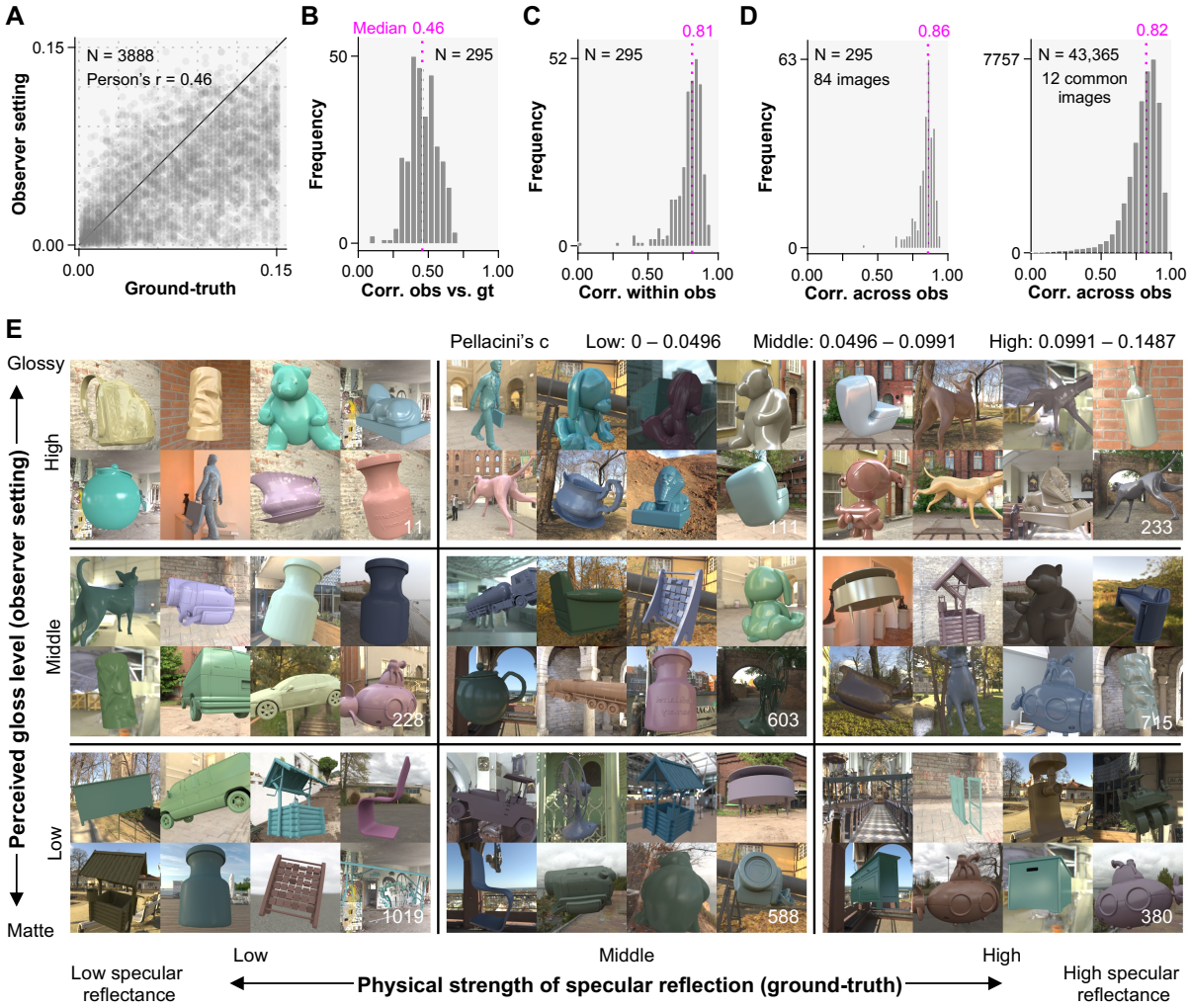

Figure 2. (A) Each data point shows a specific stimulus image, and the plot shows all 3,888 images. The horizontal and vertical axes show physical ground-truth reflectance values and observer settings (perceived gloss) averaged over observers. **(B)** Histogram of human vs. ground-truth correlation. The plot shows the distribution of Pearson's correlation coefficients between each observer's settings and the corresponding ground-truth values. **(C)** Histogram of within-observer correlations. For each observer, Pearson's correlation coefficient was computed between session 1 and session 2 to quantify the intra-observer variability. **(D)** Histogram of correlation across observers. For the left plot, the correlation was computed between one observer and the average across all other observers across 72 images unshared across image sets. For the right plot, the correlation was computed over 12 images for every pair of observers (i.e. (295×294)/2 = 43,365 pairs). **(E)** Each grid shows 8 example images. The horizontal and vertical positions in the grid show the physical specular reflectance assigned to each object and observer settings (perceived gloss level). The number of images in each grid is shown at the right bottom of each grid.

