Unsupervised Learning Predicts Human Perception and Misperception of Gloss

Authors and affiliations:

Katherine R. Storrs[1] & Roland W. Fleming[1,2]

1. Department of Experimental Psychology, Justus Liebig University Giessen
2. Centre for Mind, Brain and Behaviour (CMBB), University of Marburg and Justus Liebig University Giessen

Contact details for corresponding author:

Katherine R. Storrs,
Justus-Liebig University Giessen,
Fachbereich 06 - Psychologie und Sportwissenschaft,
Abteilung Allgemeine Psychologie,
Otto-Behaghel-Strasse 10F,
35394 Giessen, Germany
tel. +49(0)641 9926127
email: katherine.storrs@gmail.com

Abstract

Reflectance, lighting, and geometry combine in complex ways to create images. How do we disentangle these to perceive individual properties, like the glossiness of a surface? We suggest that brains disentangle properties by learning to model statistical structure in proximal images. To test this, we trained unsupervised generative neural networks on renderings of glossy surfaces and compared their representations with human gloss judgments. The networks spontaneously cluster images according to distal properties such as reflectance and illumination, despite receiving no explicit information about them. Intriguingly, the resulting representations predict the specific patterns of ‘successes’ and ‘errors’ in human perception. Linearly decoding specular reflectance from the model’s internal code predicts human gloss perception better than ground truth, supervised networks, or control models, and predicts, on an image-by-image basis, illusions of gloss perception caused by interactions between material, shape, and lighting. Unsupervised learning may underlie many perceptual dimensions in vision, and beyond.

Introduction

A photograph of a glass of water might consist of large bright patches, sparkling dots, and low-contrast blurs—yet we immediately see these as reflections on the water's surface, bubbles within, and smudges on the glass. This ability to identify the distal physical causes of proximal image features is widely considered the central challenge of vision (1–7). Yet how we infer the outside world from ambiguous sensory data remains mysterious. Explanations often invoke 'inverse optics'—the idea that the brain models physical processes and estimates the most likely distal world given the image and some prior assumptions (4,8–11). It is not yet clear how brains might acquire such models in the first place, though,, since they cannot access the true state of the world (12–14). Indeed, for purely visual properties like glossiness, no other sensory modality could provide the 'ground truth' required to train the visual system.

Here we present a solution to this long-standing dilemma, by showing that it is possible to learn representations that approximately disentangle physical properties in distinctively human-like ways, simply by learning about the statistical regularities in images themselves, with no additional information about the scenes which gave rise to them. We take gloss perception as a paradigmatic case of perceptual inference in which multiple physical effects—lighting, shape and reflectance—must be teased apart. Importantly, while humans are generally good at gloss perception, they also exhibit some well-known errors (3,5,15–19) that could provide cues to the underlying computations. We find that by learning to efficiently compress and spatially predict images of glossy surfaces, an unsupervised generative deep neural network (DNN) not only clusters inputs by factors like material and illumination, but also strikingly reproduces many characteristic "misperceptions" of human observers.

For an image to look like a glossy surface in the first place, certain photo-geometrical constraints must be satisfied, e.g., highlights being correctly aligned with the surface geometry (7,58,59). However, within images satisfying these conditions, there are also large variations in the apparent *degree of glossiness.* The apparent glossiness of surfaces not only varies with specular reflectance, as expected (22–29), but also other

factors such as lighting and shape (15–18,30–32). It is these successes and failures of 'gloss constancy' that we seek to explain.

It might seem paradoxical to suggest that simply learning statistical regularities of images could ever teach the brain to see the outside world. Yet, since the 1950s, it has been understood that statistical redundancies in sensory data could help the brain identify meaningful structures (33–35). Using unsupervised learning to identify compact codes for natural image patches strikingly predicts the response properties of neurons in early visual cortex (34–38). Unfortunately, such 'efficient coding' approaches have not yet scaled beyond the initial encoding of images, into scene understanding (i.e., inferring the distal physical causes of sense data). However, recent advances in deep learning provide powerful new tools for testing whether scene properties—such as material and illumination—can be disentangled in a purely unsupervised, data-driven fashion.

To appreciate why this might work, consider that glossy surfaces like plastics or glazed ceramics are well approximated as weighted combinations of diffuse and specular component images. Although different surfaces produce complex and widely variable images, ultimately the variance due to reflectance can be summarised as a single parameter ('glossiness') describing the relative weight of the two components (26,39) (for simplicity, this ignores additional parameters controlling the shape of the bidirectional reflectance distribution function). Inferring such a parameter would be the *most compact* representation of differences between images of surfaces, as gloss varies. We hypothesized that learning highly compressed, low-dimensional representations of a set of surface images with sufficient fidelity to generate new samples that resemble the training set, may encourage a model to discover an embedding of the stimuli that re-organizes them approximately by their reflectance, rather than their pixel similarity (see also (14)).

Such inferences are far from trivial, but DNN methods have recently emerged for learning sophisticated models of image structure in the form of latent variables, which summarise how images differ from one another (40–44). DNNs achieve complex

transformations by passing input data through layers of units that apply weighted summations and non-linearities, roughly mimicking the operations of biological neurons (45–49). During training, connections between units are iteratively adjusted, to improve the network's performance on some learning objective. In supervised learning, networks are directly told what response to output for every input image in the training dataset. Supervised DNNs have revolutionised computer vision, achieving near-human object and face recognition (50–53), and are the best extant models of late ventral stream function (45–47,54–57). However, unlike humans, they are often fragile to tiny image perturbations (58,59) and over-rely on local texture (60,61). As models of human gloss perception, it is unclear from where the necessary training labels could come.

In unsupervised learning, networks are encouraged through various learning objectives to learn statistical regularities in the training data without being given any explicit labels. For example, autoencoder networks are trained to compress training images into compact descriptions, then reconstruct them as accurately as possible (44,62,63). Here, we use a variant known as a PixelVAE (41,42), which learns to both summarise and spatially predict images, in two connected processing streams (**Figure 1**). It is a generative model that can create completely novel images with high-order statistical structure similar to the training data—in our case images of glossy surfaces.

To test whether human-like gloss perception could arise through unsupervised learning of surface appearances, we rendered over 10,000 images from a virtual world consisting of frontal views of bumpy surfaces with either high ('gloss') or low ('matte') specular reflectance. Using renderings grants tight control over the statistics of the training environment, allowing us to guarantee that reflectance could not be trivially decoded from raw images and that physical factors varied independently of one another. Each image had a different random configuration of bumps, depth of surface relief and colour, and was illuminated by one of six natural light fields (**Figure 1A-B**). We then trained ten different instances of a PixelVAE (41,42) network on this dataset, to ensure results were robust to representational differences between training instances of the same architecture (64). The model's training objective was to learn a probabilistic distribution over pixel values that captured the structure and variability in

and across the training images at all spatial scales, and that allowed it to generate plausible entirely novel samples from that visual world.

One of our central questions was the extent to which the model transforms raw image space into a feature space within which different physical causes are disentangled (12,14,65,66). Surfaces with similar reflectance properties may occupy very disparate points in raw image space, but should cluster together in the feature space of a good perceptual model. To assess this, we have to use our knowledge, as experimenters, of the ground truth properties of scenes. Here we do this (1) by colour coding visualizations (**Figure 2A**); (2) through representational similarity analysis (67) (**Figure 2D**); and (3) by using linear classifiers to measure separability (**Figure 3**). Although this evaluation step involves label-based decoding, it merely quantifies the placements of stimuli relative to one another in terms of the physical properties of interest, and is very different from using supervised training to arrive at the feature space in the first place. That this linear decoding step is not sufficient to impose knowledge about physical properties is amply demonstrated by the inability to decode scene properties from simpler feature spaces such as pixels, texture descriptions, or indeed from sophisticated supervised DNN models (see **Figure 2B-C** and **E**, **Figure 7A** and **Figure S4C**).

Because the model receives no explicit physical labels, its resulting internal representation is likely to only approximately disentangle physical parameters. We propose that the imperfections in unsupervised disentanglement might explain why human gloss judgments are erroneously affected by other physical properties like lighting and surface relief (15,16,31). We were able to test how well the model could predict human gloss judgements, by deriving ‘gloss estimates’ through linear decoding. We found that the unsupervised PixelVAE networks learned rich and flexible latent codes from which both human perception *and misperception* of specular reflectance could be predicted on an image-by-image basis.

Results

After training, all ten instances of the unsupervised PixelVAE model could generate wholly novel images that look like plausible surfaces (**Figure 1D**). The networks generate images pixel-by-pixel from the top left corner by probabilistically sampling from a learned distribution, conditioned both on previous pixels and values in the model's 10-dimensional latent code (**Figure 1C**). The representations in this latent code are a highly compressed representation of whole-image properties, and are the focus of all subsequent analyses.

As a comparison model, we also trained ten instances of a supervised ResNet (68) network to classify the same images as high or low gloss, using ground truth (high or low specular reflectance in the rendering parameters) as training labels. Its mean classification accuracy was 99.4% +/-SD 0.001. The supervised network also included a 10-dimensional fully-connected layer prior to its two-unit output classification layer, which we treated as its high-level latent code for subsequent comparisons with the unsupervised model (**Figure 1C**).

