nature human behaviour

Article

https://doi.org/10.1038/s41562-026-02445-0

Human gloss perception reproduced by tiny neural networks

Received: 30 June 2025

Accepted: 13 March 2026

Published online: 12 May 2026

Check for updates

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

A key goal of visual neuroscience is to explain how our brains infer object properties such as colour, curvature or gloss. Here we used machine learning to identify computations underlying human gloss judgements—traditionally considered a challenging inference. We rendered thousands of objects with varied shapes using a Ward reflectance model across lighting and viewpoints, then obtained gloss ratings for each image. Observers' judgements were consistent with one another, yet systematically deviated from reality. We compared these ratings with neural networks trained either to estimate physical reflectance ('ground-truth networks') or to reproduce human judgements ('human-like networks'). While estimating physical reflectance required deep networks, shallow networks accurately replicated human judgements. Remarkably, even a single-filter network could predict human judgements better than the best ground-truth network and generalized to known gloss illusions. These results suggest that gloss perception relies on simple general-purpose computations, and demonstrate the power of interpretable 'tiny' networks in understanding cognition.

In daily life, we encounter materials with a wide range of qualities, such as metal, cloth or plastic[1]. In a brief glance, we can judge their colour, lightness, shape, glossiness, translucency and other appearance characteristics. We can infer associated physical properties, such as weight, softness or roughness, and even their states, such as whether they are wet or dry, clean or dirty, solid or melting[2]. Humans use visually sampled information about materials to guide our behaviours, such as lifting heavy objects slowly or plucking ripe berries gently.

Among the various material properties that we can perceive visually, gloss is the most extensively studied[3]. The term 'gloss' refers to the subjective impression of shine or lustre, arising from mirror-like surface reflections (see Supplementary Fig. 1 for details). Gloss estimation is widely considered a quintessential, challenging perceptual inference[3–5] as it involves not only distinguishing highlights and reflections from other features (for example, bumps or surface markings) but also pooling and interpreting the reflections to arrive at a global estimate of the surface's reflectance. 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[6–9], the luminance gradient[10], the standard deviation (s.d.) of luminance over a surface[11], and various image metrics derived from specular reflection patterns[12,13]. 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. To characterize how the visual system addresses this challenge, many studies have measured gloss perception across a wide variety of viewing conditions[14–30]. Yet, no single model fully captures human gloss perception across these diverse factors.

In the present study, we aim to identify biologically plausible, image-computable models that can account for human gloss judgements across a broad set of images. Our focus is specifically on understanding the perceptual computations that support judgements of gloss from images. While this is not equivalent to gloss perception in real-world, dynamic viewing, image-based judgements provide a tractable and meaningful way to investigate the visual information and representations that correlate with perceived gloss.

[1]Department of Psychology, Justus-Liebig-Universität Gießen, Giessen, Germany. [2]Department of Experimental Psychology, University of Oxford, Oxford, UK. [3]School of Psychology, University of Auckland, Auckland, New Zealand. ✉e-mail: takuma.morimoto@psy.ox.ac.uk

'Tiny' neural networks as interpretable data-driven models

One persistent challenge for 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 judgements[31,32]. 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 judgements 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 judgements, 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 judgements, aiming to make them as shallow as possible to increase interpretability[33–35]. We then interrogated the mechanisms that emerged within the networks. This enables a data-driven, hypothesis-free approach to uncovering the computations underlying human perception across diverse viewing conditions[36].

Figure 1 summarizes 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 assigned a perceived gloss value based on human gloss judgements ('human labels'), 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 'Laboratory experiment' in the Supplementary Information). 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 grey background (sRGB = [127, 127, 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 judgements.

Our findings suggest that relatively shallow CNN architectures with as few as three convolutional layers are sufficient to predict human gloss judgements approximately as well as individual human judgements predict one another, while inferring the physical ground truth requires far more complex computations for equivalent accuracy. Indeed, we find that even a single convolutional kernel can predict human perception better than the best of the ground-truth networks. Analysis of the emergent filters in the human-like networks showed that a filter resembling a combination of a Gaussian blob and diagonal ridge-like features can effectively extract image features predictive of human gloss judgements. 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 grey circle in Fig. 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 (Pearson's $r(3{,}888) = 0.52$, $P < 0.001$). The histogram of Pearson's correlation coefficients for each of the 295 observers shows a median correlation value of 0.46 (interquartile range (IQR) 0.40–0.54), with no observer exceeding 0.75 (Fig. 2b). This indicates that gloss judgements deviated from ground truth in informative ways, probably due to the broader range of viewing conditions than in many earlier studies: by testing a wider range of conditions, we were able to identify more cases where humans 'misperceive' the specular reflectance.

However, crucially, these gloss judgements were highly consistent both within and across observers (Fig. 2c,d). Intra-observer correlations were calculated between sessions 1 and 2 for each observer over 84 images (Fig. 2c). Each inter-observer correlation was computed over 84 images between one observer and the average of the rest who judged the same set (Fig. 2d, left plot), or between one observer and another for all possible observer pairs ($295 \times 294 \times 0.5 = 43{,}365$ 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 judgements. On the contrary, observers made highly systematic and consistent patterns of agreement and disagreement with physical reflectance in gloss estimation (median inter-observer correlation 0.86, IQR 0.82–0.89 and 0.82, IQR 0.74–0.88; Fig. 2d). A separate, laboratory-based experiment validated that the online data are of comparable quality (Supplementary Fig. 6).

Each grid in Fig. 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 judgements agree. Images in the top left and bottom right grids (low–high, high–low) highlight interesting cases where human judgements and physical ground truth disagree substantially. Note that objects with low specular reflectance still have non-zero reflectance and can therefore produce visible highlights. Such images effectively provide different objectives for networks trained on physical ground truth versus human judgements. Although human judgements 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 judgements (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, we are seeking models positioned close to this distribution of green circles in this plot.

Grey diamonds show models based on statistics of luminance distributions within the image (mean, median, first quartile, third quartile, minimum, maximum, s.d., skewness and kurtosis), which have previously been implicated in gloss perception[37]. However, here, we find that they 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 significantly (grey star symbol) relative to the mean-luminance model, which showed the highest correlation to human observers among the single-statistic luminance models (two-tailed paired t-test; $t(23) = 6.18$, $P < 0.001$, Cohen's $d = 1.261$, 95% confidence interval (CI) 0.713 to 1.793). Light-green downward triangles show models that use sharpness, coverage and sub-band contrast computed from specular reflection images (a set of cues that summarize spatial variations in luminance, and that have been proposed to play a key role in gloss perception)[12]. These models correlate more with physical ground truth than with human observers, indicating that they do not exhibit the characteristic errors seen in human judgements. Combining these cues yielded no improvement (the light-green star symbol) compared

https://doi.org/10.1038/s41562-026-02445-0

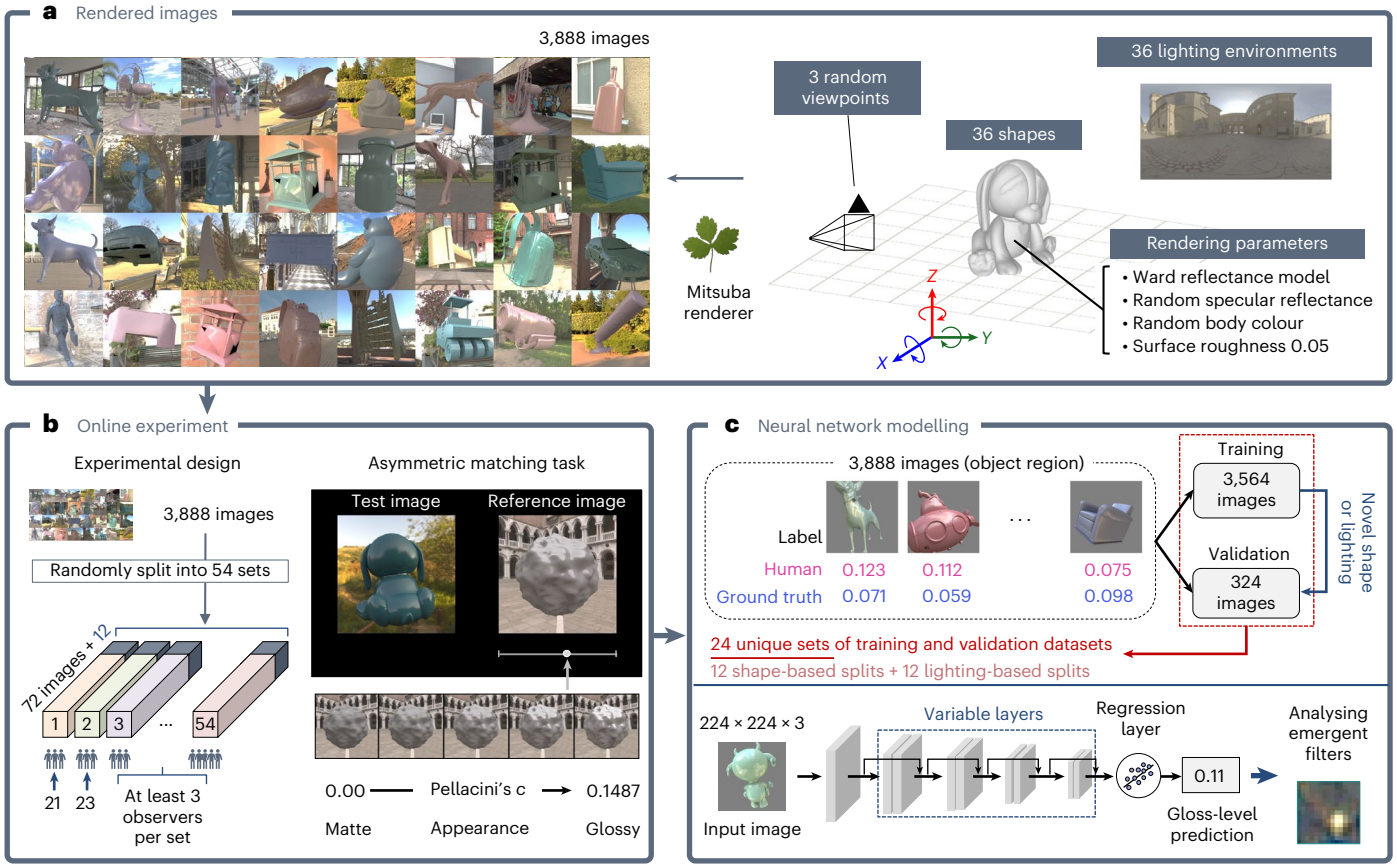

Fig. 1 | Schematic of our approach. a, The scenes contained an object with a glossy surface (Ward reflectance model[69,70], surface roughness fixed at 0.05). Random specular reflectance values and body colours were applied. The full set of 3,888 images was rendered by combining 36 shapes, 36 lighting environments and 3 random viewpoints per shape–lighting pair (36 × 36 × 3). Images were generated in the CIE *XYZ* (1931) colour space, converted to linear sRGB and then displayed using standard sRGB gamma correction ($\gamma = 2.4$). **b**, Left: design of online experiment. The 3,888 images were randomly divided into 54 sets of 72 images each, plus 12 additional images that were shared across sets and rated by all online observers. These shared images, taken from our previous study (see fig. 5a in ref. 30), were used solely to assess inter-observer variability and were not included in the main set of 3,888 images or used for training the network models. Two image sets were also tested in a lab-based experiment (see 'Laboratory experiment' in the Supplementary Information), 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 (that is, Pellacini's *c* (ref. 73)) of the reference object to match the perceived gloss level of the two objects. **c**, For cross-validation, the 3,888 images were split such that each validation set consisted of a block of either three novel shapes (3 shapes × 36 lighting environments × 3 viewpoints = 324 images) or three novel lighting environments (36 shapes × 3 lighting environments × 3 viewpoints = 324 images), which were excluded from the corresponding training set. There were 12 such shape-based splits and 12 lighting-based splits, resulting in 24 unique, non-overlapping combinations of training (3,564 images) and validation (324 images) datasets. We trained CNNs with varying numbers of intermediate layers using images labelled by human gloss judgements ('human-like networks') or by physical ground-truth labels ('ground-truth networks'), where each 'label' consisted of a continuous value of perceived gloss or physical specular reflectance as captured by Pellacini's *c*. Gloss levels predicted by the networks were compared against human responses and physical ground-truth labels. The trained networks were analysed to understand the computational mechanisms that emerged within them.

with the sub-band contrast model (two-tailed paired *t*-test; $t(23) = 0.07$, $P = 0.947$, Cohen's $d = 0.014$, 95% CI −0.386 to 0.414).