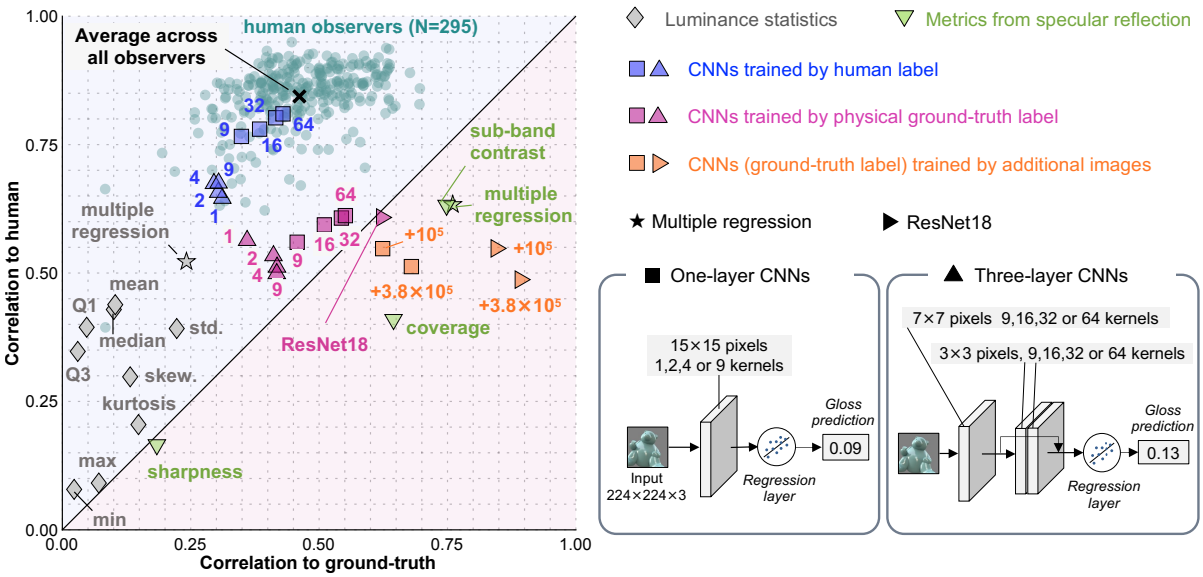

Figure 3. Dark green circles show individual human observers (N=295), where the x-axis and the y-axis show the correlation over 74 images to the physical ground-truth and the correlation to the average of the rest of the observers, respectively. Other data points show candidate computational models. Gray diamonds show low level luminance image statistics models, and the gray star shows a multiple regression model based on the luminance statistics (best fit to human data). Light green downward triangles show models based on statistics computed from specular reflection images (sharpness, coverage, and sub-band contrast), and the light green star symbol shows a multiple regression based on the specular metrics (best fit to human data). Triangles and squares show one-layer models with different kernel numbers (1, 2, 4 and 9; 9, 16, 32 and 64), where blue symbols refer to the networks trained on images labeled by observer judgments and pink symbols show networks trained on images labeled by physical ground-truth labels. Magenta rightward triangle shows ResNet18 trained on physical ground-truth label. Orange symbols represent CNNs trained with additional images, generated using novel shapes and lighting conditions not included in the 3,888 images. Orange squares correspond to the three-layer model with 64 filters, and orange rightward triangles show ResNet18. The orange numbers on the plot show the number of additional images used.

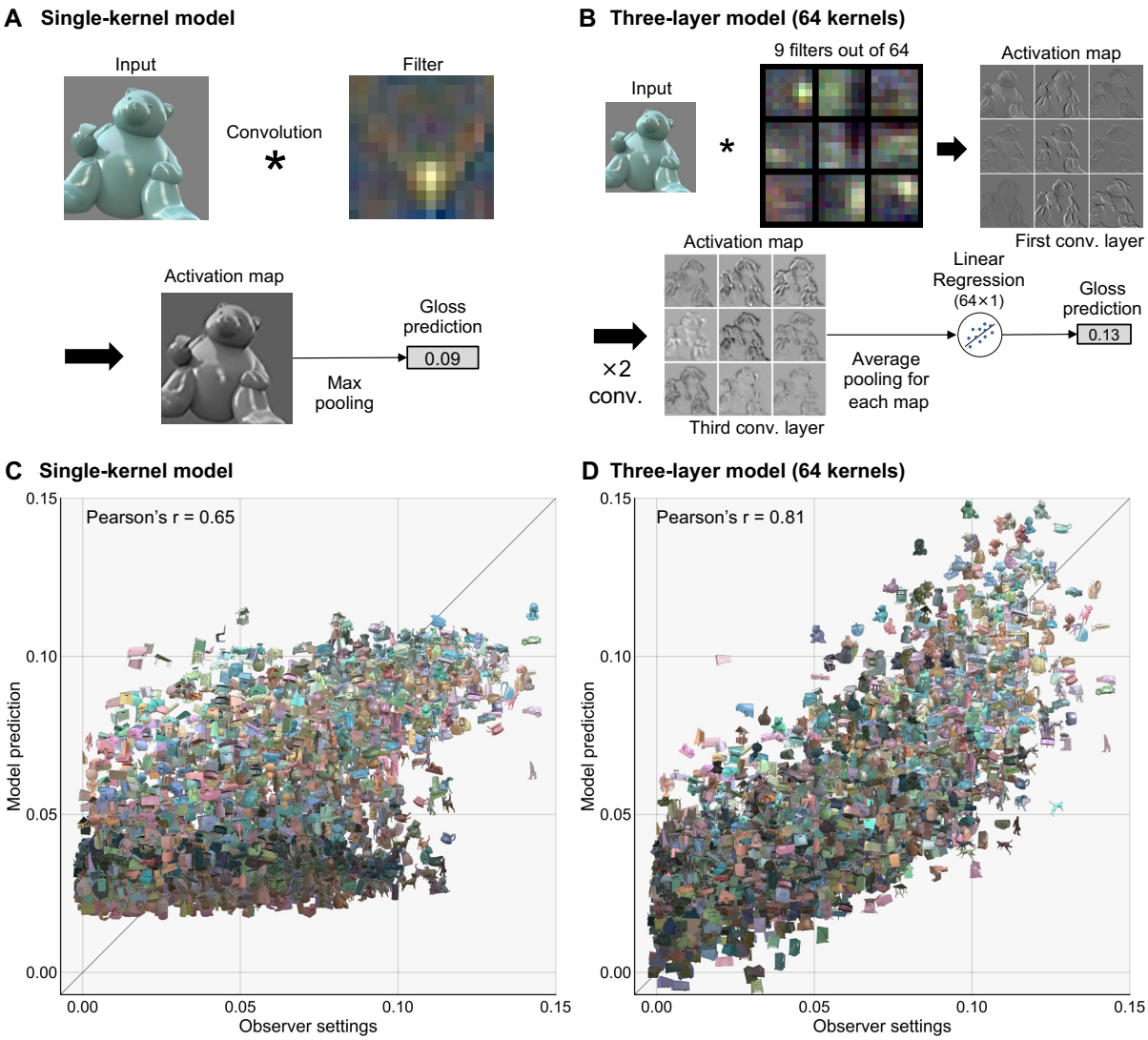

Figure 4. (A) Visualization of the internal computations for a one-layer neural network model trained on images with human labels. The input image was convoluted by a filter with 15 by 15 pixels and max pooling was taken on the convoluted activation image. **(B)** Visualization of a three-layer model with 9 filters (7 by 7 pixels) trained on images with human labels. In the first convolutional layer, the input image was convolved by a set of 9 kernels. In the second and third convolutional layer, activation maps from the first convolutional layer were further convolved by the two sets of kernels subsequently and average pooling was taken for each of 9 activation maps from the third convolutional layer. Then the pooled features were used as input for a linear regression model, which predicted gloss level. **(C)** Scatter plot of observer settings versus model predictions for the one-layer model. This plot combines responses from all 24 networks (see Training and Validation Procedure in Methods and Materials), ensuring that each image was never seen by the network during training. **(D)** Corresponding plot for the three-layer model.