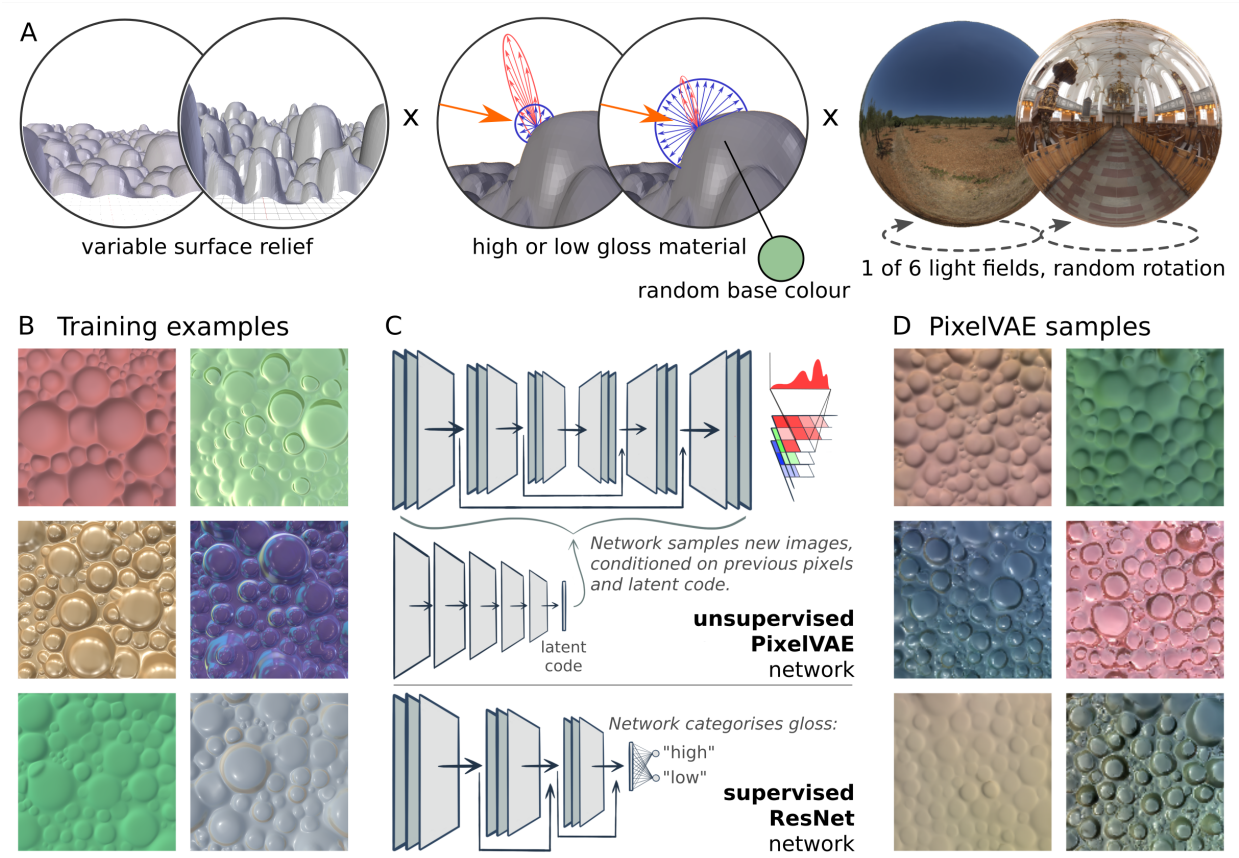


Figure 1: An unsupervised neural network learns to generate plausible novel images from a simulated visual world.
(A) The simulated physical world consisted of a bumpy surface viewed from above. The scene varied in depth of surface relief (left), surface base colour and whether the surface material had high or low gloss (middle), and illumination environment (right).
(B) Six examples from a training set of 10,000 images rendered from the simulated world by randomly varying the world factors.
(C) The dataset was used to train an unsupervised PixelVAE network (upper panel). The network learns a probability distribution capturing image structure in the training dataset, which can be sampled from one pixel at a time to create novel images with similar structures. A sub-stream of the network learns a highly compressed 10-dimensional latent code capturing whole-image attributes, which is used to condition the sampling of pixels. The same dataset was used to train a supervised ResNet network, which output a classification of “high gloss” or “low gloss” for each input image. The penultimate layer of the supervised network is a 10-dimensional fully-connected layer, creating a representation of equivalent dimensionality for comparison to the unsupervised model’s latent code.
(D) Six example images created by sampling pixels from a trained PixelVAE network, conditioning on different latent code values to achieve diverse surface appearances. These images are not reconstructions of any specific images from the training set, but are completely novel samples from the probability distribution the network has learned.

An unsupervised generative model disentangles distal scene properties

The unsupervised PixelVAE model's training objective deals only with proximal image data. Yet we found that after training on the rendered dataset, distal scene properties—such as gloss and lighting—spontaneously clustered within the networks' 10-dimensional latent codes (cf. (41,68,69)). Visualising in two dimensions using tSNE (70) strikingly reveals that low-gloss images cluster together, while high-gloss images form multiple tight clusters, corresponding to different light fields (**Figure 2A**). Within each light-field cluster, images are arranged by the angle of illumination, as well as by surface relief, with flatter surfaces occupying nearby points, and bumpier surfaces more distant points. This shows that without explicit labels, the unsupervised model learns to at least partially reorganise stimuli by their physical properties, one of the core challenges of mid-level vision.

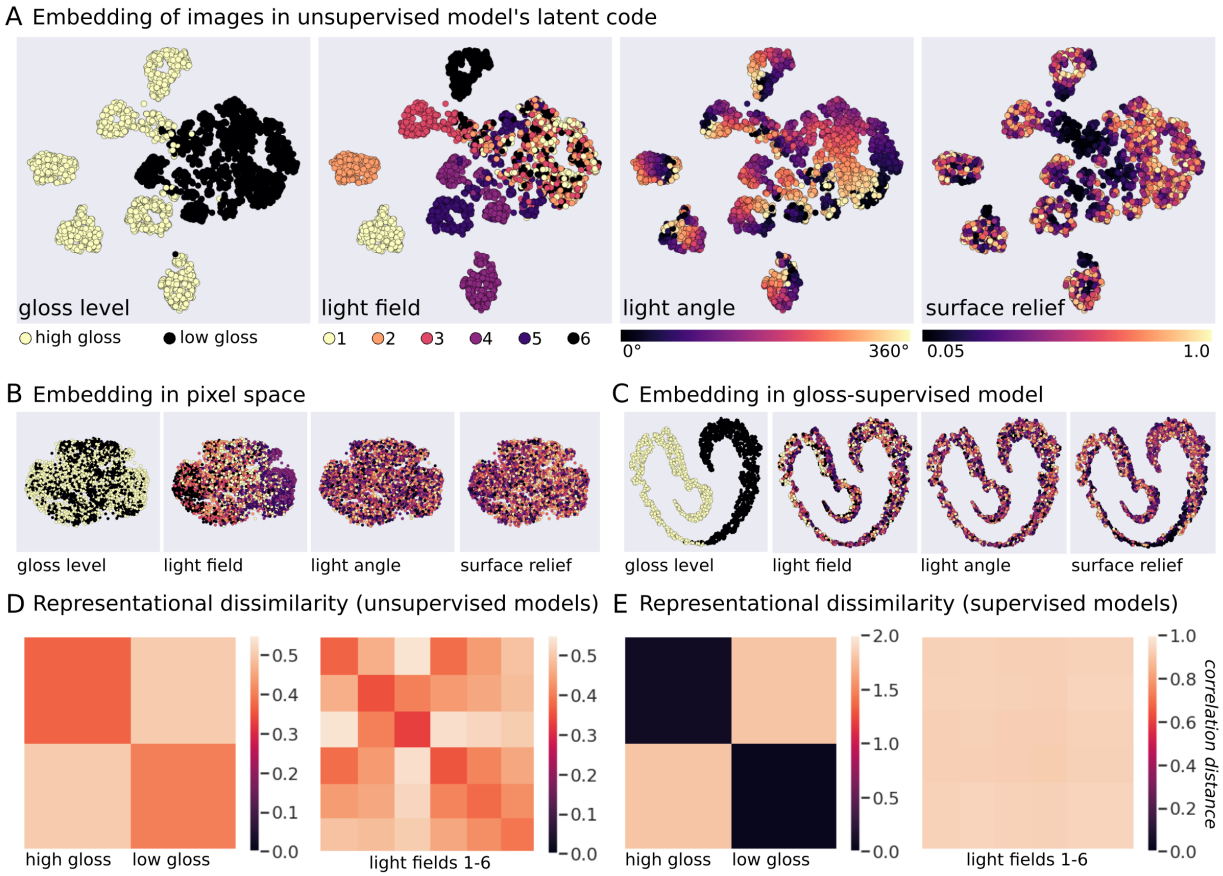


Figure 2: World factors are disentangled in the unsupervised model's latent code.

(A) Visualisations of distances between 4,000 images, in the 10D latent code of one unsupervised PixelVAE network, projected into two dimensions using tSNE. Unsupervised learning spontaneously disentangles the underlying world factors, arriving at a nested representation. Images of low-gloss surfaces form one large cluster while images of high-gloss surfaces form multiple small subclusters (left panel), according to the light field used to render the image (centre left panel). Lighting direction varies smoothly within each subcluster (centre right panel), and surfaces with low relief are closer to one another than are those with high relief (right panel).

(B) tSNE visualisation of the same images in raw pixel space, where world factors are thoroughly entangled.

(C) tSNE visualisation of the same images in the 10D latent code of one gloss-supervised ResNet network, showing a less rich representation. High- and low-gloss images form two clearly-separated clusters, but other world factors are entangled.

(D) Dissimilarity matrices showing correlation distance (1-*r*) between all pairs of images in the latent representations of all PixelVAE networks, averaged across surfaces with the same (left) gloss level, or (right) illumination. Pairs of images belonging to the same gloss/lighting condition (diagonal blocks in each matrix) were represented more similarly than pairs of images belonging to different conditions (off-diagonal blocks).

(E) Corresponding dissimilarity matrices calculated from the representations in the 10D latent codes of the supervised networks reveals clustering only by the task-relevant dimension of gloss.

That this rich clustering of images by scene properties emerged in the model is far from trivial. It was not caused by raw image similarities, since tSNE visualisation of the same images in raw pixel space showed a tight entangling of scene properties (**Figure 2B**). Other linear and non-linear pixel embeddings such as MDS and LLE (71) also failed to separate low- from high-gloss surfaces (**Supplementary Figure 4C**). Similarly, when we, applied the same visualisation to the 10D layer of the gloss-supervised models, we observed a neat separation of high and low gloss images, but the other world factors were intermixed (**Figure 2C**). Similar qualitative patterns held for all ten instances of both unsupervised and supervised models.

To quantify these clustering effects, we used representational similarity analysis (67) (**Figure 2D**). The results support the tSNE visualisations. Pairs of images belonging to the same gloss condition (both high-gloss or both low-gloss), corresponded to closer points in the unsupervised models' 10D latent codes than pairs of images belonging to different gloss conditions (two-tailed repeated-measures t-test comparing average distances between same- vs different-gloss image pairs, across network training instances: $t_9 = 11.09$, $p = 9\text{x}10^{-10}$). Likewise, pairs of images illuminated by the same light field had more similar latent representations than those lit by different light fields ($t_9 = 19.7$, $p = 1\text{x}10^{-27}$). However, in the supervised models, only clustering by gloss was apparent (**Figure 2E;** gloss $t_9 = 112.2$, $p = 3\text{x}10^{-28}$; whereas lighting $t_9 = 1.77$, $p = 0.08$). Thus, while the supervised model optimizes disentanglement of the single physical property on which it is trained, the unsupervised model spontaneously discovers multiple latent variables and seeks to capture the extent to which they each contribute to observed image structure.

The unsupervised model predicts human gloss perception for novel images

Our central question was how well the spontaneous separation of high and low gloss images in the unsupervised model corresponded with human gloss judgments. To derive quantitative gloss predictions from the models, we next trained a linear support vector machine (SVM) classifier to find the hyperplane in the 10D latent code of each network that best separates high from low specular reflectance images (**Figure 3A**). In

neuroscience, information that is available directly via a linear readout from units or neurons is generally considered to be explicitly represented by a model or brain region (e.g. (65,72–76)). Importantly, this is simply a formal way of quantifying the degree and form of the disentanglement. It does not provide the model with any new information, but merely measures the relative placement of images in the existing feature space.