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 judgements across all observers (black cross). Yet, it is extremely intriguing and noteworthy that CNNs with only one kernel (single-kernel model) come close to the edge of the human distribution, with a mean correlation coefficient of 0.65 (s.d. 0.064) across 24 independently trained models from 24-fold cross-validation. Given the correlation coefficient across observers was around 0.85, these very light models reach 75.3% of this performance ceiling and indeed approximate human perception better than any of the models trained on physical ground-truth labels that we tested.

To further contrast the human-trained models with an inverse-optics framework, we also trained networks on physical ground-truth labels and additional images, assuming that with sufficient training a supervised neural network can approximate a near-optimal observer at inferring specular reflectance from the images in our training set. Note that 'near-optimal observer' does not imply perfect performance, as some stimuli can be fundamentally ambiguous in that they contain insufficient information to infer reflectance accurately. Comparing human-like and physically trained networks in terms of their architectural complexity and training data requirements helps reveal how the underlying computations differ between recovering physical surface properties and predicting human perceptual judgements. Put another way, it allows us to test the extent to which human behaviour resembles a 'near-optimal' observer. Interestingly, we find that it does not.

In contrast to the CNNs trained on human labels, the networks trained on physical ground-truth labels, shown by magenta symbols, struggle to achieve high performance in inferring physical reflectance. Although trained to recover physical labels, some physical ground-truth networks are positioned near human-like networks, and

https://doi.org/10.1038/s41562-026-02445-0

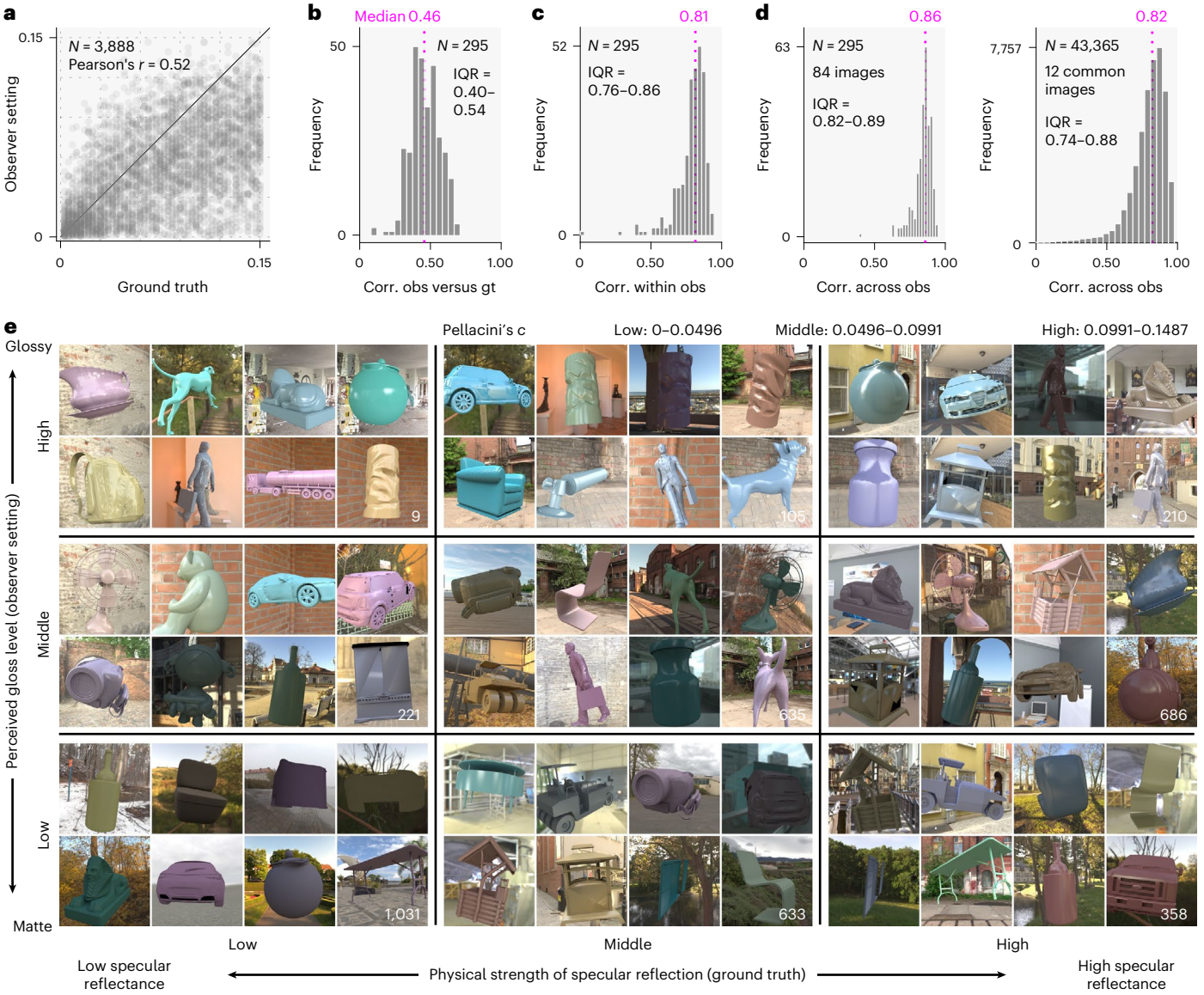

Fig. 2 | Summary of behavioural results. a, Each data point shows a specific stimulus image, and the plot includes data on all 3,888 images. The horizontal and vertical axes show physical ground-truth reflectance values and observer settings (perceived gloss) averaged over observers. **b**, Histogram of human versus ground-truth correlation. Corr., correlation; obs, observer; gt, ground truth. 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. For the right plot, the correlation was computed over 12 images that were shared across all 54 image sets for every pair of observers (that is, $(295 \times 294)/2 = 43{,}365$ pairs). **e**, Each grid shows eight 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.

even the three-layer models cluster around the diagonal unity line in the plot. A more complex architecture, such as ResNet18 (magenta rightward triangle), is also located near the unity line. 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 squares) still show a limited correlation below 0.70, even with an additional 3.8×10^5 images. A ResNet18 trained with these additional images achieves a correlation of around 0.9 with the physical ground truth, highlighting the computational challenges involved in estimating physical ground-truth. Importantly, however, its correlation with human responses remains below 0.5. The intriguing aspect of this observation is that it suggests 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 inverse optics approaches. Instead, it appears that humans use relatively simple computations to intuit glossiness. Having established that lean neural network models can be trained to approximate human responses, we next sought to analyse the inner workings of these networks as a means to gain insights into the computations underlying human gloss judgements.

Analysis of single-kernel models

As shown in Fig. 4a, the single-kernel model first applies a $15 \times 15 \times 3$ convolution to the input image, followed by max pooling and the addition of a bias term to predict a gloss level. To evaluate the variability of the emergent kernel, we analysed 24 networks trained as part of the cross-validation process.

https://doi.org/10.1038/s41562-026-02445-0

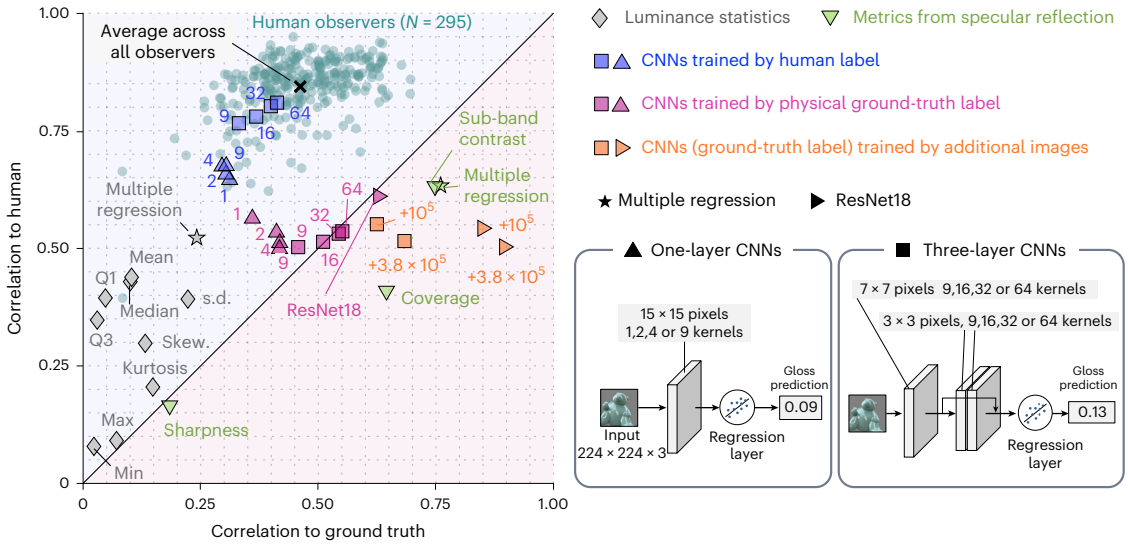

Fig. 3 | Summary of computational model behaviours. Dark-green circles show individual human observers ($N = 295$), where the x axis and the y axis show the correlation over 74 images to the physical ground truth and the correlation to the average of the rest of the observers, respectively. Other data points show candidate computational models. Grey diamonds show low-level luminance image statistics models, and the grey 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 labelled by observer judgements and pink symbols show networks trained on images labelled by physical ground-truth labels. The magenta rightward triangle shows ResNet18 trained on physical ground-truth labels. 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.

We found that the emergent kernels were highly consistent across the 24 networks (Fig. 4b). Their typical spatial structure consisted of a bright central blob surrounded by a darker region, which is well suited for detecting highlights based on their characteristic spatial intensity profiles. Notably, the bright region often exhibits a yellowish tint, while the surrounding area appears bluish. This mirrors the typical colour relationship in natural scenes between direct illumination (such as sunlight) and indirect illumination (such as skylight), allowing the kernels to exploit hue differences for highlight detection. It is also notable that each blob is abutted by one or two brownish diagonal ridges—typically consisting of a brown upper region, a dark middle region and blueish lower region—presumably positioned to detect oriented, lower-contrast highlights. These ridges are bands of positive weights running at approximately 45° or 135°, which enable the kernel to respond strongly to local intensity changes at similar angles. As a result, the kernels can detect diagonal boundaries and curvatures typical of specular highlights on curved surfaces, as well as capture how highlights spread or extend in specific directions. Interestingly, while the 'human-like' single-kernel networks (that is, those trained on human labels) exhibited these distinctive diagonal ridges, the 'ground truth' networks did not (see Supplementary Fig. 7a for example kernels). This suggests that the oriented ridges in the kernels truly capture something specific to human gloss perception—rather than being generically useful features for estimating reflectance. This may reflect the fact that the human visual system develops with a more varied visual diet and has to support many more visual tasks than simply estimating specular reflectance. Presumably, when asked to identify gloss, humans rely especially on features that distinguish specularity from other sources of image contrast.