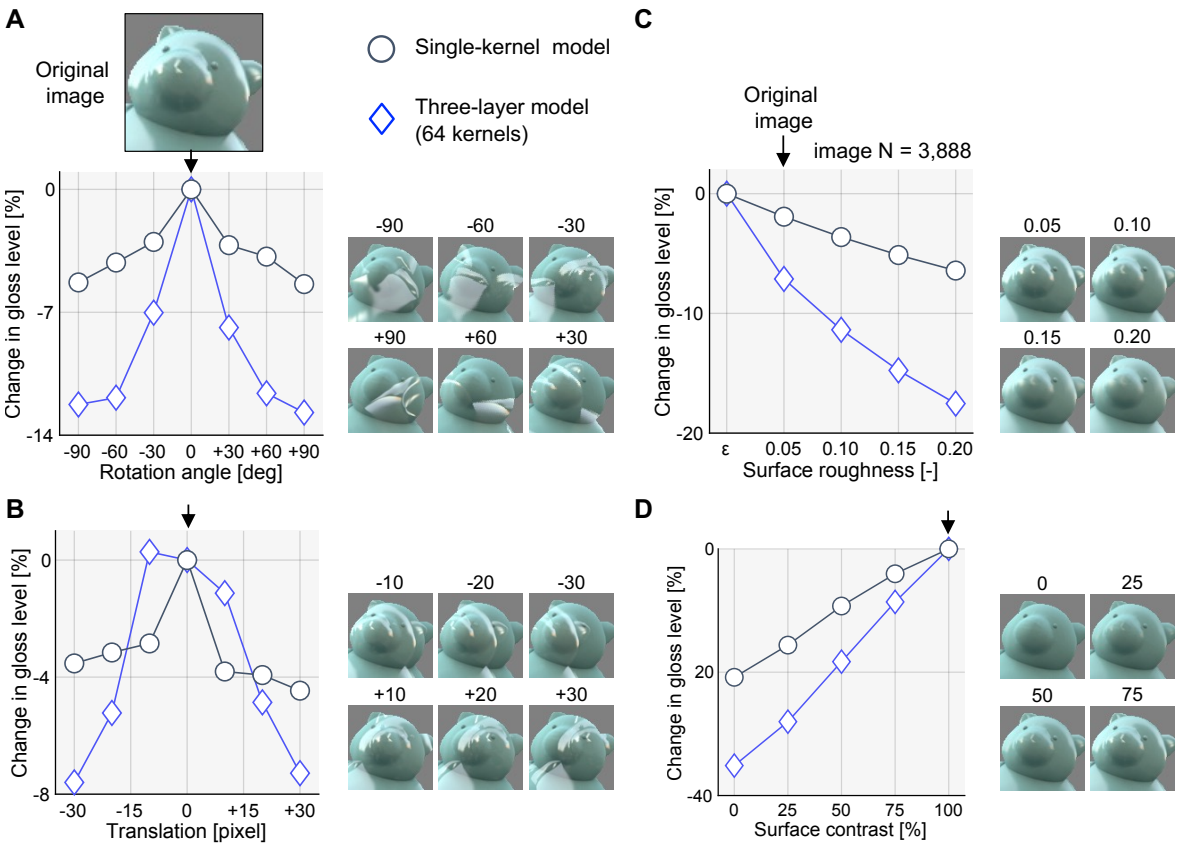

Figure 5. Evaluation of single-kernel and three-layer network models for three known perceptual effects. The dark blue circles and blue diamonds show model responses for the one-layer model and three-layer models, respectively. (A) Effect of rotating specular structure. The top image shows the original image where no manipulation was applied. The rotation 0 indicates the original image. (B) Effect of horizontally translating specular reflection. The translation 0 indicates the original image. (C) Effect of changing surface roughness level. The surface roughness 0.05 indicates the original image. (D) Effect of changing surface contrast level. The surface roughness 100% indicates the original image. Contrast was manipulated by changing Pellacini's c.

Supporting Information

Object geometries and lighting environments.

To generate test images, we selected 36 three-dimensional meshes of everyday objects from Evermotion (https://evermotion.org) and 36 image-based illumination maps from various databases, including the Debevec database [1], Southampton-York Natural Scenes [2], and Freebies (https://hdrmaps.com/freebies/; accessed May 2021), as shown in panels A and B of **Figure S1**. For a reference image, we chose an Uffizi light probe, as in a previous study [3], and generated the geometry of a bumpy sphere using the 3D modeling software Blender (version 2.79b). These lightings and shapes were selected to cover a wide range of characteristics, with different complexities, which we thought is key to identifying unique human behaviors that decorrelate from physical reflectances. Panel C shows the chromatic distribution of each lighting environment. There is a tendency for pixel colors to spread along the vertical blue-yellow axis, known as a daylight locus [4].