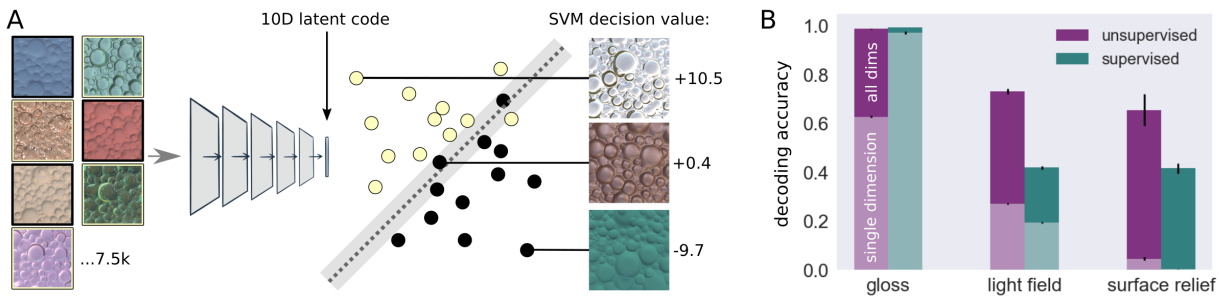


Figure 3: Disentangled world factors are available via a simple linear readout.
(A) To quantify how explicitly gloss was represented in each model, we trained linear SVM discriminants to classify gloss from the representations of images within the latent code of each network. As well as obtaining decoding accuracy, this allowed us to use decision values (i.e. distance from SVM hyperplane) to make continuously-valued predictions from the model regarding the gloss of novel images (three examples shown), which was critical in subsequent experiments.
(B) Average decoding accuracy for each world factor from latent codes of unsupervised and supervised networks (y-axis: proportion correct for two-way gloss classification and six-way light field classification, and R^2 for surface relief regression). Error bars: standard error of the mean over ten training instances of each model type. Pale lower regions of bars indicate average decoding accuracy when using the best single latent dimension, and give an impression of how distributed the representations are.

Based on this linear decoding, we find that gloss-classification accuracy for novel renderings across the ten unsupervised models was extremely good at 99.3% (+/-SD 0.002)—practically as good as decoding gloss from 10D latent code of the supervised models (99.4% +/- 0.002; **Figure 3B**). Light field and surface relief could also be decoded well above chance from the unsupervised networks (**Figure 3B**), and significantly better than from the supervised networks (two-tailed independent-measures t-test comparing light field decoding between unsupervised and supervised models: $t_9 = 23.25$, $p = 7\text{x}10^{-15}$; surface relief: $t_9 = 3.30$, $p = 0.004$). Thus, like the tSNE visualisations and representational similarity analysis, linear decoding demonstrates

that the unsupervised networks learn a compact representation that summarises information about not only surface material, but other holistic scene properties such as illumination and surface relief. The analysis also revealed that representations were distributed rather than sparse. The full latent code predicted scene properties much better than any individual dimension could (**Figure 3B** and **Supplementary Figure 1C**).

Crucially, through linear decoding, we could now derive a continuous predicted gloss value for any image by inputting it to a network and calculating the SVM decision value for its corresponding point in latent space (i.e., signed distance of point from network's gloss-separating hyperplane; **Figure 3A**). This allowed us to compare the model against human gloss perception.

For Experiment 1 we rendered 50 new images of surfaces with random magnitudes of specular reflectance, sampled uniformly from almost completely matte to almost mirror-like. Twenty observers rated the apparent gloss of each surface, from 1 (matte) to 6 (glossy). We compared their ratings to the gloss values predicted by the unsupervised model. **Figure 4A** shows that agreement was excellent (r = 0.96, p = $7\text{x}10^{-29}$), and substantially better than for the supervised model (r = 0.70). Notably, the unsupervised model even predicted human ratings better than ground truth (specular magnitude within the rendering engine; two-tailed repeated-measures t-test t_{49} = 8.17, p = $1\text{x}10^{-7}$).

We also considered a number of alternative models (**Figure 4A** bargraph), the best of which was a feature space consisting of the 1,000 final-layer features from a ResNet DNN trained on 1.2 million images to classify objects (52,68). This is consistent with previous findings that high-level representations in object-recognition DNNs capture perceptually-relevant features of textures and images (55,77–79). Other, less well-performing, models included a multi-scale texture description comprising 1,350 feature dimensions (80); the 4,096 latent features from a relatively simple image autoencoder; 10-dimensional embeddings of raw images via tSNE (70), MDS or LLE (71); and luminance histogram skewness—a measure previously proposed to predict human gloss perception (81) (see **Methods**).

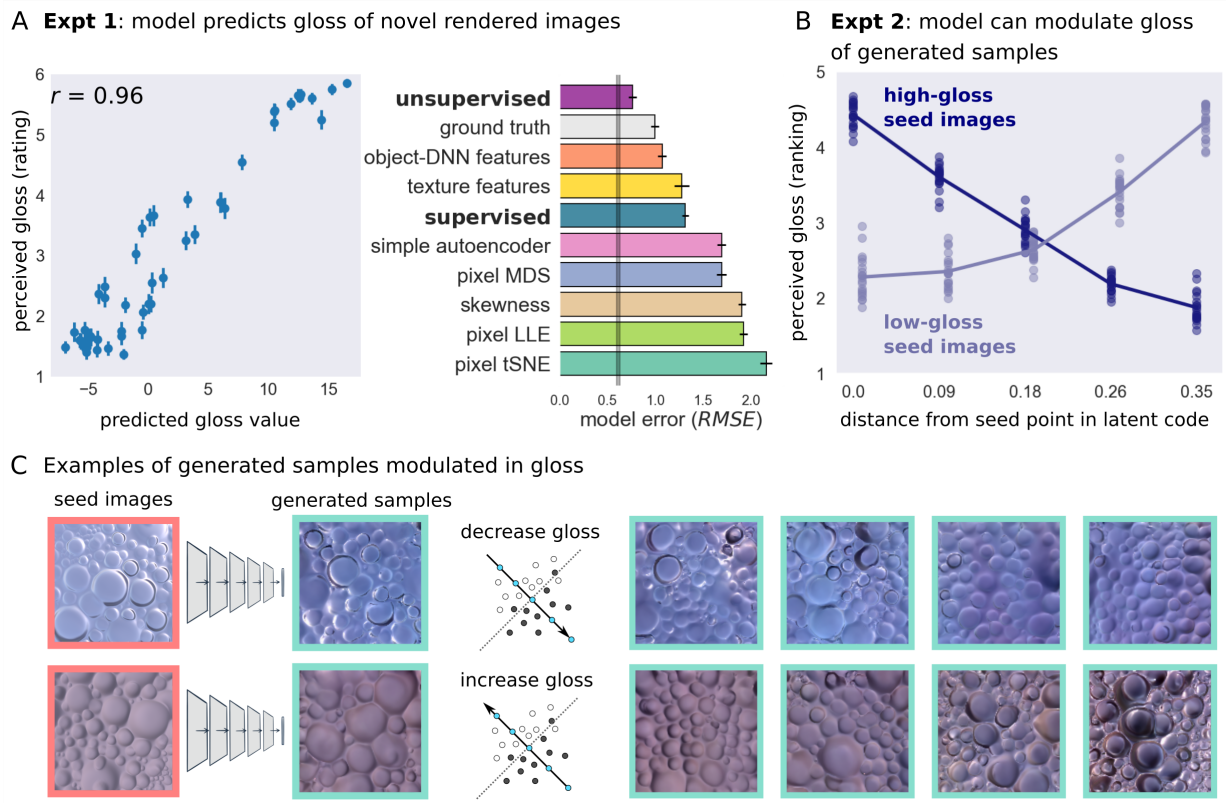


Figure 4: The unsupervised model can be used to modulate the gloss of generated samples, and predicts human-perceived gloss of new surfaces.
(A) Gloss values in the unsupervised networks well predict human gloss ratings for 50 novel rendered surfaces of random gloss levels. Scatterplot shows decision value on a glossy-vs-matte SVM classifier for each image in the PixelVAE networks on the x-axis (averaged over ten training instances), and human gloss ratings on the y-axis. Barplot shows average error in predicting each individual participant's ratings, after normalising model predictions into the range 1-6, for the unsupervised model and for diverse alternative predictors (see text and **Methods** for details). Vertical grey line indicates how well individual human ratings can be predicted from data of other participants, giving the minimum possible model error. Error bars: SEM over 20 participants.
(B) Moving along the gloss-discriminating axis in a network's latent code as shown in (C) successfully modulates perceived gloss, both when adding gloss to initially matte surfaces (pale blue line), and when removing it from initially glossy surfaces (dark blue line). Dots show each participant's gloss rankings of each of five modulation steps, averaged over four test sequences starting from low-gloss initial seed images, and four starting from high-gloss seed images. Lines indicate mean over 20 participants.
(C) Example gloss-modulated sequences. A glossy or matte rendered image (left) is input to a network and its 10D latent representation recorded. New images are generated from the model, conditioned first on the original ("seed") point in latent space, and after stepping by small amounts along the model's gloss-discriminating axis (i.e. the axis orthogonal to the decision plane of a glossy-vs-matte SVM classifier), either towards the "matte" (top) or "glossy" (bottom) directions.

Since PixelVAE networks are generative models, we can create novel images by sampling from them (**Figure 1D**). In Experiment 2, we used such images to test whether perceived gloss varied systematically with the location in the model's latent space. From each of the model training instances, we generated four sequences of five images by starting at "low gloss" points in the latent space and moving in the higher-gloss direction (i.e., along the axis orthogonal to the gloss-separating hyperplane). We also created another four sequences moving from high to lower gloss points (**Figure 4C**). The same 20 observers sorted the images within each sequence from least to most glossy. **Figure 4B** shows that moving along the model's 'gloss discriminating axis' systematically reduced the apparent gloss of generated images when moving in the matte direction (one-way repeated-measures ANOVA $F_{4,19}$ = 244.11, p = 2x10^{-42}) or increased it when moving in the high-gloss direction ($F_{4,19}$ = 649.82, p = 6x10^{-58}).

Thus, despite never being given information about distal properties during training, we find the unsupervised networks develop internal representations that not only disentangle distal causes that are impossible to tease apart in the raw input (**Figure 2B**, **Figure 6A** and **Supplementary Figure 4C**), but also, more remarkably, predict human gloss perception better than the true physical reflectance properties of surfaces.