Figure 4c shows the chromatic distribution of an example kernel from a human-like model in $L^*a^*b^*$ colour space. To obtain these coordinates, we converted the sRGB values of each kernel pixel to $L^*a^*b^*$ using a D65 white point ($X = 95.0$, $Y = 100$, $Z = 108.9$). The resulting distribution is tightly clustered along the International Commission on Illumination (CIE) daylight locus, where natural illuminant colours typically fall[38]. As illustrated in Supplementary Fig. 3b, the environmental illuminations used in this study vary in gamut area but generally exhibit an elongated colour distribution along the daylight locus, which directly influences the colour of specular highlights as reflections of incident light. This suggests that the filters exploit this regularity to help extract specular reflections from object surfaces. To test this, we generated 3,888 new images with lighting environments in which pixel colours were rotated by +90° in the a^*b^* plane, while L^* (lightness) was kept constant. All other rendering parameters matched the main dataset, and we assumed human gloss judgements would remain the same. As shown in Fig. 4d, the chromatic distribution of the resulting kernel in the human-like model also rotated by roughly 90°, confirming that the kernel's chromatic properties are shaped by the chromatic illumination statistics of the training environment.

To test whether certain geometric factors in the rendering process gave rise to the 45° and 135° ridge detectors, we manipulated object geometry and illumination to isolate their underlying contributions. When we rotated all objects by +90° relative to their original orientation (Fig. 4e), the angle of the ridge shifted to 155°, corresponding to a 20° change. By contrast, when we lowered the illumination elevation by 90°, moving the primary light source from the upper hemisphere towards lower regions, the kernel pattern exhibited little change (Fig. 4f). Therefore, the learned kernel ridge does not simply align with absolute lighting direction and overall is more affected by the projected surface geometry in the image set. See Supplementary Fig. 7b,c for extended analyses.

As shown in Fig. 4g, we found that the spatial properties of the bright blob and accompanying ridges can be represented by a two-dimensional (2D) Gaussian function, a type of receptive field observed as early as retinal ganglion cells[39] and one-dimensional (1D) oriented Gaussian ridges, respectively. The 2D Gaussian has five free parameters: amplitude (a_g), x-centre coordinate (x_{g0}), y-centre coordinate (y_{g0}), s.d. along the x axis (σ_{gx}) and y axis (σ_{gy}). Each 1D Gaussian ridge has five

https://doi.org/10.1038/s41562-026-02445-0

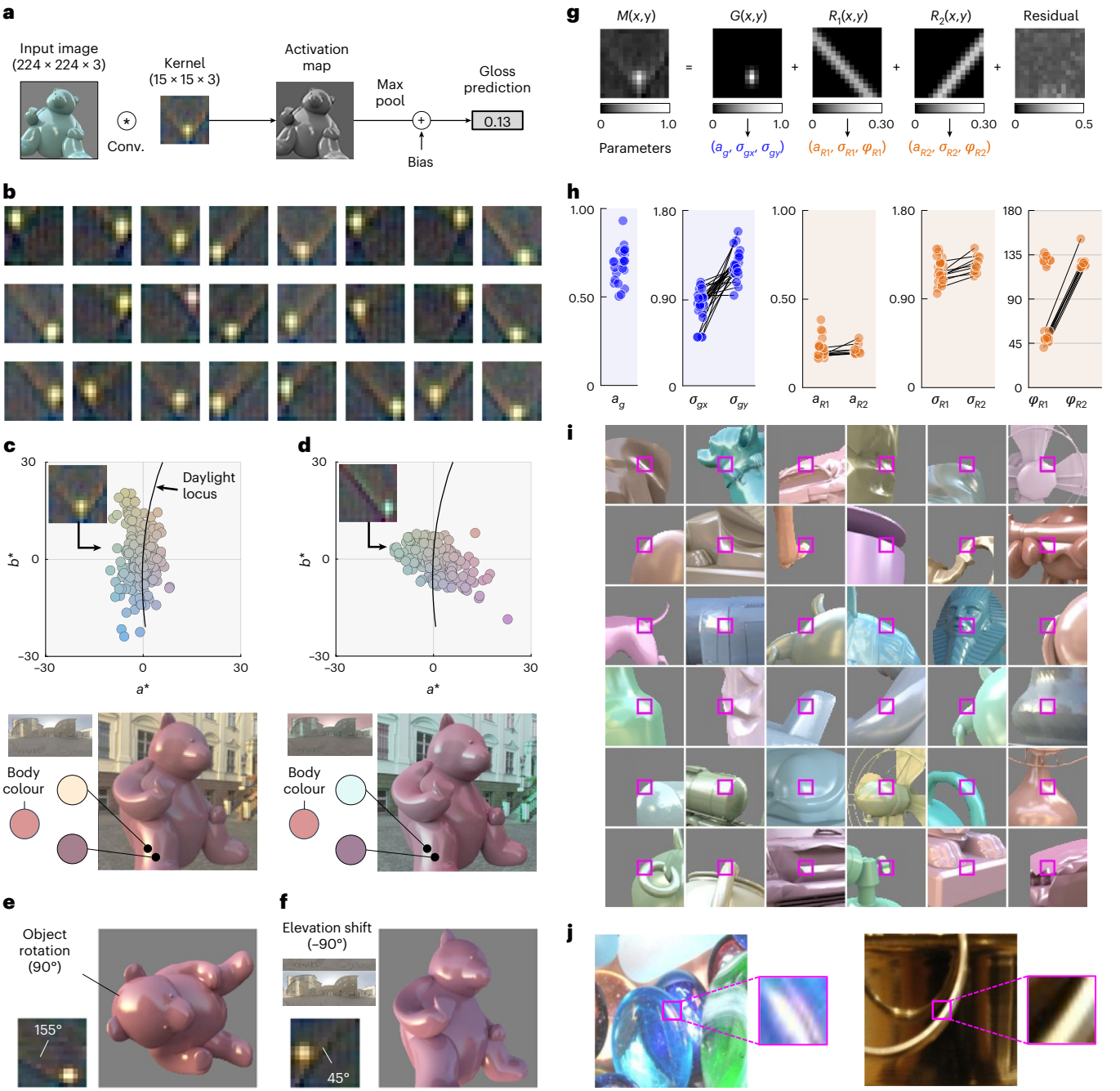

Fig. 4 | Analysis of single-kernel models trained on human gloss judgements. **a**, Computational steps of the single-kernel model. Conv., convolution. The input image is convolved with a 15 × 15 × 3 kernel, the resulting activation map is max-pooled to a single value, and a bias term is applied to produce the gloss prediction. **b**, Consistency check across 24 networks trained in cross-validation. Although the training splits partially overlapped across folds (and, thus, the networks are not fully independent in terms of the data they saw), the networks were trained separately 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 brown–black–blue ridges. **c**, Analysis of chromatic tuning of the kernel. The sRGB colour of each pixel in an example kernel was converted to the $L^*a^*b^*$ colour coordinates, and their a^*b^* distribution is shown. We found that kernel colours are closely aligned along the daylight locus, probably reflecting the typical colour of specular highlights and their contrast with the surrounding bluish ambient light (from skylight reflections in shadows) seen in training images (see example image at the bottom). **d**, Chromatic distribution of an emergent kernel from a human-like network trained on a new set of 3,888 images under 90° gamut-rotated environmental illuminations. The kernel clearly adapts to the chromatic statistics of specular highlights present in the training data. **e**, Kernel that emerged from a human-like network trained on images in which the object was rotated by 90° relative to the original orientation. **f**, Kernel that emerged from a human-like network trained on object images illuminated by light probes whose elevation was lowered by 90°. **g**, Fitting a 2D Gaussian filter and one or two Gaussian ridges to the kernel's spatial intensity distribution. For simplicity, fitting was performed on the luminance image after mean subtraction. The bright blob region is well approximated by the 2D Gaussian filter. The diagonal stripe pattern is captured by a 1D Gaussian ridge, and the residual component is also shown. The difference in scale between the two components is indicated by the colour bars. **h**, Distribution of the fitted parameters for the 2D Gaussian components and ridges across the 24 kernels shown in **b**. **i**, Thirty-six image regions that most strongly activated the example kernel for images where both humans and the model judged the surfaces as highly glossy. Although the regions capture a variety of specular reflection geometries, diagonal highlights are comparatively more prevalent. **j**, Examples of real material photographs, showing enlarged regions of oriented specular highlights that maximally activate our single kernel.

free parameters: amplitude (a_{Rk}), x-centre coordinate (x_{Rk0}), y-centre coordinate (y_{Rk0}), s.d. (σ_{Rk}) and orientation (φ_{Rk}), where k represents ridge 1 or 2. Although having both x-centre and y-centre is redundant (because the ridge functions are infinitely long), they were retained for consistency with the 2D Gaussian blob formulation. The correlation between the emergent CNN kernel and the fitted Gaussian blob and ridge model was 0.89 on average across 24 filters (s.d. 0.017), implying that remaining local features in the residuals (Fig. 4g, rightmost image) contribute minimally to the overall structure.

Figure 4h shows the fitted parameters for each function across the 24 kernels displayed in Fig. 4b, excluding x_{Rk0} and y_{Rk0}. The 2D Gaussian blobs exhibited high amplitude values (0.5–1.0) and s.d. indicating a slight vertical elongation. Among the 24 kernels, 9 were fitted with two ridges. The amplitudes (a_{Rk}) of these ridges were consistently lower than those of the Gaussian blobs, typically ranging from 0.2 to 0.3, and σ_{Rk} values were similar between ridges. The orientation parameter (φ_{Rk}) clustered clearly around 45° and 135°. When two ridges were present, they were separated by 90°, indicating a systematic orientation preference in the filter.

We examined the local image features that most strongly activate the 'blob-and-ridge' kernel (Fig. 4i). Each image shown was judged as highly glossy by both the model and human observers, and the magenta square marks the local patch that maximally activates the kernel. These examples reveal that strong activations typically occur for regions containing an elongated, collinear highlight 'ridge,' most often oriented diagonally (~45° or 135°). Unsurprisingly, the kernel also responds to more circular highlight structures ('blobs') (for example, fifth row, third column). Another notable observation is that the filter is sensitive to cases where the highlight terminates abruptly at an object boundary or sharp surface edges. Such configurations are rarely produced by diffuse (Lambertian) shading in natural lighting environments, which instead generates smoother intensity gradients. At the same time, these examples also illustrate that the kernel can respond to a variety of specular highlight patterns.

We emphasize that the strong orientation selectivity of the kernel is unlikely to be merely an artefact of our training image set. Oriented specular highlights are frequently observed in everyday materials that contain roughly cylindrical structures, in which the highlight is necessarily elongated, as demonstrated by examples taken from the Flickr Material Database[40] (Fig. 4j). This pattern is thus characteristic of real-world surfaces, not just synthetic images. Further validation of the single-kernel model's performance on real-world photographs is presented in a later section.

In summary, the kernels observed within the tiniest CNN are composite filters that encode multiple fundamental visual features, including luminance and chromatic contrast, orientation and shape. This allows them to capture both the presence and geometric structure of specular highlights, supporting robust extraction of highlights across a range of different geometries. Although such a simple filter was far from the best of the models we tested for predicting human gloss judgements, it is striking that its correlation coefficient reaches 75.3% of the upper limit set by inter-observer consistency, yet is highly interpretable, yielding insights into key features that observers probably rely on to make their gloss matches.