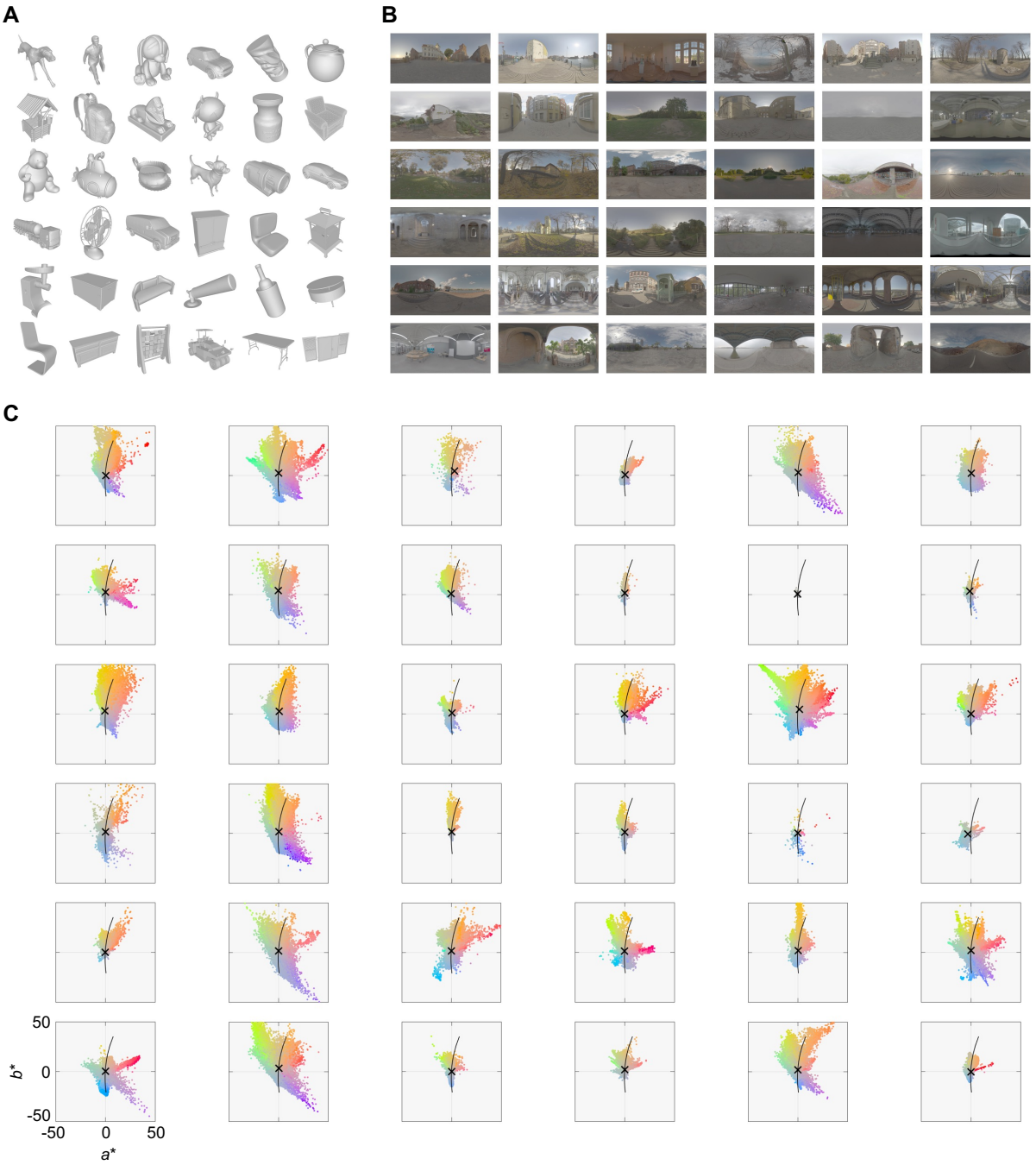

Figure S1. (A, B) Object geometries and lighting environments selected to generate test images. They were selected to be as diverse as possible. (C) Chromatic distribution of each lighting environment in L*a*b* color space. We randomly sampled 10% of pixels to reflect the density of distribution. The black cross symbols show mean chromaticity and black curve shows daylight locus. The white point was set to D65. The layout is the same as panel B.

Observer settings for each lighting environment and shape.

Figures S2 and **S3** show mean observer settings compared to the physical ground-truth, grouped by each lighting environment and each shape, respectively. Each data point represents one image, with each plot containing 108 images (36 lighting environments or 36 shapes × 3 viewpoints). The aim of this analysis is to determine whether a specific lighting environment or shape consistently induces high or low gloss levels. For lighting environments, although correlations between human settings and physical ground-truth vary between 0.40 and 0.68, observer setting patterns were generally similar, with no noteworthy trends. In contrast, there are clearer differences across shapes. Shapes ranked high tend to be round, presumably because they generate highly visible specular reflections, while flat objects ranked low tend to produce low perceived gloss. This is visually intuitive, as specular highlights for the flat objects are widely spread on their surfaces and not easily visible. Overall, these findings show that when focusing on a particular shape or lighting environment, perceived gloss somewhat correlates with the underlying physical specular reflectance. It is the use of diverse shapes and lighting environments that decorrelates human responses from the physical ground-truth labels in our behavioural data.