The unsupervised model predicts failures of human gloss constancy

Importantly, while human gloss perception generally aligns well with surface specular reflectance, it also exhibits some well-documented 'errors'. For example, bumpier surfaces tend to look glossier than flatter surfaces, and specific combinations of lighting and surface relief yield specific patterns of misperception (**Figure 5A**; (15,16)). A key test of any computational model of biological vision is that it mimics perceptual errors as well as successes (82). We therefore used systematic failures of gloss constancy to decouple human perception from true scene properties, and challenge competing models.

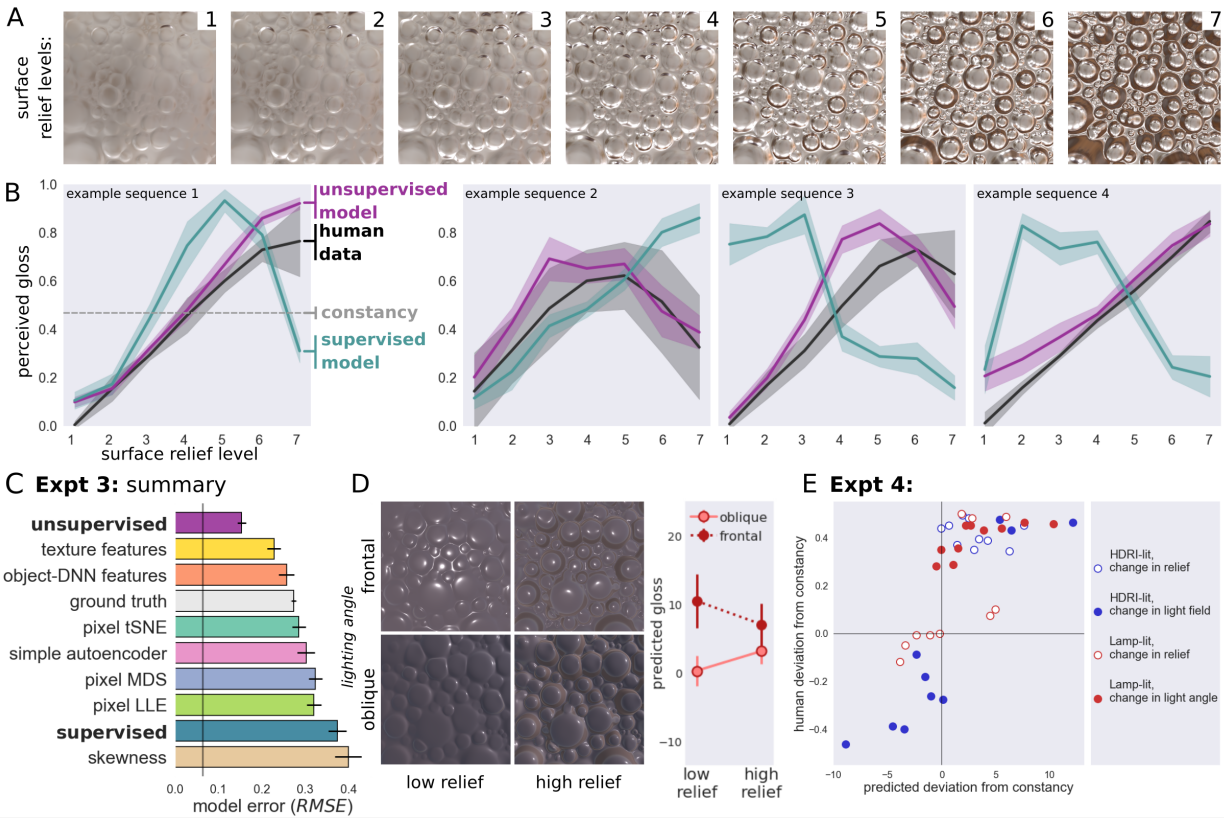


Figure 5: Unsupervised learning predicts human perception and misperception of gloss on an image-by-image basis.

(A) Material properties of surface are identical in all seven images, but surface relief increases. People often see bumpier surfaces as glossier than flatter ones, constituting a failure of gloss constancy (15,16).

(B) Examples of various patterns of human gloss constancy failure for four sequences of images increasing in surface relief. Leftmost panel shows data for sequence in (A). Black: human data (proportion of times each image step was reported as being glossier than others within each sequence); Purple: prediction of unsupervised model; Teal: prediction of supervised model. Veridical gloss perception would be flat horizontal line. Shading indicates standard deviation over 20 human participants or 10 model instances. See **Supplementary Figure 2** for full data.

(C) Summary of average error of each model's prediction of each participant's data, averaged across all twenty image sequences. Conventions are as in Figure 4B.

(D) Changes in lighting also often change apparent material properties for human observers. The quartet of images shows the same surface material rendered on a low- vs high-relief surface (left vs right columns), and illuminated straight-on vs at a 30-degree angle (top vs bottom rows). Right: unsupervised model predictions for these four images (error bars: standard deviations over training instances). The model predicts a larger difference in apparent gloss for the low- than high-relief images, and an interaction between changes in lighting and relief.

(E) Magnitude of deviations from constancy for 40 pairs of images depicting the same surface material with different scene properties (proportion of times image A of a pair was seen as glossier than image B). Scenes were illuminated by either natural lighting (blue dots) or directional lamp (red dots). Image pairs differed in surface relief (unfilled

dots), light field (blue filled dots), or lighting angle (red filled dots). Y-axis: raw proportion minus 0.5, so that zero indicates good constancy; +0.5 indicates image A always ranked glossier than image B; -0.5 indicates opposite. Data are data averaged over 20 observers. X-axis: difference in predicted gloss values between the images in each pair, averaged over 10 training instances of the PixelVAE model.

To do this, we rendered sequences in which reflectance and other scene properties were fixed, but surface relief increased in seven steps (**Figure 5A**). It is important to note that although the supervised model is near-perfect at coarsely categorising surfaces as having high or low specular reflectance, it still predicts different *degrees* of glossiness for different images, sometimes including failures of constancy when different images depict the same surface material with different geometry. It is the detailed patterns of these failures of constancy which we are interested in comparing between the unsupervised and supervised models. For stimuli in Experiments 3a and 3b we selected two sets of 10 sequences for which (a) both the unsupervised and supervised models predicted deviations from constant gloss, and (b) the models made *different* predictions about the particular pattern of deviations (see **Methods**). The rationale behind this is that 'hard cases' in which models disagree provide the strongest power to test which is the superior model (83,84).

With these image sequences in hand, in Experiments 3a and 3b, two groups of 20 observers provided gloss judgements via a paired-comparison task (see **Methods**). If observers correctly estimated reflectance, all surfaces should appear equally glossy, yet we find that they do not. Observers showed strong failures of gloss constancy, usually reporting surfaces with deeper relief to be glossier, although seven of the twenty sequences showed non-monotonicities, with perceived gloss highest for intermediate reliefs (four examples shown in **Figure 5B;** complete data: **Supplementary Figure 2**). The unsupervised model, despite never being explicitly trained to represent gloss, and without being fit to human data, predicted patterns of failures of gloss constancy remarkably well (**Figure 5C**, median R^2 = 0.77). The model correctly predicted the qualitative pattern of constancy failure (monotonic vs non-monotonic) for 18 out of the 20 stimulus sequences. The supervised model completely failed to predict human gloss constancy. For almost all sequences it made predictions that were anti-correlated with human patterns (median R^2 = -1.36). Out of the

alternative models, a gloss classifier based on texture features (78) provided the next best performance to the unsupervised model (median R^2 = 0.54), but was significantly poorer (two-tailed repeated-measures t-test t_{19} = 5.66, p = 4x10^{-6}).

Human gloss constancy ranges from good to bad depending on interactions between lighting and shape. We next tested whether unsupervised learning could account for this range of behaviours by rendering 800 candidate image pairs depicting surfaces with identical material, but differing in surface relief or lighting (**Figure 5D**; **Methods**). From these, we selected 40 image pairs for which the unsupervised model predicted a wide range of degrees of gloss constancy, from excellent (near-identical predicted gloss for both images within a pair) to very poor (images received very different predicted gloss values; see **Methods**).

Twenty observers indicated which surface in each pair appeared glossier, with each pair repeated eight times. For these images with diverse scene properties, the unsupervised model predicted the degree and direction of human (failures of) constancy reasonably well (R^2 = 0.55, p < 5x10^{-8}, **Figure 5E**) and outperformed all alternative models. Fitting a simple logistic function to relate model and human gloss values further improves the prediction (R^2 = 0.73). Together, these findings show that the arrangement of stimuli in the unsupervised model's latent space predicted both the broad successes and the specific patterns of errors of human gloss perception.

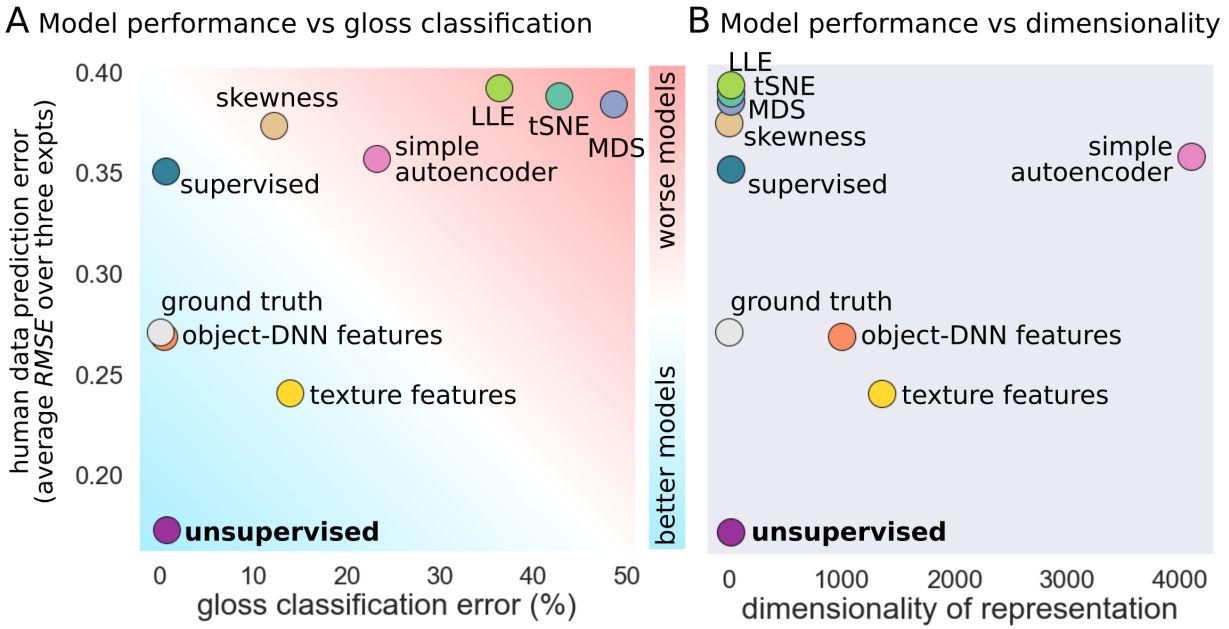


Figure 6: The unsupervised model outperforms diverse alternative models.
(A) Relationship between error at classifying surfaces as glossy or matte (x-axis) and error in predicting human psychophysical gloss judgements (y-axis) for all models. Data are averaged over the three psychophysical experiments in which the ground truth gloss of stimuli is defined (Experiments 1, 3 and 4).
(B) Relationship between the dimensionality of each model's representation, and its error in predicting human judgements (y-axis shared with A). Not only does the unsupervised PixelVAE model best explain human data, but it does so with around two orders of magnitude fewer dimensions than the nearest competing models (a 1,350-dimensional texture feature space, and a 1,000-dimensional layer from an object-recognition supervised DNN).