Analysis of three-layer models

Figure 5a illustrates the computation flow of the example three-layer model with 64 kernels. The model processes the input image through multiple layers of convolution, pooling, normalization and nonlinear activation, ultimately using linear regression to predict the gloss value. The leftmost images show nine example kernels in the first convolutional layer, which show Gaussian blobs with different spatial scales, contrasts, colour tunings and polarities.

We examined this three-layer network, 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, using 324 images that were not included in the training dataset for this network. We first extracted activation maps after each convolutional layer (marked by stars) and aggregated the signals over the map using max pooling for the first layer and average pooling for the third layer. This resulted in 64 scalar values per layer, which were plotted on a 2D plane using t-distributed stochastic neighbour embedding (t-SNE; Fig. 5b). 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 (the plot does not label specific shapes). 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. To quantify this progression, we trained a linear regressor with 10-fold cross-validation to predict human gloss judgements from each stage's readout. Pearson's correlations (mean ± s.d.), averaged over 10 folds and 24 networks, were 0.13 ± 0.13 for pixel space, 0.64 ± 0.074 for layer 1, 0.67 ± 0.076 for layer 2 and 0.79 ± 0.0586 for layer 3. A one-way repeated-measures analysis of variance revealed a main effect of layer ($F(3, 69) = 307.1$, $P < 0.001$, Cohen's $f = 3.65$, 95% CI 2.99 to 4.30). Bonferroni-corrected pairwise comparisons ($\alpha' = 0.0083$) confirmed the following significant differences: pixel space < layer 1, layer 2 and layer 3; layer 1 < layer 3; and layer 2 < layer 3. This suggests that the first convolutional layer most strongly contributes to gloss judgements, with subsequent layers making small but significant adjustments to improve prediction, as well as arranging objects by their shape characteristics (note the bulbous-spiky organization[41,42] in the layer 3 representation)

Taken together with the analysis of the single-kernel network, these observations suggest our gloss judgements can be predicted well by a combination of well-known, low-level filtering mechanisms within the visual system.

Evaluation of model generalization across supplementary image sets

We conducted three additional generalization tests using (1) object images with manipulated specular highlights, (2) an additional set of 42,120 rendered images, and (3) real-world photographs, as detailed below.

Gloss illusion caused by manipulated specular highlights

Surprisingly, given the simplicity of the internal representations, we also found that the human-like networks predicted a number of known gloss effects[8,43,44] as shown in Fig. 6. For instance, the networks accurately anticipated the decrease in perceived glossiness when the specular highlight is rotated (Fig. 6a), moved horizontally (Fig. 6b), when surface roughness increases (Fig. 6c) and when the underlying specular reflectance weakens (Fig. 6d). The decrease was steeper for the three-layer models compared with 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 increases or specular contrast decreases, as this correlates with a decrease in local intensity of the specular component. It is somewhat 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, this local intensity profile probably arises from the combined contribution of both specular highlights and diffuse components when they are correctly aligned[45]. 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

https://doi.org/10.1038/s41562-026-02445-0

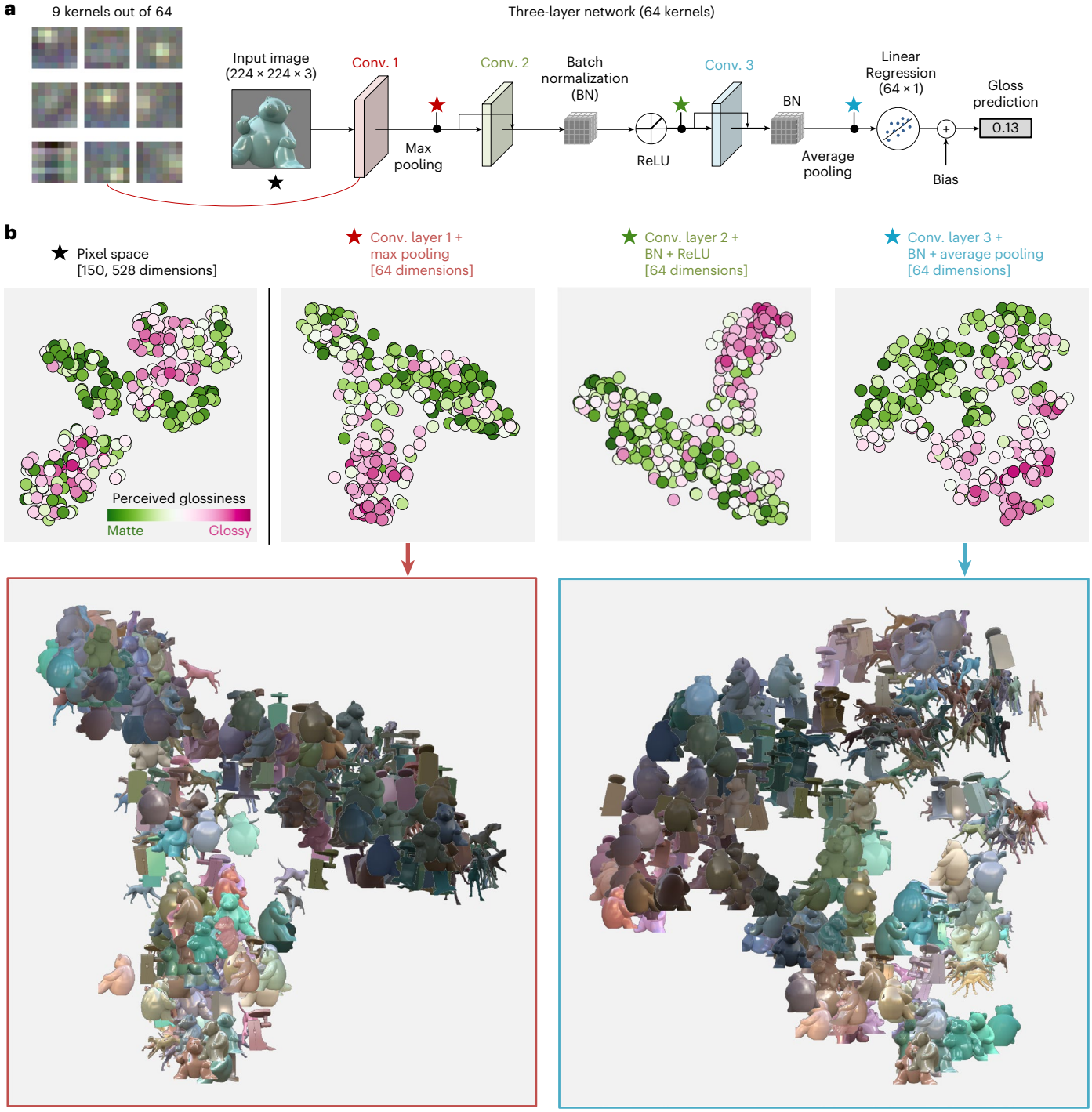

Fig. 5 | Analysis of a three-layer model with 64 kernels trained on images with human labels. a, In the first convolutional layer, the input image was convolved by a set of 64 kernels, 9 of which (7 by 7 pixels) we visualize on the left side. In the second and third convolutional layer, activation maps from the first convolutional layer were further convolved by the two sets of kernels subsequently and average pooling was taken for each of 64 activation maps from the third convolutional layer. Between convolutional layers 2 and 3, batch normalization and/or ReLU are applied. There were skip connections between layers to bypass the information if beneficial. Then the pooled features were used as input for a linear regression model, which predicted gloss level. **b**, Example of internal representation in a three-layer network model. 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*-SNE to project onto a 2D plane. The bottom two figures show representations from the first and third convolutional layers. For comparison, we also performed *t*-SNE analysis on the same images in the pixel space (leftmost plot), where matte and glossy objects are not separated and objects are primarily clustered by shape (not shown here).

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[5,9,12,13,46,47], a significant portion of gloss perception may actually be accounted for by more basic mechanisms.

https://doi.org/10.1038/s41562-026-02445-0

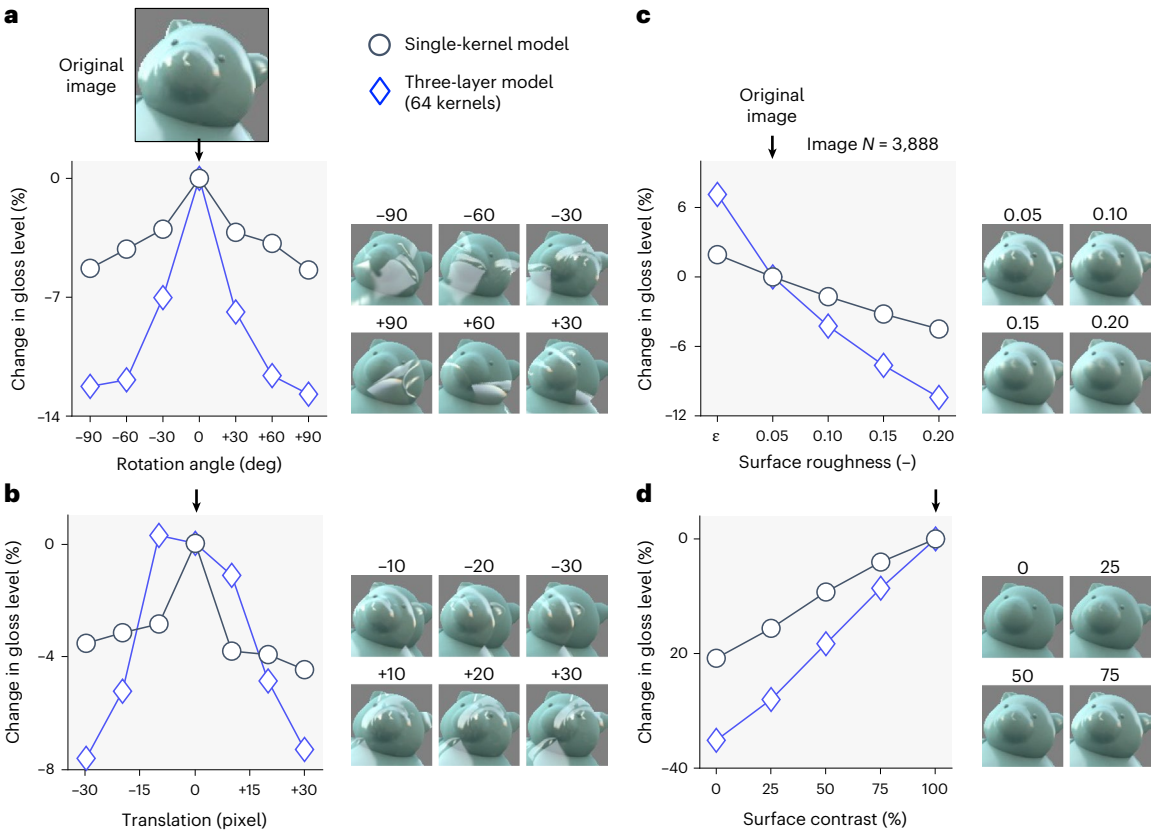

Fig. 6 | Evaluation of single-kernel and three-layer network models for three known perceptual effects. The dark-blue circles and blue diamonds show model responses for the one-layer model and three-layer models, respectively. **a**, Effect of rotating specular structure. The top image shows the original 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*.

Additional rendered images

Serrano et al.[48] generated 42,120 images from 9 lighting environments (Fig. 7a), 9 object geometries (Fig. 7b) and 520 bidirectional reflectance distribution functions. For each image, they collected ratings of various perceptual attributes including glossiness through crowdsourcing, providing an excellent validation set for our models. We tested our single-kernel model and three-layer model with 64 kernels, both trained on all 3,888 images. To prepare the input, we used object masks to exclude non-object regions, filling them with a mid-grey colour (sRGB = [127, 127, 127]), and then fed the resulting images into our models to generate gloss predictions.