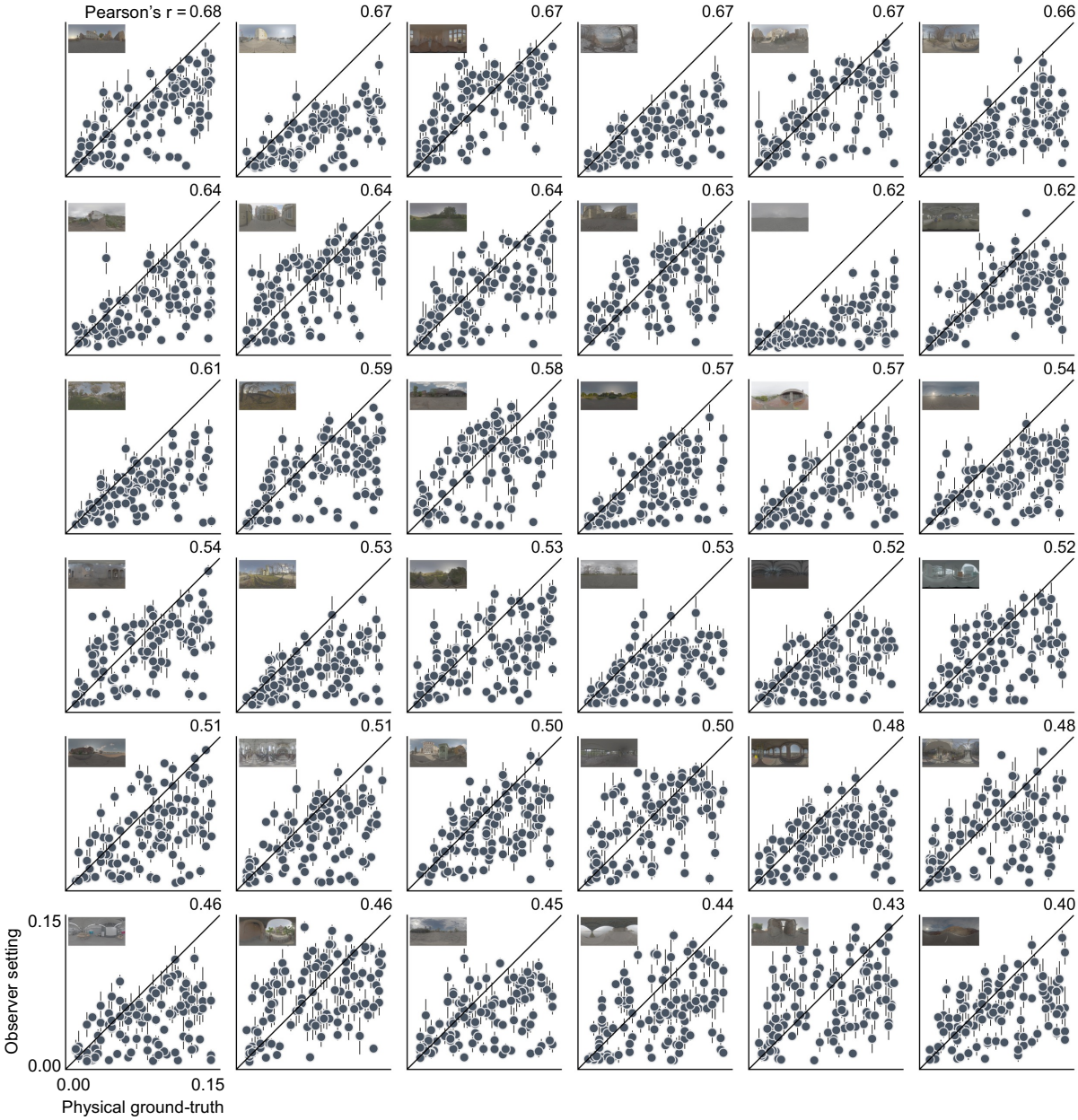

Figure S2. Average observer settings compared to physical ground-truth for images grouped by lighting environment. All images for each specific lighting environment were included in each panel, resulting in 108 data points (36 shapes × 3 viewpoints). The upper right value shows Pearson's correlation coefficient, used to rank each panel. The vertical error bar represents ±S.E. across observers.

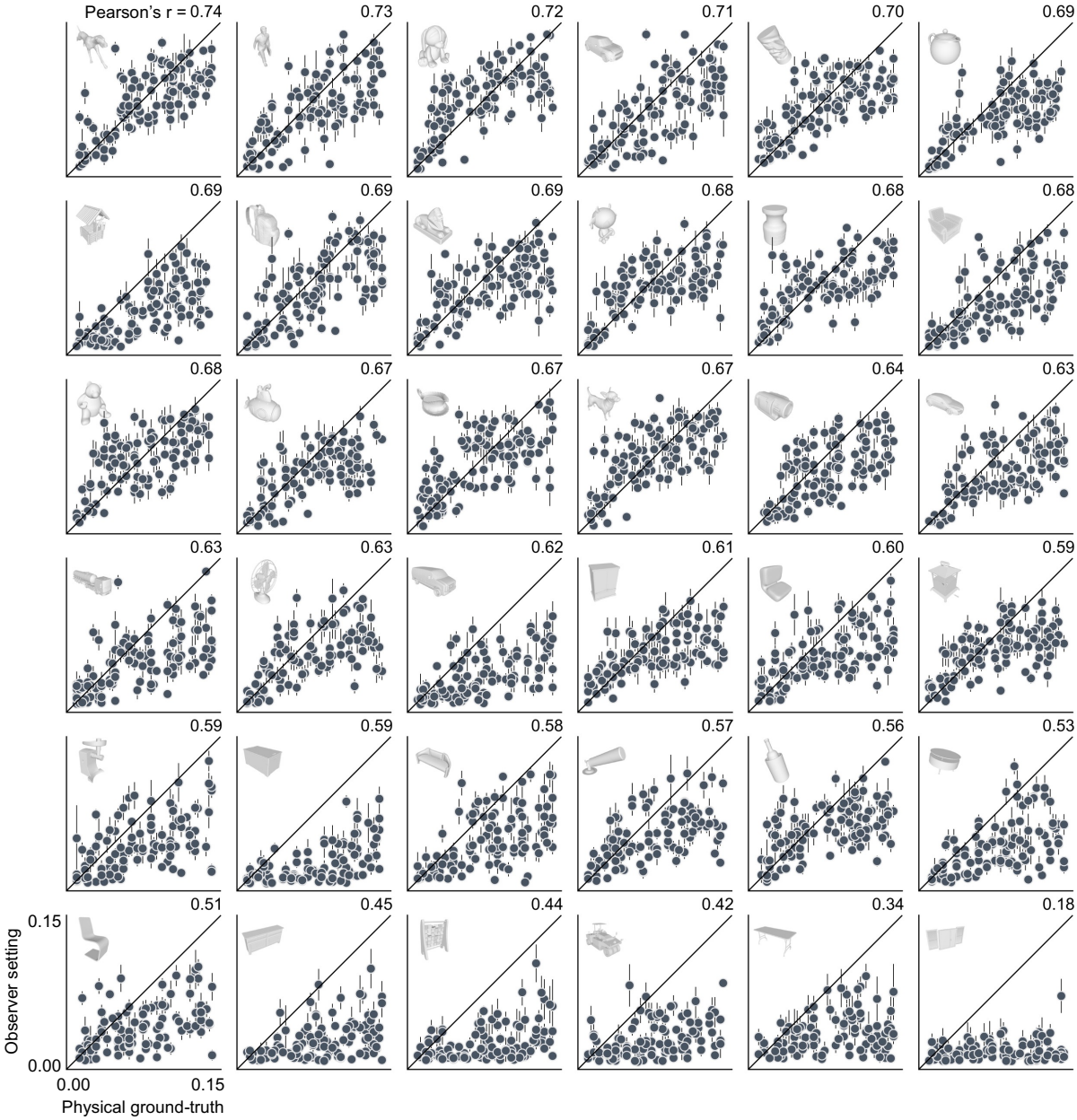

Figure S3. Similar plot to Figure S2, but data points are grouped by object shapes. Each plot contains 108 images (36 lighting environments × 3 viewpoints).

Laboratory experiment

Motivation

Our main experiment was conducted online to effectively collect a large number of perceptual labels. However, there are general concerns with online experiments [5], such as the lack of monitor calibration, absence of an experimenter, and lack of verbal communication, which may affect task comprehension. To assess the quality of the online data, we conducted a validation experiment in a controlled laboratory environment as follows.

Observers and ethics.

Twenty observers (13 females and 7 males) were recruited. Observers' age ranged from 18 to 36 and mean ± standard deviation was 23.4±4.06 years. All observers had normal or corrected-to-normal visual acuity and normal color vision, screened using Ishihara Pseudo-isochromatic color plates [6]. Informed consent was obtained before the experiment. Before the experiment, the experimenter explained the task of the experiment to the observers. All procedures were approved by the local ethics committee at Justus Liebig University Giessen.