Aggregating results across the three experiments in which people made gloss judgements about rendered images for which we had ground-truth specular reflectance (Experiments 1, 3, and 4), the unsupervised model outperforms all others (**Figure 6**). It achieves near-perfect ground-truth gloss-classification accuracy, while still predicting the idiosyncratic errors of human gloss perception (**Figure 6A**), and achieves this with a feature space two orders of magnitude more compact than the next best model (**Figure 6B**). Overall, the next best model was a set of texture features hand-engineered to efficiently capture higher-order statistical structure in images (80) However, the unsupervised PixelVAE networks were able to discover superior feature spaces, both for classifying gloss and for predicting human gloss perception, by summarising image structure arising from surfaces with varying gloss levels. Simple image statistics, such as the skewness of the luminance histogram (81) failed to

capture human gloss perception under our deliberately challenging stimulus conditions, where differences in specular reflectance must be de-confounded from differences in lighting and surface shape.

Effects of training set and model generalization

The composition of the training dataset has profound effects on all machine learning models. However, a good model of human mid-level vision should make predictions that generalise to some degree over changes to the training or test sets. Our unsupervised networks were trained on a simulated environment with a bimodal distribution of glossiness (near-matte or high-gloss), but we found that they could nevertheless predict gloss in new datasets containing continuously-varying gloss, as well as generalising across large changes in scene properties such as lighting and surface shape. We established this by creating four new surface geometries, and rendering 500 images from each, as well as the original geometry, using a new set of light fields never seen during training, and sampling gloss level continuously rather than bimodally (**Figure 7A**). Using linear regression, we could well predict the glossiness in the continuously-varying images (mean R^2 = 0.79 +/- SD 0.03), and predictions for the four new geometries were at least as good as for the familiar geometry (**Figure 7A**). tSNE visualisation of the 2,500 images' embedding in latent space indicated a smooth organisation by specular reflectance (**Supplementary Figure 4B**). Remarkably, given their constrained training environments, the models even seem able to categorise close-up photographs of diverse real-world surfaces as being of high or low gloss (**Supplementary Figure 4A**).

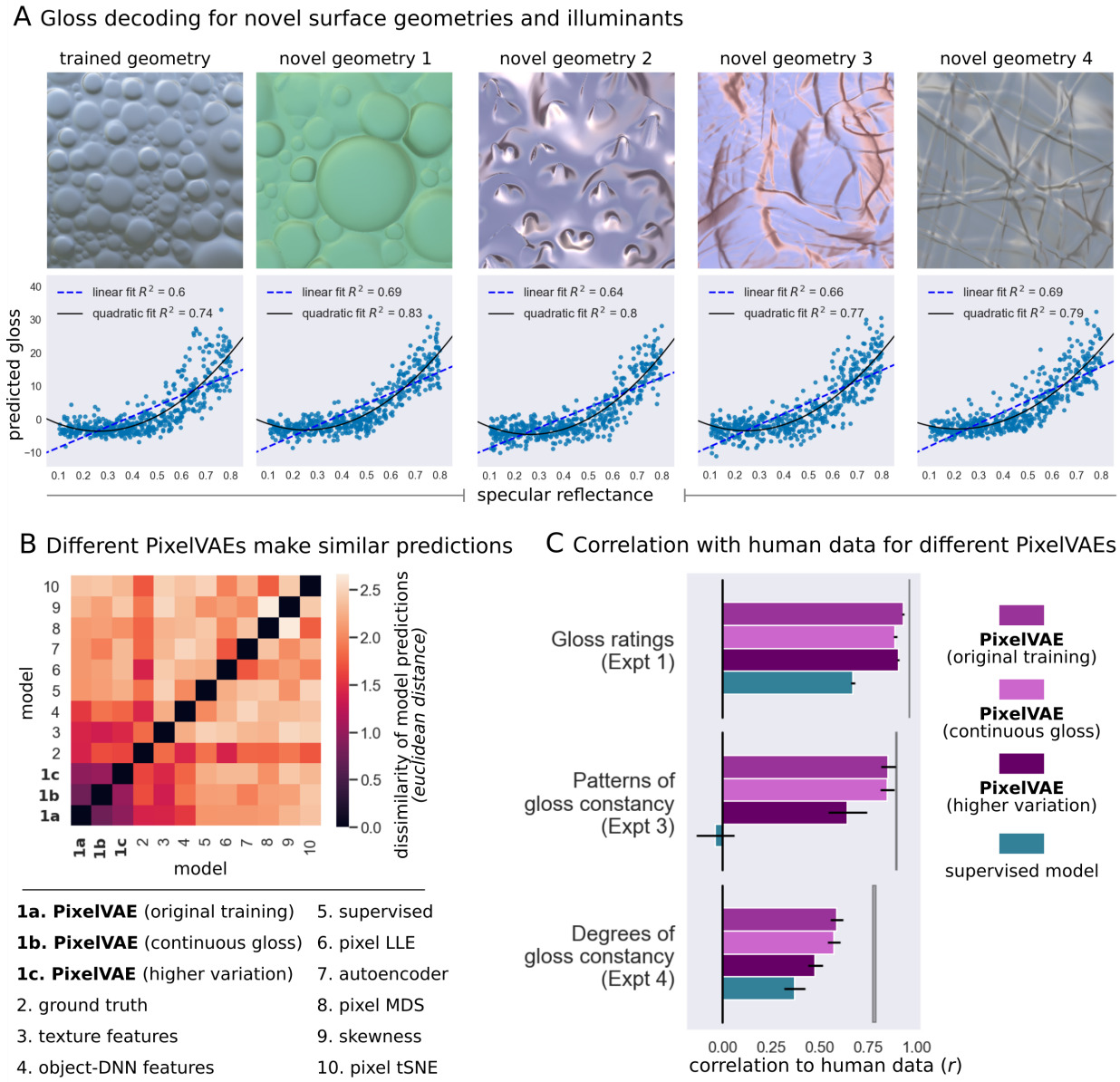


Figure 7: The unsupervised model generalises well to novel scenes, and makes similar gloss predictions after retraining on different visual environments.
(A) Scatterplots show predicted gloss (x-axis) and ground-truth specular reflectance (y-axis) for sets of 500 new images rendered with random continuously-sampled gloss levels, using the trained surface geometry (leftmost panel) or four novel geometries unseen during training (right panels). A single example image from each set is shown above plots. All scenes were illuminated by light fields never seen during training.
(B) New instances of the PixelVAE model with substantially different visual experience during training all make similar predictions for the gloss levels of experimental stimuli. Matrix shows the dissimilarity between the gloss predictions made by (1a) the original set of ten PixelVAE networks, averaged over training instances; (1b) five PixelVAE networks trained on a new set of 10,000 images using the same surface geometry and light fields, but with specular reflectance continuously sampled, averaged over training instances; (1c) a PixelVAE network trained on a third set of 10,000 images using ten

surface geometries and 50 light fields, non-overlapping with the geometry and light fields used in the main training dataset; and (2-10) alternative models, ordered by the similarity of their predictions to that of the original PixelVAE. Gloss predictions are for all stimuli used in Experiments 1, 3 and 4, and were were normalised into the range 0-1 for comparability across models.
(C) Average correlation between model-predicted gloss and individual human judgements in each psychophysical experiment for PixelVAE implementations described in (B). Performance of the supervised model is included for reference. Vertical grey lines indicate how well data can be predicted from other participants, giving the highest possible model correlation. Error bars show standard error of the mean across participants.

To test the robustness of our model to different training datasets, we trained five new PixelVAE networks on 10,000 additional renderings in which gloss was sampled continuously rather than bimodally ("continuously-sampled gloss training dataset"; **Methods**). Next, we trained an additional PixelVAE on a third dataset of 10,000 renderings in which surface geometry and lighting varied far more widely ("higher-variation training dataset"; **Methods**), with each scene comprising one of 10 novel surface geometries combined with one of 50 novel light fields. Both new training environments produced models with latent codes that could well subserve gloss classification on the original bimodal gloss dataset (mean accuracy = 96.7% and 91.4%, respectively). Importantly, we found that the three versions of the PixelVAE model (original, continuous-gloss training, and higher-variation training) made highly similar predictions regarding the relative gloss levels of experimental stimuli. This indicates that the ability to predict human perception is not an artefact of the training set. To quantify this, we calculated the dissimilarity between the gloss predictions made by each pair of models for all rendered experimental images across Experiments 1, 3 and 4 (**Figure 7B**). The gloss predictions made by the three versions of the unsupervised PixelVAE model trained on non-overlapping datasets were more similar to one another than they were to those made by any of the ten diverse alternative models (dark cluster of low dissimilarity values in the bottom left of **Figure 7B**). All three training environments led to unsupervised models that reasonably well predicted human data (**Figure 7C**), both for gloss ratings of novel rendered images (Experiment 1), and for the more challenging task of predicting patterns (Experiment 3) and degrees (Experiment 4) of (failures of) gloss constancy. All three unsupervised model versions predicted human data significantly better than the supervised model in all experiments

(ps = 1×10^{-17}–2×10^{-3} across nine repeated-measures t-tests), and were significantly better than the most promising alternative model, texture features (80), in eight out of nine comparisons (ps < 5×10^{-3}, bonferroni corrected). Unsupervised learning in the PixelVAE models therefore appears to converge on a representation that captures key aspects of human gloss perception, and is robust to changes in the particular statistics, illuminations, or geometries of the training set.