Figure 7c shows Pearson's correlation between the single-kernel model predictions and human gloss ratings. The single-kernel model, despite its minimal architecture, achieves a mean correlation of 0.63 (s.d. 0.099) and 0.64 (s.d. 0.061) across lighting environments and shapes, respectively, performing comparably to our own dataset (mean $r = 0.65$). There is a general trend that geometries with smoother surfaces are better predicted by the model, although there are some exceptions, such as the bunny (geometry 4). By contrast, the three-layer model shown in Fig. 7d—which outperforms the single-kernel model on our dataset (two-tailed paired *t*-test; $t(23) = 8.49$, $P < 0.001$, Cohen's $d = 1.733$, 95% CI 1.087 to 2.364)—shows mean correlations of 0.53 (s.d. 0.056) across lighting environments and 0.65 (s.d. 0.058) across shapes. Similarly, the deep ResNet-52 gloss model by Serrano et al., when applied to our 3,888 images, shows a weak correlation of 0.29 (not shown), highlighting a trade-off between model complexity and generalization. Clearly human perception does not suffer from such generalization limits, yet this probably relates to our vastly larger visual diets.

Real-world photographs

We further evaluated generalization to real-world photographs of everyday materials using 185 images compiled from multiple sources, including the Flickr Material Database[39], visuo-haptic studies by Baumgartner et al.[49,50], a functional magnetic resonance imaging study by Jacobs et al.[51] and a perceptual material classification study by Wiebel et al.[52]. The dataset covers a diverse range of material categories, such as metal, glass, plastic, fur and leather, fluids and fabric. All images had been consistently classified as either matte or glossy by all eight observers in a previous study[11].

Figure 8a shows the predictions of the single-kernel model and the three-layer model, plotted on the *x* axis and *y* axis, respectively. The classification accuracy, based on the threshold indicated by the dashed lines in the plot, was 91.9% for the single-kernel model ($d' = 2.79$; a bias-free measure of discrimination sensitivity) and 71.9% for the three-layer model ($d' = 1.08$). This again demonstrates the superior generalization performance of the simpler model. Notably, the misclassified images lie close to the category boundary, with a mean distance of 0.0079 (s.d. 0.0081), which corresponds to approximately 7% of the model's prediction range.

This result is intriguing in that even a simple filter-based model can capture the diverse patterns of specular reflection found across different real-world materials (Fig. 8b) with high accuracy, while remaining robust against being misled by the texture patterns present in many matte images (Fig. 8c). Interestingly, the three-layer model, which performed well on our own dataset, showed reduced accuracy in this test. This suggests that even relatively shallow models, such as those with only three layers, can be prone to overfitting, pointing to the often-overlooked value of using simpler computational strategies.

https://doi.org/10.1038/s41562-026-02445-0

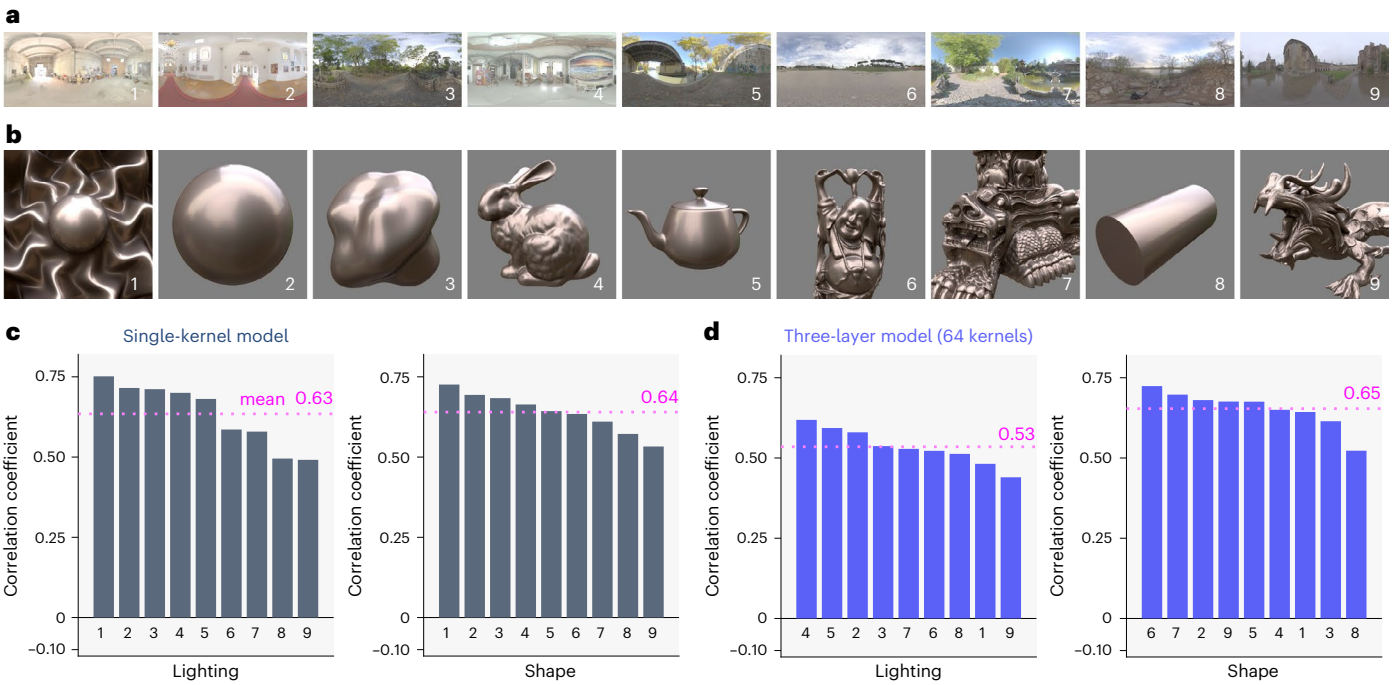

Fig. 7 | Results of a generalization test using Serrano dataset. a,**b**, Nine lighting environments (**a**) and shapes (**b**) used to generate rendered images. **c**, Pearson's correlation coefficient between the model response of the single-kernel model trained by all 3,888 images and perceptual rating measured in crowd-source experiment[47] for each lighting and shape. Each bar shows the correlation across 4,680 images. **d**, The case for a three-layer model trained by all 3,888 images, which shows slightly lower generalization ability compared with the single-kernel model.

Together with the results on manipulated specular highlights and Serrano's 42,120 images, this validation shows that the single-kernel model achieves an effective balance between simplicity and performance, making it a strong candidate for gloss computation across diverse image sets.

Discussion

This study is inspired by a fundamental question in behavioural 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 judgements of object gloss from images 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 judgements from object images probably do not involve sophisticated inverse optics computations. This notion has been suggested previously[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 filter with a bright blob and ridges is intriguing as it appears to capture a wide range of specular structures despite its simplicity. This finding directly supports the idea that gloss perception relies, to a non-trivial extent, on relatively simple image features, as suggested by Motoyoshi et al.[6]. Similarly, previous studies suggested that the visual system may use the steepness of the intensity gradient to separate shading into shape and glossiness components[53,54]. Much as cone photoreceptors provide foundational signals for colour perception and other visual functions, bright blob and ridge detectors may serve as a key component of gloss perception, and perhaps other distal perceptual attributes too[55–57]. In light of this, one hypothesis is that low-level computations supply a compact set of general-purpose image features that can be efficiently recombined through flexible, task-dependent selection mechanisms—gloss perception in the present case—underpinning the remarkable versatility of human material perception.

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. We would caution against interpreting our findings as suggesting the human visual system contains 'gloss detector' filters resembling the kernels that emerged in our models. Representations in the human visual system have to underpin far more tasks than just gloss perception. Instead, we see the emergent kernels as a data-driven means to identify the relevant image information that observers draw on to perform the gloss match task while noting that multikernel models (3-layer, 64 kernels) better predict human judgements overall. Yet, the combination of controllable computer-generated training sets, large-scale human perceptual data and deep learning with small, interpretable models seems to be a promising avenue for unravelling the computations underlying other 'mid-level' visual inferences.

The diagnosticity of diagonal specular reflections is also consistent with an ecological perspective, given the distribution of orientation signals in natural images[58]. Because vertical and horizontal edges are pervasive in the environment, these axis-aligned contrasts are so common that their presence alone provides little information about gloss. By contrast, diagonal signals are rarer, so their presence in the image is a stronger cue that the cause is a specular reflection. Interestingly, the geometry 1 used in Serrano's dataset (Fig. 7) is known to produce a particularly strong gloss impression[59], and its spatial pattern resembles the feature captured by our single kernel. Consistent with this, our model predicted human gloss judgements most accurately for this geometry. Finally, although the present study focused on isotropic materials, many

https://doi.org/10.1038/s41562-026-02445-0

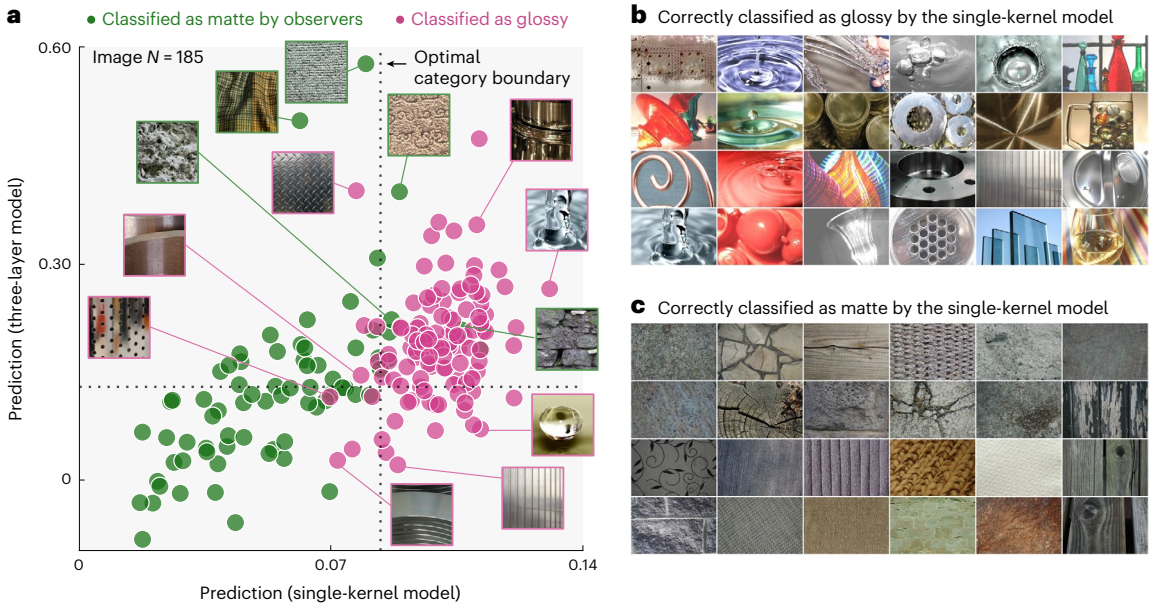

Fig. 8 | Generalization performance on real-world material photographs. **a**, A scatter plot comparing model responses for each image labelled as either matte or glossy based on behavioural data from a previous study[11]. The *x* axis represents the single-kernel model, and the *y* axis represents the three-layer model. Green and pink circles show images consistently rated as matte and glossy, respectively, by all eight observers. The dotted lines indicate the optimal category boundaries that achieve the highest classification accuracies: 91.9% (d' = 2.79) for the single-kernel model and 71.9% (d' = 1.08) for the three-layer model. **b**, Example images correctly classified as glossy by the single-kernel model. **c**, Example images correctly classified as matte.

natural and artificial materials—such as brushed metals, silk fabrics and polished wood—are anisotropic and often produce directionally elongated specular highlights, suggesting that the mechanisms captured by our kernel model may be applicable to these material types as well.