Test images.

We used two of the image sets that had been shown in the main experiment to 21 and 23 online observers. Each set contained 72 images, excluding 12 common images. All lab-based observers judged all images from both sets.

Apparatus.

Images were displayed on a 24-inch LCD monitor (ColorEdge CG2420, 1920×1200 pixels, frame rate 60 Hz; EIZO, Ishikawa, Japan) that allows for 10 bits per color channel (red, green, and blue). Gamma correction and spectral calibration were performed using measurements from a spectroradiometer (CS-2000; KONICA MINOLTA, Inc., Tokyo, Japan). The code to run the experiment was prepared in MATLAB using custom functions as well as functions provided in PsychToolbox-3 [7]. A chin rest was used to maintain the viewing distance.

Procedures.

The experiment was conducted in a dark room. Observers placed their forehead and chin on the chinrest, to keep the viewing distance constant at 49 cm from the LCD monitor. Observers viewed the monitor binocularly. In each session observers completed 144 trials (72×2 images), and all observers completed 2 sessions, totaling 288 trials. The images were presented in a random order. Otherwise, the experimental procedure and the task of the observer were identical to the online experiment. Median response time was 6.18 sec per trial across — longer than in the online experiment (4.89 sec per trial).

Results.

Panel A in **Figure S4** compares observer settings between laboratory and online experiments, where each data point represents the average across all observers. In both image sets, we found a near-perfect correlation. These high correlations likely result from averaging data from a large number of observers. We also created a histogram of correlations for all possible pairs of individual observers (Panel B). In this plot, we observed that the median correlation was higher than 0.8, which is not dramatically lower than the within-observer correlation shown in Panel C. In summary, these comparisons demonstrate excellent consistency between online and laboratory-based experiments, suggesting that online data collection may be feasible for certain research questions and visual tasks.

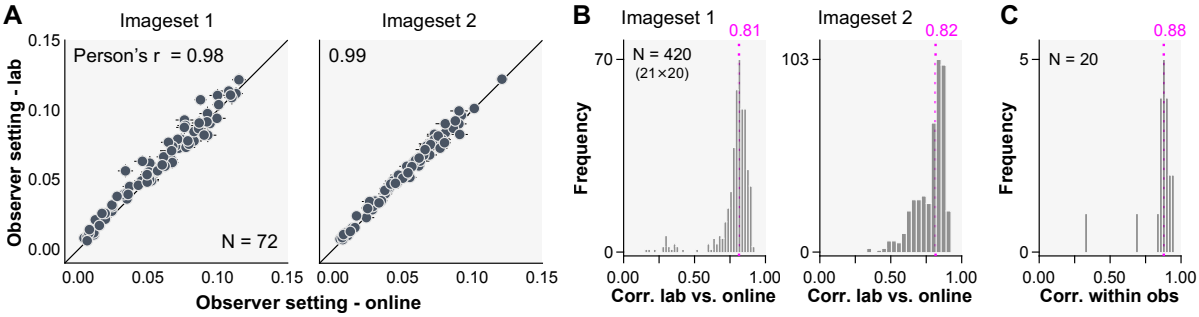

Figure S4. Results of the online experiment. (A) Laboratory experiment data plotted against online experiment data, with each data point representing the average across observers. (B) Histogram of correlations across individual observers between laboratory and online experiments. The median is shown in magenta. (C) Histogram of within-observer correlations computed between sessions 1 and 2 over 144 images.

Analysis of CNN-based models

One-layer model.

Panel A in **Figure S5** shows that the emergent kernels are quite consistent across 24 networks. While the position of the bright blob (and the accompanying brown slant) differs, the relative spatial structure and chromatic properties appear very similar. Panel B displays the chromatic distribution of an example kernel in the L*a*b* color space. We converted the sRGB values of each kernel pixel to L*a*b* coordinates by setting the white point to D65 (X=95.0, Y=100, Z=108.9). We found that the chromatic distribution is tightly clustered around the CIE daylight locus, where natural illuminant colors typically distribute [4]. As seen in Panel C of Figure S1, the chromatic distribution of the environmental illuminations used in this study also exhibits this trend, which governs the color distribution of specular highlights, a direct reflection of incident illumination. It is likely that the filters emerged to reflect this regularity and utilize it to extract specular reflections from an object surface.

The upper subpanel of Panel C shows a 3-D visualization of the kernel. The middle subpanel shows a fitting to the difference of Gaussians (DoG) model, with 13 free parameters: amplitude, x-center coordinate, y-center coordinate, rotation angle, standard deviations along the x-axis and y-axis for both the first and second Gaussian, and an offset value. We found that the DoG filter provided a slightly better fit (R^2=0.79, sum of residuals = 0.0481) than a single Gaussian filter (R^2=0.76, sum of residuals = 0.055). Residuals appear to have a systematic periodic pattern, so we fitted second and higher-order derivative Gaussian filters, but the fitting revealed no easy way to decompose the residuals into finer filters. Furthermore, independent component analysis using the residuals from 24 filters extracted no systematic patterns. Nevertheless, these filters might possess additional functions beyond the DoG filter, even though the value of the residuals is much smaller than the DoG component.