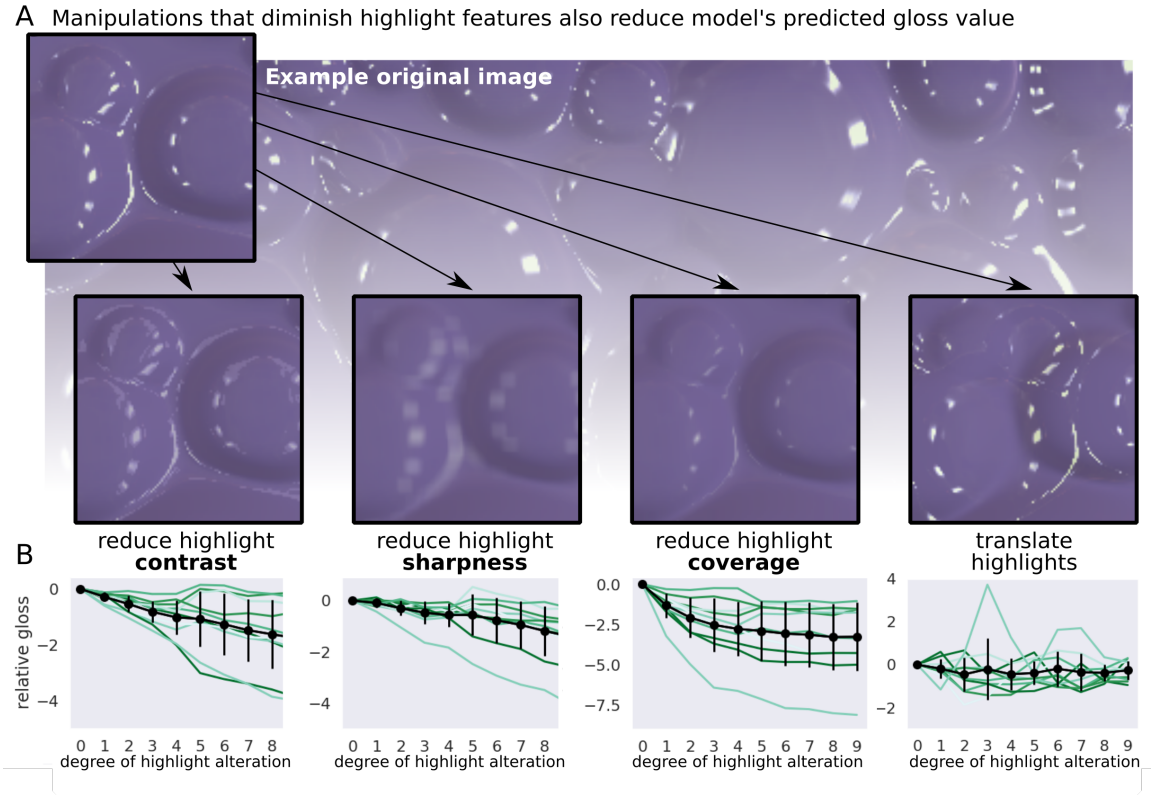


Figure 8: Reducing the coverage, sharpness or contrast of highlights reduces model-predicted gloss.
(A) We separately rendered the specular and matte components for ten probe images, and manipulated the specular component by either reducing contrast, blurring, shrinking coverage via an image erosion algorithm, or laterally translating it, in ten progressive steps. We then recombined the two components to produce ten sequences of ten full images, and presented them to the unsupervised networks. Insets show each of these manipulations applied to part of an example image.
(B) Model-predicted gloss declined when any of the three highlight features was reduced (left three plots). Green lines show the predicted gloss for each step in each sequence, relative to that of the original un-manipulated image, averaged over training instances of the model; black lines show mean and standard deviation over all sequences. Translating the specular components (rightmost plot), which does not affect the coverage, sharpness, or contrast of highlights, did not affect predicted gloss.

Features underlying gloss representation in the model

Previous research (16) identified specific image features—the coverage, contrast, and sharpness of specular highlights—that predicted perceived gloss for surfaces like ours. To test whether the PixelVAE model was also sensitive to these cues, we measured the coverage, contrast, and sharpness of highlights in 10,000 new renderings of surfaces with specular reflectance varying continuously from near-matte to near-mirror. All three cues could be decoded from the latent code of a PixelVAE trained on these

images (mean R^2 = 0.71), and could be increasingly well decoded from successive convolutional and fully connected layers (**Supplementary Figure 3A**). A linear combination of the three cues in the layer immediately preceding the 10D latent code correlated with gloss predicted from the latent code (*r* = 0.80; **Supplementary Figure 3B**). Moreover, manipulating the highlights in images to weaken each cue also reduced predicted gloss (**Figure 8A-B**).

Interestingly, the model's predicted gloss did *not* decrease when the specular component of images was shifted so that highlights were misaligned with the geometry of the depicted surfaces **(Figure 8B)**. This manipulation preserves coverage, contrast and sharpness, yet infringes photogeometric constraints that are a precondition for humans to identify highlights as highlights, and therefore to see gloss at all (6,20,21). The fact that the model predicts human gloss constancy despite not being sensitive to such photogeometric constraints suggests that although these constraints are crucial to many aspects of surface perception (7,21,85–87), the *degree* of perceived gloss, within surfaces that are seen as having highlights, may be largely explainable in terms of image features.

Discussion

Efficient representation of sensory data has long been hypothesised as a central goal of neural coding (33,34,36,88,89). Yet, while such approaches predict many aspects of sensory coding, to date, they have not explained how we visually infer properties of the outside world. Unsupervised learning objectives in modern DNNs, such as data compression and spatial prediction, offer powerful new implementations of these statistical learning principles. Our findings show for the first time that mid-level perceptual dimensions, like gloss—which appear to (imperfectly) estimate properties of the physical world—can emerge spontaneously by learning to efficiently encode sensory data. Thus, unsupervised deep learning provides a bridge between the view that mid-level vision primarily extracts summary statistics (e.g. (90)), and the view that it achieves an explicit decomposition of images into distinct physical causes (4). Our results suggest that rather than framing perception as the optimal *estimation* of

physical quantities, it may be more usefully considered as a data-driven process of *disentanglement* in multidimensional feature spaces that capture regularities in sensory data (12,14,33–35,65,88,89,91,92).

Does the PixelVAE model “understand” gloss, the way human vision does? Its most obvious failure is an insensitivity to certain photogeometric constraints known to affect human surface perception, such as the alignment of specular highlights with diffuse shading (7,20,21) (**Figure 8**). Without a detailed representation of 3D shape, the model certainly cannot explain all aspects of human gloss perception (85,86,93). Yet, conversely, accounts that focus on highlight-shading consistency cannot explain variations in perceived glossiness when highlights do align with shading—the conditions we consider here. This includes both expected changes in gloss when specular reflectance varies (23–25,27,29,30), and illusory changes as a function of lighting and shape (e.g. 7,41). Indeed, to the best of our knowledge, no extant accounts of gloss constancy and its failures (e.g. 44) take photogeometric constraints into consideration. It is striking that a model with visual experience limited only to static fronto-parallel views of surfaces already well predicts degrees of apparent gloss across a wide range of images, including cases when specular reflectance is held constant. An important direction for future research is testing whether unsupervised DNNs can also learn photogeometric relationships if training sets provide additional information about shape (e.g. through motion, stereo, occlusion, or larger variations in geometry). There are some grounds for optimism. For example, partially view-invariant representations of 3D objects have been shown to emerge through unsupervised learning on movies, via Hebbian learning (94), competitive neural networks (95), or slow feature analysis (96).

In using deep learning models, we do not wish to imply that all material perception is learned during an individual lifetime. Unsupervised learning principles can also operate on an evolutionary timescale. For example, V1 cell receptive fields are predicted by simple unsupervised learning models such as independent components analysis (36) and these seem to be present at birth in macaques (97). There is evidence that 5-8 month old infants can distinguish between matte and specular surfaces (98), but also that material recognition and classification are still developing in 5-10 year old children

(99,100). Even 5-8 months of visual experience provides a huge dataset for statistical regularity learning (101). It could be that approximate versions of many perceptual dimensions are rapidly learned during the first days, weeks and months of life.

Although we do not propose the PixelVAE architecture as a physiological simulation of visual cortex, its computational principles are eminently achievable in biological brains. Data compression and spatial prediction are learning objectives that require no additional information beyond the incoming sensory data, and there are several mechanisms by which brains could represent the probability distributions used within the PixelVAE network (102–104). At the same time, the brain certainly does not learn statistical distributions over RGB pixels. If the visual system uses learning objectives like the ones investigated here, they presumably operate on representations of images that have undergone substantial optical and retinal processing (105,106).

In conclusion, unsupervised DNN models provide an ecologically feasible solution to the problem of how brains seem to represent properties of the distal world without access to ground truth training data (5,12–14,91,92). Non-linear transformations reorganise inputs according to high-order regularities within and across images, allowing the visual system to better summarise and predict sensory data. Yet, because regularities in images are caused by underlying physical objects and processes, these new configurations often end up (partially) disentangling physical properties from one another. It is the imperfect nature of the disentanglement that yields the characteristic ‘errors’ humans make. Failures of constancy, which are rife in vision, therefore offer tell-tale clues to how we learn to see. Unsupervised learning may account for them, not just in gloss perception but in perception more broadly.

Methods

Participants

Three groups of 20 human naïve observers reported perceived gloss across five experiments (mean age 23.7). All participants had normal or corrected-to-normal visual acuity. Two male participants self-reported poor red-green colour vision. Six individuals participated in two different experimental groups, but received no information about the experimental design or hypothesis after the first session. Experiments were conducted in accordance with the Declaration of Helsinki, with prior approval from the local ethics committee. Volunteers were paid €8 per hour.

Stimuli

Stimuli were 800×800pixel images of bumpy surfaces rendered using Unity3D (version 2018.2.3). A virtual 40×40cm sheet with irregularly positioned bumps was placed in a scene and illuminated by one of six high dynamic range (HDR) light probes. Light probes were white-balanced 8192×4096pixel images of four exterior (beach, bay, woodland, savanah) and two interior (church, conference hall) environments, captured by Unity Technologies (107). A virtual camera (60° field of view) hovered 12cm above the sheet, viewing directly downwards. By randomly varying the camera's location and orientation, an extremely large number of unique images could be generated from the scene.

Main training dataset. 10,000 images were rendered to create the main training dataset for neural network models. For each rendering, one of the six HDRIs was randomly selected to illuminate the scene. Surface relief was scaled multiplicatively in depth by a uniformly-sampled random scaling factor, so that the distance between the lowest and highest points on the surface was between 0.1–2.5cm. The surface's reflectance properties were controlled via the Unity3D standard shader, using the 'specular setup'. The surface's base colour was chosen by selecting RGB values randomly uniformly within the interval 0.3 to 0.7. The surface was either low or high gloss, with equal probability. For low gloss surfaces, specular reflectance was selected randomly uniformly between 0.1–0.3, and the concentration of the specular component

was randomly uniformly between 0.2–0.4. For high gloss surfaces, specular reflectance was between 0.3–0.5, and specular concentration between 0.75–0.95. The training dataset therefore had a bimodal distribution of approximately 50% low- and 50% high-gloss surfaces, with small variations in reflectance properties within each group. The same dataset was also used when training classifiers to decode gloss and other world factors from models.