One key aspect of our approach is the use of large-*N* psychophysical experiments. Because 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 behaviours with sufficient resolution. Conversely, when the analysis was limited to a small number of shapes or lighting environments, some cases happened to show relatively high correlations between perceptual gloss judgements and the physical ground truth (Supplementary Figs. 4 and 5). This correlation is not inherent to our perceptual judgement but is influenced by the stimulus set. Moreover, it is worth noting that the behavioural error pattern we observed shows a strong asymmetry: as illustrated in Fig. 2e, there are far fewer cases of 'false positives' (objects that are not physically shiny but appear so) than 'misses' (high-reflectance objects that do not look glossy). This probably arises from basic optical constraints—it is difficult to produce salient specular highlights on low-reflectance objects, whereas it is relatively easy to eliminate highlights on high-reflectance ones; for example, placing a glossy object in a diffuse lighting environment or rotating a planar glossy object so that it reflects the ground rather than the sun can make a shiny object appear matte. Recent studies have made significant efforts to collect extensive human judgement data across a large number of images[60,61]. To understand visual mechanisms, we must account for both successes and errors in perception, and large-scale measurement helps capture these diagnostic behavioural 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 behaviour. Combining our approach with unsupervised learning has also shown some potential for reducing the cost of collecting labelled images[62–64].

There is a general concern about the lack of calibration and control over viewing conditions in online experiments. Our stimuli were presented assuming an sRGB display profile, but individual observers' monitors may deviate from this assumption (for example, in chromaticity or gamma of each RGB phosphor). Importantly, however, our glossiness was measured using an asymmetric matching task—a relative judgement in which both test and reference images were viewed under the same display conditions. Thus, even if an observer's monitor had an atypical gamma or contrast profile, it would have affected both images similarly and is unlikely to have systematically biased their settings. Moreover, as shown in Supplementary Fig. 6 (laboratory experiment), we observed a high degree of consistency between online and offline data, suggesting that variations in viewing conditions did not substantially influence observers' gloss matches.

A related point is that neither online nor laboratory-based experiments used a high-dynamic-range (HDR) scene. To ensure compatibility with standard monitors, we deliberately avoided extremely high dynamic range lighting environments and viewpoints producing strong specular highlights. The maximum specular reflectance was capped at 0.0999 (9.99%), which still yielded highly glossy appearances. Using a higher value could allow a wider range of glossiness to be explored, as demonstrated in recent HDR display work[65]. This constraint is shared by most material perception studies and shows an avenue for future investigation.

Our model validation results highlight the generalization challenges inherent to complex models. While deep learning has addressed many challenges in visual perception, it remains uncertain whether such models capture the mechanisms underlying human vision[66]. A key strength of our approach is its capacity to systematically explore specific filter designs and their combinations, enabling a more comprehensive search through the space of low-level computations. Our results may provide insight into long-standing questions in vision science, specifically whether complex perceptual systems like material perception originate from low-level visual processes[67].

One outcome of this study is the creation of a publicly available database containing 3,888 images. While many databases exist for labelled 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.

To conclude, our study has provided a data-driven perspective on the computational mechanisms in human gloss judgements using large-scale perceptual measurement and CNN-based modelling. We found a relatively simple and biologically plausible linear model that predicts the idiosyncrasies of human gloss constancy, but also several established perceptual gloss effects beyond the training range, as well as gloss perception for photographs of real-world objects.

Methods

Stimuli

The detailed procedure for stimulus generation is explained in our previous study[30]. All reference and test images were generated using the physically based rendering software Mitsuba v 0.6[68]. Images were generated at 512 × 512 pixels in *XYZ* 1931 format. For the online experiment, the image was converted to sRGB, normalized using the 97th percentile pixel value, and gamma correction was applied. The spatial resolution of 512 × 512 pixels was used for the psychophysical experiment, but reduced to 224 × 224 pixels for the network's inputs to shorten the training time. All images contained a single object at the centre of the scene, to which we applied diffuse and specular reflectance using the Ward reflectance model[69,70] with a fixed surface roughness level of 0.05. The specular reflectance value was randomly sampled from 0.0029 (weakly glossy) to 0.0999 (highly glossy). The 36 objects consisted of three-dimensional models of varying complexity, including both three-dimensional scans of real objects and artist-created models, spanning natural and human-made artefacts and enabling us to capture how different geometries influence highlights and other indicators of gloss. We applied image-based environmental illumination[71] to the scene. Each lighting environment had been photographically captured from a specific real-world location, with each pixel representing the colour and intensity of light arriving from a particular direction at a single point in space—thereby capturing illumination from all directions. We used 36 lighting environments, collected from the Debevec database[71], Southampton-York Natural Scenes[72] and Freebies (https://hdrmaps.com/freebies/; accessed May 2021). Further details are provided in Supplementary Fig. 3. Image-based environmental illumination was set at infinity, with the camera positioned at the object's height and facing it directly. Interreflections were included in the rendering. The object was randomly rotated around the *Z* axis, while rotations around the *X* and *Y* axes were limited to ±15°, maintaining an upright orientation but allowing for some variation (see Fig. 1a for the *XYZ* rotation axes). We deliberately avoided scenes with extremely high dynamic range lighting, allowing the images to be displayed on standard monitors without substantial tone compression. 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* (ref. 73), a perceptually linear characterization of Ward specular reflectance associated with the object to render a set of images, 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 colour was used, with fixed lightness at 50, and random chroma and hue between 8–26 and 0°–360° in the $L^*a^*b^*$ perceptually uniform colour space[74], respectively. A full list of object shapes and environmental illuminations is shown in Supplementary Figs. 2 and 3.

To generate the textured object images used for validation, we selected 12 object geometries from the original set of 36 that had smooth, continuous surfaces suitable for texture mapping. We then created 1,296 images by applying achromatic blob textures with either the same luminance as the body colour or 1.5 times higher (Supplementary Fig. 9a).

For the online experiment, the 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. At least three observers were recruited for each set. For two of these sets, we recruited 21 and 23 online observers, respectively, and compared their responses with those of 20 observers from the laboratory-based experiment to validate the online data (see 'Laboratory experiment' in the Supplementary Information).

Experiment

The main experiment was conducted online using PsychoPy (version 2022.1.4)[75] 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 Supplementary Fig. 6).

Observers and Ethics

A total of 467 observers participated in the online experiment, with data from 295 analysed (see exclusion criteria in a later section). Observers had normal or corrected-to-normal visual acuity and normal colour vision (self-report). Their age ranged from 19 to 65 (mean 41.4; s.d. 12.4) years. 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 hypothesis. 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[76] 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, colour, viewpoint and lighting environment (asymmetric matching[77,78]). 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 (sRGB = [127, 127, 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, totalling 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 min (4.89 s per trial).

Exclusion criteria

For quality control, we set two exclusion criteria for online experimental data. 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 median response time shorter than the

median minus two s.d. of laboratory-based experiments. The rationale is that a matching task requires comparison and adjustment; decisions at implausibly short latencies are a strong indicator of inattentive or automated responding. 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. In contrast to the online experiment, no data were excluded from the laboratory validation experiment.

Computational modelling

Fitting of single-kernel models. To quantitatively describe the spatial structure of the filters developed in network models, we modelled them as the sum of a 2D Gaussian $G(x,y)$ and multiple oriented ridge functions $R_k(x,y)$ as defined in equation (1):

$$M(x,y) = G(x,y) + \sum_{k=1}^{n} R_k(x,y) \quad (1)$$

The 2D axis-aligned Gaussian function was defined as equation (2):

$$G(x,y) = a_g \times \exp\left(-\left[\frac{(x-x_0)^2}{2\sigma_{gx}^2} + \frac{(y-y_0)^2}{2\sigma_{gy}^2}\right]\right) \quad (2)$$

where a_g is amplitude, (x_0,y_0) is the centre, and σ_{gx}, σ_{gy} are the s.d. along each axis.

Each oriented ridge was defined as equation (3):

$$R_k(x,y) = a_{R_k} \times \exp\left(-\frac{[(x-x_{0,R_k})\sin\varphi_{R_k} - (y-y_{0,R_k})\cos\varphi_{R_k}]^2}{2\sigma_{R_k}^2}\right) \quad (3)$$

where a_{Rk} is amplitude, $(x_{0,Rk}, y_{0,Rk})$ is the ridge centre, σ_{Rk} is the width perpendicular to the ridge and φ_{Rk} is the orientation. The index k takes values 1 and 2.

We converted each kernel from sRGB to *XYZ* and used the *Y* (luminance) channel for fitting to simplify the analysis. Optimization was performed using MATLAB's lsqcurvefit function. The functions were fitted sequentially: first the primary ridge, then the secondary ridge and finally the 2D Gaussian. If a second ridge component was not identified through the optimization, only a single ridge was fitted to the kernel.

Luminance statistics models. The *Y* (luminance) channel from raw *XYZ* images was extracted. We implemented nine models to determine glossiness using the mean, median, s.d., 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—derived from the specular reflection component of each image, inspired by an influential prior study[12]. Each model involved free parameters, which were optimized to align most closely with human gloss judgements. 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[79]. 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 cut-off 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[80]. We then computed the root-mean-square contrast, equivalent to the s.d. 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 judgements.

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° in 30° steps, using the image centre as the rotation axis. For translation, we shifted the specular component horizontally from −30 to 30 pixels in 10-pixel steps relative to the image size of 224 × 224 pixels. 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 ε (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 each manipulated specular image with the corresponding diffuse image.

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

Network architectures. We chose a ResNet architecture for its effectiveness in object recognition tasks[81] and its modular design, which enables systematic adjustment of 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 nonlinear operations such as rectified linear unit (ReLU) or batch normalization. The three-layer model begins with a convolutional layer followed by max pooling, then two additional convolutional layers, batch normalization, and ReLU with a skip connection that allows the input to bypass the signals from the previous layer if beneficial. This is followed by average pooling and, finally, a regression layer that computes a weighted sum of the averaged kernel outputs, where the weights are learned free parameters, along with a bias term to predict continuous gloss values.

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. One-layer networks were trained for 30 epochs, and three-layer networks for 90 epochs, with both trained until their performance reached a plateau. Training began with an initial learning rate of 0.01, which was reduced 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 graphics processing units each with 80 GB video memory, Dual AMD Rome 7742 central processing unit, 128 cores total, 2.25 GHz base) running Ubuntu 22.04 LTS. While each individual model was lightweight, the system

enabled efficient parallel training of the many networks required for cross-validation.

Training and validation procedure. Our evaluation procedure for the one-layer and three-layer models 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 had been trained on 33 shapes (3,564 images) and tested on the 3 novel lighting environments (324 images). Again, this had been 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 Fig. 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. All images used here are 8-bit sRGB images, the same format as those used in the online experiments.

ResNet18 models (with or without additional training images) and three-layer models with additional training images were not evaluated using 24-fold cross-validation owing to the substantial computational time required for training each model. Instead, for these models, one of the 24 training–test pairs was randomly selected, and performance was assessed using a single evaluation (for models plotted with orange symbols in Fig. 3, the training set included additional images).