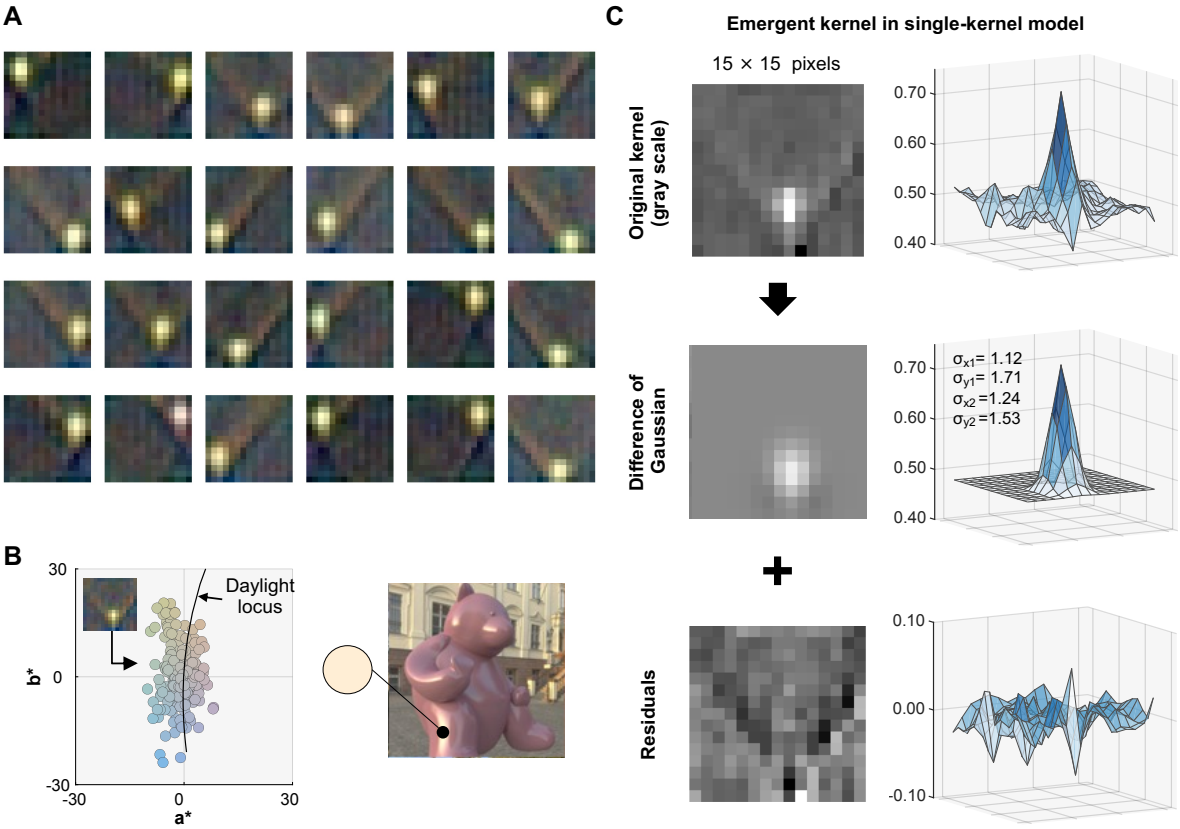

Figure S5. Kernel analysis of single-kernel networks. **(A)** A consistency check across 24 networks. Although they were trained on different (but overlapping) datasets with different random weight initializations, all networks developed a similar pattern of kernels, with a bright blob located at various positions, superimposed on straight elongated, diagonally-oriented yellow-black-blue features. **(B)** Analysis of chromatic tuning of the kernel. The sRGB color of each pixel in an example kernel was converted to the L*a*b* color coordinates, and their a*b* distribution is shown. We found that kernel colors are closely aligned along the daylight locus, likely reflecting the typical color of specular highlights seen in training images (see example image on the right). This occurs because specular reflection directly mirrors the illumination, reflecting the chromatic properties of the lighting environment shown in **Figure S1C**. **(C)** Fitting a two-dimensional difference of Gaussian (DoG) filter to the kernel's spatial intensity distribution. For simplicity, the fitting was performed using the luminance image. The bright blob region is well approximated by the DoG filter, with the residual component shown at the bottom.

Three-layer model.

We examined a three-layer network with 64 kernels, the best of all the models we tested, to see how glossiness is coded within the network and how the representation progressively changes over layers. We first extracted activation maps after each convolutional layer and aggregated the signals over the map using max pooling for the first and second layers, and average pooling for the third layer following the original design of ResNet. This resulted in 64 scalar values per layer, which were plotted on a 2-dimensional plane using t-distributed stochastic neighbor embedding (t-SNE; **Figure S6**). For comparison, we performed the same analysis using the raw pixel values of the input image.

Results revealed that in pixel space, objects are primarily clustered by their geometry. In the first layer, object images are already roughly grouped into glossy and matte categories, with some glossy objects misclassified as matte and vice versa. The second and third layers only slightly refine this representation. However, the overall representation does not change drastically from the first layer. This suggests that the first convolutional layer most strongly contributes to gloss judgments, with subsequent layers making small adjustments to improve predictions, perhaps by separating overall luminance level (e.g., albedo) and glossiness.

One difficulty in testing this hypothesis further with our dataset is that mean luminance and perceived glossiness correlate (correlation coefficient of around 0.45), so we expect trained networks did not fully learn to separate these components. Nevertheless, it is intriguing that first-layer kernels, such as those approximated by DoG filters, play a vital role in gloss perception.