Continuously-sampled gloss training dataset. To verify that the representations learned by our models were not artefacts of the bimodal gloss sampling in the main dataset, we rendered 10,000 new images in which specular reflectance was sampled randomly uniformly between 0.1–0.8, and specular concentration was set to the same value. All other scene properties were varied as in the main dataset.

Higher-variation training dataset. To verify that less constrained visual diets could lead to similar gloss representations, we created a third training dataset. 10,000 images each randomly combined one of ten novel surface geometries with fifty novel light fields. Novel geometries were virtual 40×40cm sheets with different shapes and sizes of irregularly placed ridges, bumps, and indentations (examples in **Figure 7A**). Novel light fields were fifty 4096×2048pixel HDR images (25 exterior, 25 interior) from the HDRI Haven database (http://www.hdrihaven.com). The distance of the virtual camera above the surface varied randomly uniformly between 8–12cm to introduce scale variation. All other scene properties varied as in the main dataset.

Experiment 1: Renderings with continuously sampled gloss levels. Stimuli were 50 new rendered images with specular reflectance chosen randomly uniformly between 0.2–1.0, and specular concentration set to the same value. Other attributes were varied as in the main dataset.

Experiment 2: Gloss-modulated images generated from PixelVAE models. For each of the PixelVAE networks, we first determined the axis in 10D latent space along which gloss could be most strongly discriminated (see **Data analysis**). Eight images (four low-gloss, four high-gloss) from the main training dataset were input to each

network, and their corresponding latent points were recorded. These were used as 'seed points' to generate eight 'gloss-modulated' sequences for each network. The first step in each sequence was generated by conditioning the model's pixel sampling process on the respective seed point in latent space, and sampling a 128×128pixel image (see **Neural network architectures and training**). The conditioning point was then moved 0.07 units in latent space along that model's gloss-discriminating axis, either in the matte (for high-gloss seed images) or glossy (for high-gloss seed images) direction, and a second image was generated. This was repeated three more times, yielding five-step sequences for each seed image from each model training instance (400 experimental images in total).

Experiments 3a and 3b: Sequences of renderings with increasing surface relief. To create a strong test of the different models, we wanted to probe human gloss constancy using stimuli (a) for which there were clear differences between constancy patterns predicted by unsupervised vs supervised models, and (b) which were likely to produce diverse patterns of failures of constancy in human observers. We therefore first generated candidate stimuli, then selected those that best satisfied these desiderata (83)

For each experiment, we rendered 200 different candidate sequences of 7 images with fixed random material, illumination, and camera position, but exponentially increasing surface relief (maximum bump heights of {0.10, 0.18, 0.31, 0.55, 0.97, 0.70, 3.0}cm). All surfaces had relatively high gloss. In Experiment 3a, specular reflectance was randomly uniformly between 0.2–1.0, and specular concentration 0.7–1.0. In Experiment 3b, reflectance was between 0.1–0.3, and concentration 0.8–1.0. Other attributes were varied as in the main dataset. All 1400 images in each set were input to each of the unsupervised PixelVAE and supervised ResNet networks, to obtain predicted gloss values (see **Data analysis**). For each candidate sequence of 7 images, we then simulated the responses from each network under a two-alternative forced-choice (2AFC) experiment in which each possible pair of images was compared, and the image for which the model predicted a higher gloss value was selected. This was

analogous to the “which is more glossy?” task performed by human observers (see **Psychophysical procedures**).

First, for each sequence we performed a 2×7 (model × surface relief) ANOVA between unsupervised and supervised model predictions, averaged over training instances of each model. Sequences were sorted according to the *F* interaction statistic, prioritising those with strong disagreements between model predictions (83). In Experiment 3a, we visually inspected the top-ranked sequences and selected 10 with diverse appearances and subjective failures of gloss constancy.

In Experiment 3b we selected test sequences in an entirely automated manner by classifying sequences into four qualitative groups based on the average constancy pattern predicted by the unsupervised models:

- Group 1: model predicts approximate constancy (range of predicted “proportion glossier” values < 0.25, and a linear function fit achieved $R^2 > 0.70$);
- Group 2: model predicts an approximately linear increase in gloss with bump height (range > 0.50, and a linear fit achieved $R^2 > 0.90$);
- Group 3: model predicts a non-monotonic failure of constancy (range > 0.50, and a quadratic fit of $R^2 > 0.90$ with a negative coefficient of the squared term);
- Group 4: model predicts an upward-swinging non-linear failure of constancy (range > 0.50, and a quadratic fit of $R^2 > 0.90$ with a positive coefficient of the squared term).

We then selected the top sequences in each group (ranked by *F* interaction term), in proportion to the size of each group (1 “constant,” 4 “linear,” 2 “non-monotonic,” and 3 “non-linear” sequences).

Experiment 4: Pairs of renderings with differing surface relief or lighting. Here we sought pairs of stimuli (a) with diverse scene differences (some differing in surface relief, some in lighting environment, and some in angle of illumination), and (b) for which the unsupervised model made a wide range of predictions, ranging from good constancy to strong failure of constancy.

We rendered 800 candidate image pairs with the same random material properties and camera position, but differing in either surface relief (0.40 vs 1.5cm maximum bump height) or illumination. Pairs differing in illumination were either 'naturally' lit (differing in HDR lighting environment, from among the six used in the ***main training dataset***), or illuminated by a directional lamp (at a 30° vs 90° angle). Candidate pairs were generated in four groups of 200:

- Group 1: identical natural light field; different surface relief
- Group 2: identical lamp angle; different surface relief
- Group 3: identical surface relief; different natural light field
- Group 4: identical surface relief; different lamp angle

Average predicted gloss values were obtained for each image, across training instances of the unsupervised PixelVAE model (see **Data analysis**). Within each group, we ranked pairs by the absolute difference in predicted gloss of the images in the pair, and selected pairs lying at each of the 10th percentiles, yielding 40 test pairs.

Real-world images. To test generalisation to real-world images (**Supplementary Figure 4A**), we took ten 128×128pixel colour close-up photographs of common glossy surfaces (e.g. tomatoes, plastic, metallic foil) and ten of matte surfaces (e.g. chalk, fleece, styrofoam).

Geometry and lighting generalisation test set. To test generalisation of models trained on the main dataset, we rendered five sets of 500 new images, one using the original surface sheet, and four using novel surface geometries (**Figure 7A**). Illumination was randomly one of eight novel 4096×2048pixel HDR light probes (six exterior, two interior) from the Dutch Skies 360º HDRI Project (http://www.dutch360hdr.com/). Specular reflectance was varied as in the ***continuously-sampled gloss training dataset***.

Highlight-feature manipulated images. To manipulate highlight features (**Figure 8**), we manipulated the specular component image in four different ways, before combining with the diffuse component image by addition:

- (1) Translated highlights: specular component was shifted rightwards in ten 5-pixel steps. To avoid edge artefacts after translation, images were rendered at 1024×1024 resolution, and then cropped to the lower right 800×800 pixels.
- (2) Reducing highlight coverage: we applied image erosion to the specular component using the OpenCV package for Python. Kernel size was 2×2. Up to ten iterations were applied to create ten progressively reduced highlight maps.
- (3) Reducing highlight sharpness: specular image was convolved with a Gaussian filter, with step size ranging from 1×1 to 11×11.
- (4) Reducing highlight contrast: we created a 'highlight mask' by identifying pixels with values >4 (from 0-255 range) in greyscale version of specular image. Standard deviation of pixel intensity within the highlight map was multiplicatively reduced in ten steps, from 1 to 0.1 times their original SD, while retaining the original mean.

Neural network architectures and training

DNNs were implemented and trained in Python 2.7 with Tensorflow 1.14 on a GPU-accelerated machine using 1-4 GTX1080 graphics cards. Networks were trained on the first 9,000 images of the ***main training dataset***, using the next 500 for validation, and the final 500 for testing accuracy of the supervised networks. All images were input at 128×128 pixel resolution. For both unsupervised and supervised models, ten independent instances of the same architecture were trained from different random initial weights and with different random sampling of training batches, to ensure robustness to representational differences between training instances (64). Additionally, we trained five instances of the PixelVAE network on the ***continuously-sampled gloss*** dataset and one on the ***higher-variation*** dataset (see **Stimuli**), dividing training and validation data in the same proportions. No data augmentation was used. All architectures used rectified linear activation functions.

Unsupervised PixelVAE model. We used the implementation from (41) of the PixelVAE architecture (42), available at https://github.com/ermongroup/Generalized-PixelVAE. The architecture consists of two streams of convolutional layers, which learn

jointly via backpropagation. One stream is a 'conditioning network,' which is a convolutional variational autoencoder that takes an image as input and outputs a 10-dimensional smooth latent representation. We chose a 10D latent code as being approximately the most compact representation that still allowed the network to learn a good model of the training images, based on pilot data (see **Supplementary Figure 1A-B**). The other stream is an autoregressive PixelCNN++ (108,109)(108,109) model that learns the structure of the training data in terms of a polynomial probability distribution over pixel values, and takes as inputs both the image and the latent code output by the conditioning network. To generate new images, the autoregressive stream chooses an RGB value for each pixel, working from the top left to the bottom right of the image. Each pixel is sampled from the learned probability distribution, conditioning on both the values of pixels generated so far (which constrain the local structure of the image), and on the values provided in the latent code (which constrain more holistic image properties).

The conditioning network consisted of three convolutional layers of 64, 128, and 256 feature maps, followed by a fully connected layer of 512 units and a 10-unit latent code layer. The autoregressive network consisted of six 'residual blocks' of three layers of 64 convolutional feature maps each, with skip connections linking the first and sixth, and second and fifth blocks. Pixel likelihood distributions were modelled with a mixture of 12 logistic functions, and networks were trained with a batch size of 5 for 200 epochs, around which point validation error plateaued. No regularisation was used during training, to encourage the network to depend on information in its latent code (41).

Supervised ResNet model. We used the Tensorflow implementation of the ResNet architecture (68) from https://github.com/wenxinxu/resnet-in-tensorflow. Networks consisted of three 'residual blocks' each made up of three layers of 56 convolutional feature maps, with skip connections between each. The output of the final layer was passed to a 10-unit fully-connected layer, which we treated as the 'latent code' of the model for analysis purposes. This passed, via one more non-linearity to a two-unit softmax output layer. Ten networks were trained, from different random initial weights, to classify images as renderings of high or low specular reflectance surfaces (see

Stimuli above). Networks were trained with a batch size of 32 and a learning rate of 0.001 to minimise the sparse softmax cross-entropy between outputs and correct labels. Networks were trained for 75 epochs, by which point validation accuracy plateaued above 99%.