For each training session, the 3,564 images were augmented to 320,760 images through horizontal flipping and random size cropping, where a randomly selected subregion (80–100% of the image) was rescaled to the network's input size. 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 judgements.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All behavioural data, stimulus images and model data are available via GitHub at https://github.com/takuma929/gloss_tinynetworks under a non-restrictive MIT license and via Zenodo at https://doi.org/10.5281/zenodo.19511809 (ref. 82). Behavioural data and model data from the Serrano dataset, which were used to validate our models, are available at https://mig.mpi-inf.mpg.de/ (behavioural data) and via GitHub at https://github.com/Hans1984/material-illumination-geometry (model data).

Code availability

All custom analysis codes used to reproduce figures in this Article are available via GitHub at https://github.com/takuma929/gloss_tinynetworks and via Zenodo at https://doi.org/10.5281/zenodo.19511809.

References

1. Fleming, R. W. Material perception. *Annu. Rev. Vis. Sci.* **3**, 365–388 (2017).
2. Nishida, S. Image statistics for material perception. *Curr. Opin. Behav. Sci.* **30**, 94–99 (2019).
3. Chadwick, A. & Kentridge, R. The perception of gloss: a review. *Vision Res.* **109**, 221–235 (2015).
4. Blake, A. & Bülthoff, H. Does the brain know the physics of specular reflection?. *Nature* **343**, 165–168 (1990).
5. Anderson, B. L. Mid-level vision. *Curr. Biol.* **30**, R105–R109 (2020).
6. Motoyoshi, I. et al. Image statistics and the perception of surface qualities. *Nature* **447**, 206–209 (2007).
7. Sharan, L. et al. Image statistics for surface reflectance perception. *J. Opt. Soc. Am. A* **25**, 846–865 (2008).
8. Anderson, B. L. & Kim, J. Image statistics do not explain the perception of gloss and lightness. *J. Vis.* **9**, 10 (2009).
9. Kim, J. & Anderson, B. L. Image statistics and the perception of surface gloss and lightness. *J. Vis.* **10**, 3 (2010).
10. Sawayama, M. & Nishida, S. Y. Material and shape perception based on two types of intensity gradient information. *PLoS Comput. Biol.* **14**, e1006061 (2018).
11. Wiebel, C., Toscani, M. & Gegenfurtner, K. R. Statistical correlates of perceived gloss in natural images. *Vis. Res.* **115B**, 175–187 (2015).
12. Marlow, P. J., Kim, J. & Anderson, B. L. The perception and misperception of specular surface reflectance. *Curr. Biol.* **22**, 1909–1913 (2012).
13. Marlow, P. J. & Anderson, B. L. Generative constraints on image cues for perceived gloss. *J. Vis.* **13**, 2 (2013).
14. Fleming, R. W., Dror, R. O. & Adelson, E. H. Real-world illumination and the perception of surface reflectance properties. *J. Vis.* **3**, 347–368 (2003).
15. Zhang, F., de Ridder, H., Barla, P. & Pont, S. A systematic approach to testing and predicting light-material interactions. *J. Vis.* **19**, 11 (2019).
16. 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.* **20**, 13 (2020).
17. Gigilashvili, D. & Islam, A. J. The role of shape in modeling gloss. *Proc. 30th Color Imaging Conf.* **30**, 271–276 (2022).
18. Gigilashvili, D., Thomas, J., Pedersen, M. & Hardeberg, J. Y. On the appearance of objects and materials: Qualitative analysis of experimental observations. *J. Int. Colour Assoc.* **27**, 26–55 (2021).
19. Vangorp, P., Laurijssen, J. & Dutre, P. The influence of shape on the perception of material reflectance. *ACM Trans. Graphics* **26**, 77 (2007).
20. Chen, B. et al. The effect of geometry and illumination on appearance perception of different material categories. *Vis. Comput.* **37**, 2975–2987 (2021).
21. Wendt, G., Faul, F., Ekroll, V. & Mausfeld, R. Disparity, motion, and color information improve gloss constancy performance. *J. Vis.* **10**, 7 (2010).
22. Adams, W. J., Kucukoglu, G., Landy, M. S. & Mantiuk, R. K. Naturally glossy: Gloss perception, illumination statistics, and tone mapping. *J. Vis.* **18**, 4 (2018).
23. Honson, V. et al. Effects of shape, roughness and gloss on the perceived reflectance of colored surfaces. *Front. Psychol.* **11**, 485 (2020).
24. Cheeseman, J. R., Ferwerda, J., Morimoto, T. & Fleming, R. Gloss discrimination: Toward an image-based perceptual model. *J. Vis.* **25**, 6 (2025).
25. Wijntjes, M. W. A. & Pont, S. C. Illusory gloss on Lambertian surfaces. *J. Vis.* **10**, 13 (2010).
26. Radonjić, A., Cottaris, N. P. & Brainard, D. H. The relative contribution of color and material in object selection. *PLoS Comput. Biol.* **15**, e1006950 (2019).
27. Norman, J. F., Todd, J. T. & Phillips, F. Effects of illumination on the categorization of shiny materials. *J. Vis.* **20**, 2 (2020).
28. Ho, Y. X., Landy, M. S. & Maloney, L. T. Conjoint measurement of gloss and surface texture. *Psychol. Sci.* **19**, 196–204 (2008).
29. Doerschner, K., Maloney, L. T. & Boyaci, H. Perceived glossiness in high dynamic range scenes. *J. Vis.* **10**, 11 (2010).
30. Morimoto, T. et al. Color and gloss constancy under diverse lighting environments. *J. Vis.* **23**, 1–25 (2023).
31. Prokott, K. E., Tamura, H. & Fleming, R. W. Gloss perception: searching for a deep neural network that behaves like humans. *J. Vis.* **21**, 14 (2021).

32. Tamura, H., Prokott, K. E. & Fleming, R. W. Distinguishing mirror from glass: a 'big data' approach to material perception. *J. Vis.* **22**, 4 (2022).
33. Goncalves, N. R. & Welchman, A. E. 'What Not' detectors help the brain see in depth. *Curr. Biol.* **27**, 1403–1412 (2017).
34. Rideaux, R. & Welchman, A. E. But still it moves: static image statistics underlie how we see motion. *J. Neurosci.* **40**, 2538–2552 (2020).
35. Rideaux, R. & Welchman, A. E. Exploring and explaining properties of motion processing in biological brains using a neural network. *J. Vis.* **21**, 11 (2021).
36. Anderson, B. L., Storrs, K. R. & Fleming, R. W. Where do the hypotheses come from? Data-driven learning in science and the brain. *Behav. Brain Sci.* **46**, e386 (2023).
37. Nishida, S. Y. & Shinya, M. Use of image-based information in judgments of surface-reflectance properties. *J. Opt. Soc. Am. A* **15**, 2951–2965 (1998).
38. Judd, D. B. et al. Spectral distribution of typical daylight as a function of correlated color temperature. *J. Opt. Soc. Am.* **54**, 1031–1040 (1964).
39. Young, R. A. The Gaussian derivative model for spatial vision: I. Retinal mechanisms. *Spat.Vis.* **2**, 273–293 (1987).
40. Sharan, L., Rosenholtz, R. & Adelson, E. H. Accuracy and speed of material categorization in real-world images. *J. Vis.* **14**, 12 (2014). Article.
41. Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
42. Morgenstern, Y. et al. An image-computable model of human visual shape similarity. *PLoS Comput. Biol.* **17**, e1008981 (2021).
43. Beck, J. & Prazdny, S. Highlights and the perception of glossiness. *Percept. Psychophys.* **30**, 407–410 (1981).
44. Todd, J. T., Norman, J. F. & Mingolla, E. Lightness constancy in the presence of specular highlights. *Psychol. Sci.* **15**, 33–39 (2004).
45. Marlow, P., Kim, J. & Anderson, B. L. The role of brightness and orientation congruence in the perception of surface gloss. *J. Vis.* **11**, 16 (2011).
46. Marlow, P. J., Todorović, D. & Anderson, B. L. Coupled computations of three-dimensional shape and material. *Curr. Biol.* **25**, R221–R222 (2015).
47. Anderson, B. L. & Marlow, P. J. Perceiving the shape and material properties of 3D surfaces. *Trends Cogn. Sci.* **27**, 98–110 (2023).
48. Serrano, A. et al. The effect of shape and illumination on material perception: Model and applications. *ACM Trans. Graphics* **40**, 4 (2021).
49. Baumgartner, E., Wiebel, C. B. & Gegenfurtner, K. R. Visual and haptic representations of material properties. *Multisens. Res.* **26**, 429–455 (2013).
50. Baumgartner, E., Wiebel, C. B. & Gegenfurtner, K. R. A comparison of haptic material perception in blind and sighted individuals. *Vision Res.* **115**, 238–245 (2015).
51. Jacobs, R. H., Baumgartner, E. & Gegenfurtner, K. R. The representation of material categories in the brain. *Front. Psychol.* **5**, 146 (2014).
52. Wiebel, C. B., Valsecchi, M. & Gegenfurtner, K. R. The speed and accuracy of material recognition in natural images. *Atten Percept Psychophys.* **75**, 954–966 (2013).
53. Tyler, C. W. Diffuse illumination as a default assumption for shape-from-shading in the absence of shadows. *J. Imaging Sci. Technol.* **42**, 319–325 (1998).
54. Langer, M. S. & Zucker, S. W. Shape from shading on a cloudy day. *J. Opt. Soc. Am. A* **11**, 467–478 (1994).
55. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**, 91–110 (2004).
56. Akbarinia, A. & Parraga, C. A. Colour constancy beyond the classical receptive field. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 830–845 (2017).
57. Akbarinia, A. & Parraga, C. A. Feedback and surround modulated boundary detection. *Int. J. Comput. Vis.* **126**, 583–603 (2018).
58. Girshick, A. R. et al. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat. Neurosci.* **14**, 926–932 (2011).
59. Havran, V., Filip, J. & Myszkowski, K. Perceptually motivated BRDF comparison using a single image. *Comput. Graph. Forum* **35**, 1–12 (2016).
60. Sawayama, M. et al. Visual discrimination of optical material properties: a large-scale study. *J. Vis.* **22**, 17 (2022).
61. Lagunas, M. et al. A similarity measure for material appearance. *ACM Trans. Graphics* **38**, 135 (2019).
62. Storrs, K. R., Anderson, B. L. & Fleming, R. W. Unsupervised learning predicts human perception and misperception of gloss. *Nat. Hum. Behav.* **5**, 1402–1417 (2021).
63. Guerrero-Viu, J. et al. Predicting perceived gloss: do weak labels suffice?. *Comput. Graph. Forum* **43**, e15037 (2024).
64. Liao, C., Sawayama, M. & Xiao, B. Unsupervised learning reveals interpretable latent representations for translucency perception. *PLoS Comput. Biol.* **19**, e1010878 (2023).
65. Chen, B. et al. The effect of display capabilities on the gloss consistency between real and virtual objects. *SIGGRAPH Asia 2023 Conf. Papers* **90**, 1–11 (2023).
66. Bowers, J. S. et al. Deep problems with neural network models of human vision. *Behav. Brain Sci.* **46**, e385 (2023).
67. Adelson, E. H. On seeing stuff: the perception of materials by humans and machines. In *Proc. SPIE 4299, Human Vision and Electronic Imaging VI* (eds Rogowitz, B. E. & Pappas, T. N.) 1–12 (SPIE, 2001).
68. Jakob, W. Mitsuba: physically based renderer. *Mitsuba* https://www.mitsuba-renderer.org/ (2010).
69. Ward, G. J. Measuring and modeling anisotropic reflection. In *Proc. 19th Annual Conference on Computer Graphics and Interactive Techniques* (ed. Thomas, J. J.) 265–272 (ACM, 1992).
70. Geisler-Moroder, D. & Dür, A. A new ward BRDF model with bounded albedo. *Comput. Graph. Forum* **29**, 1391–1398 (2010).
71. Debevec, P. Rendering synthetic objects into real scenes: btraditional and image-based graphics with global illumination and high dynamic range photography. In *SIGGRAPH 98 Proc. 25th Annual Conference on Computer Graphics and Interactive Techniques* 189–198 (ACM, 1998).
72. Adams, W. J. et al. The Southampton-York Natural Scenes (SYNS) dataset: statistics of surface attitude. *Sci. Rep.* **6**, 35805 (2016).
73. Pellacini, F., Ferwerda, J. A. & Greenberg, D. P. Toward a psychophysically-based light reflection model for image synthesis. In *Proc 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)* (ed. Akeley, K.) 55–64 (ACM, 2000).
74. Carter, E. C. et al. *CIE 015:2018 Colorimetry* 4th edn (CIE, 2018).
75. Peirce, J. W. et al. PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* **51**, 195–203 (2019).
76. Li, Q. et al. Controlling for participants' viewing distance in large-scale, psychophysical online experiments using a virtual chinrest. *Sci. Rep.* **10**, 904 (2020).
77. Arend, L. & Reeves, A. Simultaneous color constancy. *J. Opt. Soc. Am. A* **3**, 1743–1751 (1986).
78. Brainard, D. H. & Wandell, B. A. Asymmetric color matching: how color appearance depends on the illuminant. *J. Opt. Soc. Am. A* **9**, 1433–1448 (1992).
79. Vu, C. T., Phan, T. D. & Chandler, D. M. S3: a spectral and spatial measure of local perceived sharpness in natural images. *IEEE Trans. Image Process.* **21**, 934–945 (2012).
80. Boyadzhiev, I., Bala, K., Paris, S. & Adelson, E. Band-sifting decomposition for image-based material editing. *ACM Trans. Graphics* **34**, 1–16 (2015).

81. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
82. Morimoto, T. et al. Human gloss perception and tiny neural networks. *Zenodo* https://doi.org/10.5281/zenodo.19511809 (2026).

Acknowledgements

We thank A. Serrano for providing the associated data for the rendered image dataset, and M. Toscani for providing the real-world photographs used to validate our models. We also thank A. Zentel for her assistance with data collection in the laboratory-based experiment. T.M. was supported by a Sir Henry Wellcome Postdoctoral Fellowship from Wellcome Trust (218657/Z/19/Z). This research was also supported by the DFG under Germany's Excellence Strategy (EXC 3066/1 'The Adaptive Mind', project no. 533717223) as well as SFB-TRR-135 'Cardinal Mechanisms of Perception' (project no. 222641018; TPs C1 and C2); a Marsden Fast-Start Grant awarded to K.R.S. by the Royal Society of New Zealand (MFP-UOA2109); the UKRI–Wellcome Physics of Life programme (EP/W023873/1; PhysFEM) awarded to H.E.S.; and European Research Council Advanced Grants Color3.0 (project no. 884116) awarded to K.R.G. and STUFF (project no. 101098225) awarded to R.W.F. For the purpose of open access, we have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Author contributions

T.M., A.A., K.R.S., J.R.C., H.E.S., K.R.G. and R.W.F. conceived and designed the study. T.M. curated the data. T.M. and A.A. performed the formal analyses. T.M., K.R.S., H.E.S., K.R.G. and R.W.F. secured funding for the project. T.M. conducted the investigation and collected the data. T.M., A.A. and R.W.F. developed the methodology. T.M. managed and administered the project, with support from R.W.F. H.E.S., K.R.G., and R.W.F. provided resources for the study and supervised the project. T.M. and A.A. developed the code to implement the machine-learning models, and T.M. developed additional supporting software used in the study. T.M. wrote the original draft of the manuscript, and all authors (T.M., A.A., K.R.S., J.R.C., H.E.S., K.R.G. and R.W.F.) reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-026-02445-0.

Correspondence and requests for materials should be addressed to Takuma Morimoto.

Peer review information *Nature Human Behaviour* thanks Michael Landy, Christopher W. Tyler and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Takuma Morimoto

Last updated by author(s): Jul 31, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Online data collection PsychoPy (ver 2022.1.4): https://www.psychopy.org/ Laboratory data collection MATLAB 2021a: https://uk.mathworks.com/products/matlab.html Psychtoolbox-3 : http://psychtoolbox.org/
Data analysis	Analyses of behavioural data and model responses were conducted using MATLAB (ver. 2020a). Model training was performed using Python (ver. 3.8.20). Custom MATLAB codes to reproduce figures in the manuscript are available at the project's GitHub page (https://github.com/takuma929/gloss_tinynetworks). Additional toolboxes and modules used are listed below. MATLAB Toolboxes: Computer Vision Toolbox (ver. 9.2): https://uk.mathworks.com/products/computer-vision.html Curve Fitting Toolbox (ver. 3.5.11): https://uk.mathworks.com/products/curvefitting.html Image Processing Toolbox (ver. 11.1): https://uk.mathworks.com/products/image-processing.html Optimization Toolbox (ver. 8.5): https://uk.mathworks.com/products/optimization.html PsychColorimetric functions in PsychToolbox-3: https://www.psychtoolbox.org/

Statistics and Machine Learning Toolbox (ver. 11.7): https://uk.mathworks.com/products/statistics.html

Python Modules:
numpy (ver. 1.24.1): https://github.com/numpy/numpy
opencv-python (ver. 4.10.0.84): https://github.com/opencv/opencv-python
pandas (ver. 2.0.3): https://github.com/pandas-dev/pandas
pillow (ver. 10.2.0): https://github.com/python-pillow/Pillow
scipy (ver. 1.10.1): https://github.com/scipy/scipy
torch (ver. 2.4.1+cu118): https://github.com/pytorch/pytorch
torchvision (ver. 0.9.1+cu118): https://github.com/pytorch/vision

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All behavioural data, stimulus images, model data are available on the GitHub page (https://github.com/takuma929/gloss_tinynetworks) under non-restrictive MIT license.

Behavioural data and model data from the Serrano dataset, which were used to validate our models, are available at : https://mig.mpi-inf.mpg.de/ (behavioural data) and https://github.com/Hans1984/material-illumination-geometry (model data).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Information on biological sex was collected via self-report on both the Prolific platform (for the online experiment) and the consent form (for the offline experiment). No information regarding gender shaped by social and cultural circumstances was collected. The sex distribution was relatively balanced: 159 females and 136 males participated in the online experiment, while 13 females and 7 males participated in the offline experiment. No sex- or gender-based analyses were conducted, as the primary aim of the study was to investigate visual mechanisms common across individuals, rather than to examine differences between sexes or genders.
Reporting on race, ethnicity, or other socially relevant groupings	No data on race, ethnicity, or other socially relevant categorization variables were collected in this study, as these factors were not the focus of the research design. The main online experiment included several hundred participants recruited via Prolific, a widely used online participant recruitment platform, which draws from a broad and diverse population base. While some systematic biases related to the characteristics of Prolific users may exist, the large and diverse sample size helps minimize the risk of major systematic bias affecting the main findings.
Population characteristics	See below.
Recruitment	For the online experiment, the study was posted on Prolific, and participation was open to any registered user who chose to take part. For the offline (laboratory) experiment, participants were recruited through the departmental contact list of the psychology department at Justus Liebig University Giessen, Germany. This sample consisted mainly of younger participants, which may have introduced some bias—for example, lens density can differ between younger and older individuals. However, since the experimental task involved appearance-level judgments of glossiness (rather than fine discrimination or threshold-level decisions), any such bias is unlikely to have had a substantial effect. Informed consent was obtained from all participants prior to both the online and offline experiments. As shown in Supplementary Figure S6, the results from the online and laboratory-based experiments were highly consistent, suggesting that any potential sampling bias did not significantly impact the main findings.
Ethics oversight	This study received ethics approval from the local ethics committee at Justus Liebig University Giessen, Germany.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	A quantitative experimental study investigating human visual perception of gloss in object images.
Research sample	Online experiment: A total of 467 participants were recruited via Prolific, a platform accessible to individuals from most Organisation for Economic Co-operation and Development (OECD) countries. After applying exclusion criteria (see below), data from 295 participants (159 females and 136 males), aged 19–65 years (mean = 41.4, SD = 12.4), were included in the final analysis. Only English-speaking participants were recruited due to the language of the consent process, but there were otherwise no restrictions on participant selection. Laboratory (offline) experiment: For laboratory validation, we recruited 20 participants (13 females, 7 males), aged 18–36 years (mean = 23.4, SD = 4.06). Participants were primarily undergraduate students from the psychology department at Justus Liebig University Giessen, Germany, consistent with common practices in vision science research. No language restrictions were applied for this sample. Additional dataset: To further validate our models, we utilized a large-scale dataset comprising 215,680 perceptual responses for 42,120 combinations of material, shape, and illumination (Serrano et al., ACM Trans. Graph., 2021). This dataset was collected via Amazon Mechanical Turk and included 3,217 participants (37% female; mean age 38.1, SD = 11.96), many of whom reported backgrounds in computer graphics, design, or art. Rationale: By incorporating both an internationally recruited online sample and a laboratory-based sample, and by leveraging a large, independently collected dataset, our research sample is more demographically diverse and representative than those of most prior offline-only studies in the field of vision science. Nevertheless, the online sample does not include individuals from all geographical locations, and our restriction to English-speaking participants means that individuals outside these criteria were not represented.
Sampling strategy	Both the online and offline experiments utilized a convenience sampling approach. The study was advertised openly, and any interested individuals who met the basic inclusion criteria (normal or corrected-to-normal vision; English proficiency for the online experiment) were eligible to participate. No statistical methods were used to predetermine sample size. Instead, the sample size was determined by practical considerations related to the study design and the requirements for training machine learning models. Specifically, the goal was to obtain behavioral data across a large set of images, with each image rated by at least three independent observers to ensure data reliability. The number of participants was also constrained by the aim to limit each session to approximately 30 minutes to minimize participant fatigue and maintain data quality. The sufficiency of the sample size was assessed retrospectively, based on the ability of the trained neural network models to generalize to independent datasets that were not used during model training. Our shallow network models demonstrated robust predictive performance, suggesting that the sample size was adequate for the intended analyses.
Data collection	For the online experiment, participants used their own laptop or desktop computers, with stimuli images presented on the computer screen. Responses were made using a mouse and keyboard. All instructions were provided on-screen before the start of the experiment. Participants were instructed to perform the task in a room with the lights turned off to ensure consistent viewing conditions. For the offline (laboratory) experiment, data collection was conducted in a controlled laboratory environment. Stimuli images were presented on a computer screen, and participants responded using a keyboard. During the instruction phase for the first observer, the researcher, research assistant, and participant were present. During all subsequent sessions, only the research assistant and participant were present. The research assistant was blind to both the experimental condition and the study hypothesis during data collection.
Timing	Data collection was conducted between 2nd Nov 2022 and 21st Nov 2022.
Data exclusions	For online experiments, two exclusion criteria were applied for data quality: (1) observers failing either of the catch trials were excluded; (2) observers whose median response time was faster than the laboratory sample's median minus two standard deviations were excluded. This process resulted in 295 observer datasets (63.2%) included in the final analysis. No data were excluded from the offline laboratory experiment.
Non-participation	No participants dropped out or declined participation to the experiment.
Randomization	In the online experiment, participants were divided into 54 groups, with each group assigned to judge a different set of images. Participants were sequentially allocated to the groups in the order they joined the study (i.e., the 1st participant was assigned to

group 1, the 54th to group 54, and the 55th to group 1). Data collection continued until each group included at least three participants after applying the exclusion criteria. In the offline laboratory experiment, no such group allocation was used; all participants completed the same set of conditions (within-subject design).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.