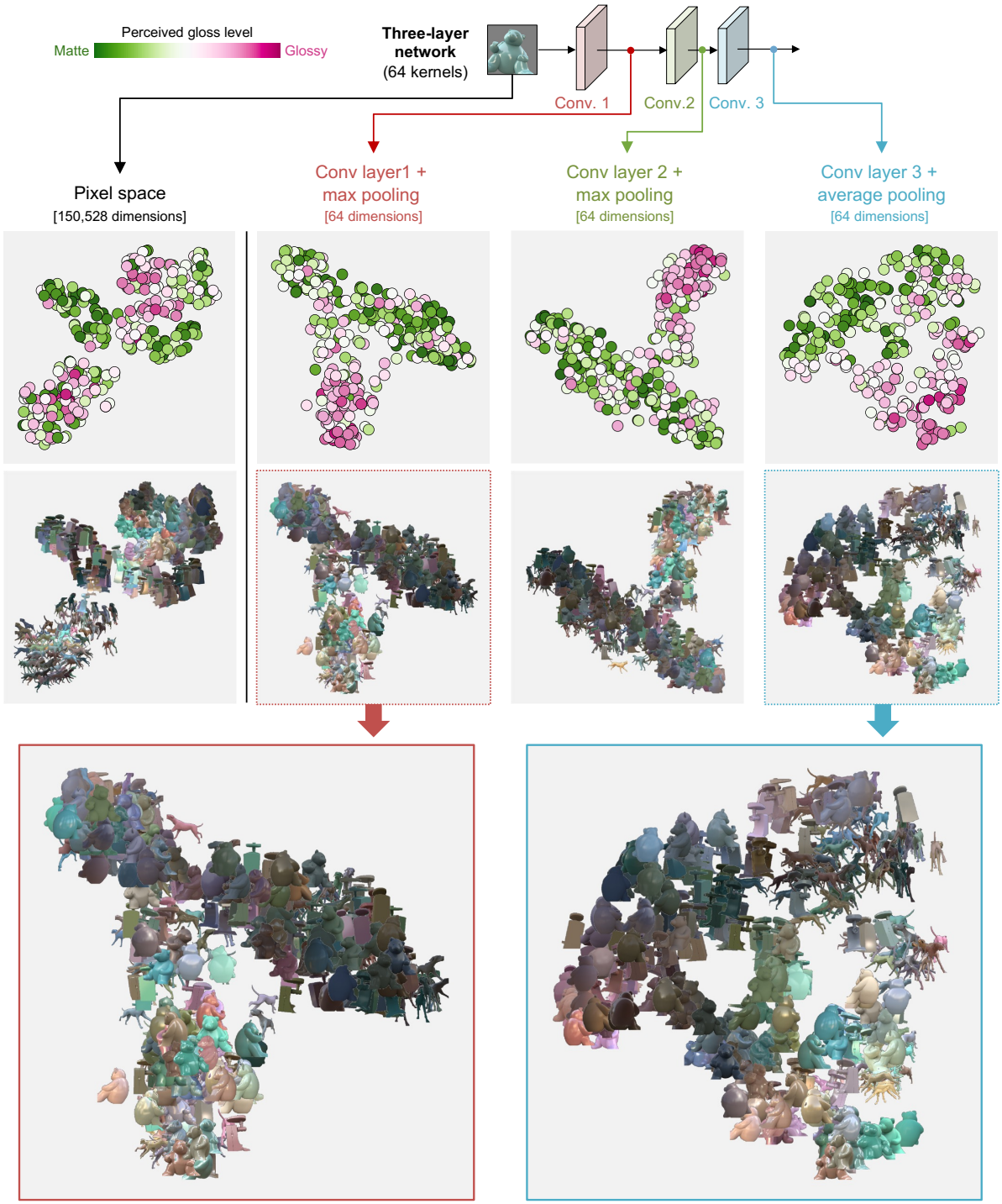

Figure S6. Example of internal representation in a three-layer network. We fed the network images it had not seen during training and extracted activation maps from each convolutional layer. The maps were aggregated by max pooling (layers 1 and 2) or average pooling (final layer) over the entire map, then used as inputs for t-distributed stochastic neighbor embedding (t-SNE) to project onto a 2-dimensional plane. For comparison, we also performed t-SNE analysis on the same images in the pixel space (leftmost plot), where objects are primarily clustered by shape. The bottom two figures show enlarged representations from the first and third convolutional layers.

CNN-based models that takes the surrounding context information into account

We removed the surrounding context from each image, set background pixels to mid-gray (R=G=B=128), and used the resulting images to train the CNN-based models presented in the main text. This decision was made because our shallow network model lacks the computational capacity to effectively segregate the background from the object while also predicting gloss levels. However, the surrounding context might contain useful information that could help CNNs better predict human responses. To explore this possibility, we added an additional stream to our existing models, incorporating information from the surrounding region into the regression layer at the end to determine gloss levels.

Figure S7 illustrates the architectures of such models and their performance in predicting human responses for the one-layer model (panel A) and the three-layer model (panel B). The results indicate that including context information in the network stream results in no performance improvement for both models (two-tailed paired t-test; $t(23)=0.572$, $p = 0.573$ for one-layer model; $t(23)=0.55$, $p = 0.588$ for three-layer model), suggesting that the surrounding context does not provide significant additional information. It is possible that useful cues are present but not captured by the networks. Nevertheless, this observation aligns with past empirical evidence that observers do not rely on surrounding contextual information when judging surface glossiness [8].

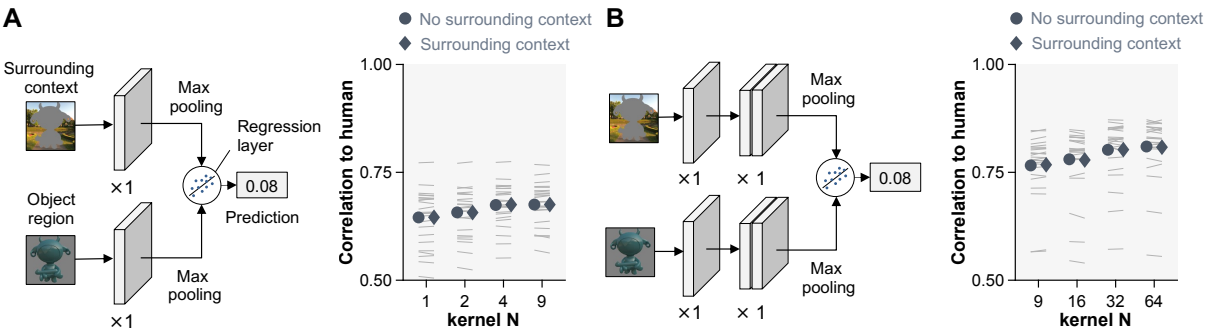

Figure S7. Testing whether surrounding context provides useful information for predicting human responses. (A) The left network illustrates the architecture of a one-layer model with an additional stream for processing surrounding context. The circles indicate the correlation coefficient with human responses, averaged across 24 networks when no surrounding context was used, as reported in the main text (Figure 3). The diamonds show the results when surrounding context was included. (B) The same analysis for a three-layer network. Lines indicate performance changes for each of 24 networks. In both cases, including surrounding context information resulted in little performance improvement.

SI References

1. Debevec P (1998) Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. SIGGRAPH 98 Proc 25th Annu Conf Comput Graph Interact Tech 189–198.
2. Adams WJ, et al. (2016) The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. Sci Rep 6:35805.
3. Morimoto T, Akbarinia A, Storrs K, Cheeseman JR, Smithson HE, Gegenfurtner KR, Fleming RW (2023) Color and gloss constancy under diverse lighting environments. J Vis 23(7):8, 1–25.
4. Judd DB, et al. (1964) Spectral distribution of typical daylight as a function of correlated color temperature. J Opt Soc Am 54:1031–1040.
5. Rodd JM (2024) Moving experimental psychology online: How to obtain high quality data when we can't see our participants. J Mem Lang 134:Article 104472.
6. Ishihara S (1973) The series of plates designed as a test for colour-blindness (Kanehara Shuppan Co. Ltd., Tokyo, Japan).
7. Brainard DH. (1997) The Psychophysics Toolbox. Spat Vis. 10(4):433-6.
8. Motoyoshi I, Matoba H (2012) Variability in constancy of the perceived surface reflectance across different illumination statistics. Vis Res 53:30–39.