Simple autoencoder. We also considered a far simpler unsupervised model, in the form of a non-variational convolutional autoencoder implemented using Keras 2.2.5 for Tensorflow. The architecture comprised four layers of 64, 32, 16, and 16 feature maps alternating with 2×2 pooling layers, leading to a 4,096-unit fully connected bottleneck layer, which we treated as the model's latent feature space for analysis purposes. The compressed code was expanded through mirrored layers of 16, 16, 32, and 64 feature maps to an output of the original dimensionality (128×128×3). The network was trained for 1,000 epochs to minimise mean absolute error between its input and output images (batch size 32, other learning parameters used default Adam optimiser values as implemented in Keras).

Object-trained deep neural network. Finally, we evaluated a pre-trained DNN supervised on object-recognition: an 18-layer Resnet (68) model available from the MATLAB 2019b Deep Learning toolbox (https://de.mathworks.com/help/deeplearning/ref/resnet18.html). The network had been pre-trained on the Imagenet Large-Scale Visual Recognition Challenge (ILSVRC) database to classify 1.2 million images into 1,000 object and animal categories (52). For analyses, we used representations in the 1,000-unit fully-connected layer immediately before softmax category readout.

Additional comparison models

Histogram skewness. Skewness was defined as the skew (3rd moment) of the distribution of pixel intensities in a greyscale version of each image.

MDS / tSNE/ LLE. Multi-dimensional scaling (MDS) is a linear dimensionality reduction technique that finds the linear projection of a dataset that best preserves distances between all data points. t-distributed Stochastic Neighbourhood Embedding (tSNE;

(70)) is a non-linear technique that preferentially preserves distances between *nearby* data points. Locally Linear Embedding (LLE; (71)) is a non-linear technique that seeks a lower-dimensional manifold that best preserves all distances.

Each dimensionality reduction algorithm was used to create a 10D embedding of the 10,000 images from the main training dataset, as well as all 270 rendered images used in experiments (50 images from Experiment 1; 140 from Experiment 3; 80 from Experiment 4). The experimental probe images were included because it is not possible to project new data points into the reduced-dimensionality solution discovered by tSNE or LLE. Additional 2D embeddings were performed to create visualisations (**Figure 2A-C** and **Supplementary Figure 4B-C**) using 4,000 images randomly selected from the main training dataset. Default parameters were used, as implemented in the scikit-learn package for Python.

Texture features. For each image we calculated a multi-scale texture feature description (80) using the TextureSynth package for MATLAB (www.cns.nyu.edu/~lcv/texture/). Images were rescaled to 256×256 pixels, and analysed at 3 scales, 4 orientations, with a spatial neighbourhood of 7, producing a description of each image in terms of 1,350 feature dimensions.

Psychophysical procedures

Psychophysical experiments were conducted in a dimly lit room, sitting at a comfortable distance from the screen. Participants could freely view the screen, and were given no fixation instructions. Stimuli were presented on an EIZO ColorEdge CG277 self-calibrating LCD monitor with a resolution of 2560×1440 and a refresh rate of 60Hz. PsychoPy 3.1 with Python 3.6 was used to present stimuli and record responses. At the beginning of each experiment, observers were shown 4-12 example experimental stimuli, randomly arranged, to familiarise them with the appearance and range of material properties they would be asked to judge. Response times were unconstrained, and stimuli were displayed until a response was recorded.

Experiment 1: Gloss ratings for novel rendered images. Fifty 800×800 pixel (18.6×18.6cm) images were presented singly in the centre of the screen. Each was repeated three times, for a total of 150 trials, presented in a random order. Participants were asked to rate the apparent gloss of each image using a 6-point scale with endpoints labelled “1 = completely matte” and “6 = extremely glossy”.

Experiment 2: Gloss rankings for gloss-modulated network samples. Eighty sets of five 128×128 pixel (2.9×2.9cm) PixelVAE-generated images were shown. The five images within each set were arrayed in a random order in a row in the centre of the screen. Participants were asked to use the keyboard number keys to sort the images into a row of initially empty boxes at the bottom of the screen, in order from least glossy to most glossy. Each set of stimuli was shown once, for a total of 80 trials, presented in a random order.

Experiments 3a and 3b: Pairwise comparisons of gloss-constancy probe sequences. Experiments 3a and 3b differed only in the specific sequences shown to participants. Both experiments consisted of 10 sequences of 7 images, within which all possible pairs were shown twice each, for a total of 420 trials. Images appeared at a resolution of 800×800 pixels (18.6×18.6cm), side by side on screen. Side of screen and trial order were randomised, with all sequences interleaved with one another. Participants were asked to report, using the left and right arrow keys, which surface appeared more glossy.

Experiment 4: Pairwise comparisons of gloss-constancy probe pairs. Each of 40 pairs of images was shown 8 times, for a total of 360 trials. Images appeared at a resolution of 128×128 pixels (2.9×2.9cm), side by side, with side of screen and trial order randomised. Participants reported, using the left and right arrow keys, which surface appeared more glossy.

Data analysis

All analyses of human and model data were performed in Python 2.7 or 3.6, using numpy v1.16.5 and/or scikit-learn v0.21.3 packages.

Representational similarity analysis. We measured average correlation distance (1 minus Pearson *r*) between the latent representations of all 10,000 images in the ***main training dataset***, grouped by whether they used the same (diagonal blocks) or different (off-diagonal blocks) gloss/lighting conditions, for each network training instance. Representational dissimilarity matrices (RDMs; **Figure 2D-E**) were created by averaging these values over training instances of each model. To visualise the similarity of model predictions (**Figure 7B**) we created a vector, for each model, of predicted gloss values for all 270 rendered images used as experimental stimuli (50 images from Experiment 1; 140 images from Experiment 3; and 80 images from Experiment 4), normalised into the range 0–1. Euclidean distance between these prediction vectors was calculated for all pairs of models.

Decoding world factors from models and deriving gloss predictions. For models with multidimensional feature spaces (i.e. all except histogram skewness), a linear support vector machine (SVM) was used to classify specular reflectance (high vs low) of rendered images from their representations within the model. SVMs were trained on a random sample of 7,500 of the main training dataset images, and tested on the remaining 2,500 to derive gloss classification accuracies (**Figure 3**, **Figure 6A** and **Supplementary Figure 3A**). To derive continuously-valued gloss predictions for experimental images, images were input to each model, and the signed distance of their representation from the model's SVM decision boundary was measured. Positive values indicate "high-gloss" classifications and negative values indicate "low-gloss", with absolute magnitude indicating the strength of evidence for that classification (**Figure 3**, **Figure 4B-C**, **Figure 5D-E**, **Figure 7A**, **Figure 8B**, **Supplementary Figures 3B** and **4A**).

For unsupervised PixelVAE and supervised ResNet models, we also trained linear SVMs to perform a 6-way light-field classification, and fitted linear regressions to predict surface relief (**Figure 3B**). Classifiers and regressions were performed once using the full 10D latent space for each network, and again using each of the network's

individual latent dimensions (**Figure 3B**; see also **Supplementary Figure 1C** for visualisation of individual dimensions in one unsupervised network).

For the histogram skewness model, raw skew values were used as gloss predictors for the purposes of model comparison. Ground-truth gloss classification accuracy (**Figure 6A**) was defined as the accuracy using an optimal threshold to separate high- from low-specular reflectance images from one another on the basis of their skewness, fitting the threshold on 7,500 main training dataset images and testing on the remaining 2,500.

Deriving predicted 2AFC experimental data from models. Model predictions for Experiment 3 were derived by simulating responses of each model in the 2AFC task performed by humans. For each sequence of 7 images, the predicted gloss values of each possible pair of images was compared, and the image for which the model predicted higher gloss was selected. Predicted responses were summarised as "proportion selected as being glossier" for each level in the sequence, as for human data (**Figure 5B** and **Supplementary Figure 2**).

Measuring highlight features. Coverage, sharpness and contrast of highlights (**Figure 8**) were measured using a MATLAB package developed by (110). Briefly: (1) Coverage is defined as the proportion of pixels with higher intensities in the full image than in the diffuse component image; (2) Sharpness is defined as the Local Phase Coherence (111) within the specular component image; (3) Contrast is defined as the sum of the RMS contrast across eight bandpass-filtered versions of the specular component image.

Psychophysical data analysis

No participants or trials were excluded from analysis. In Experiment 1 ratings were averaged across participants for each stimulus. In Experiment 2, rankings were converted to average rank position of each image step for each participant, averaging within matte and glossy seed image sequences. In Experiments 3a and 3b, pairwise comparisons were converted to the proportion of times each image within each

sequence was judged as having higher gloss, across pairings and repetitions. In Experiment 4, pairwise comparisons were converted to the proportion of times ‘image A’ (arbitrarily labelled) was judged as having higher gloss than ‘image B,’ and this proportion subtracted from 0.5 (i.e. 0 indicates equal apparent gloss, and good constancy; deviations in either direction indicate deviations from constancy).

1

References

1. Adelson EH. Lightness Perception and Lightness Illusions. In: The New Cognitive Neurosciences. 2nd ed. Cambridge: MIT Press; 2000. p. 339–51.
2. Anderson BL. Mid-level vision. Current Biology. 2020;30(3):R105–R109.
3. Anderson BL. The perceptual representation of transparency, lightness, and gloss. In: Handbook of Perceptual Organization. 2015;466–483.
4. Barrow H, Tenenbaum J, Hanson A, Riseman E. Recovering intrinsic scene characteristics. Comput Vis Syst. 1978;2(3–26):2.
5. Fleming RW. Material perception. Annual Review of Vision Science. 2017;3:365–388.
6. Todd JT. The visual perception of 3D shape. Trends in Cognitive Sciences. 2004;8(3):115–121.
7. Todd JT, Norman JF, Mingolla E. Lightness constancy in the presence of specular highlights. Psychological Science. 2004;15(1):33–39.
8. Marr D. Vision. San Francisco: Freeman; 1982.
9. Kersten D, Mamassian P, Yuille A. Object perception as Bayesian inference. Annual Review of Psychology. 2004;55:271–304.
10. Geisler WS, Kersten D. Illusions, perception and Bayes. Nature Neuroscience. 2002;5(6):508.
11. von Helmholtz H. Treatise on Physiological Optics Vol. iii. 1867;
12. Anderson BL. Can computational goals inform theories of vision? Topics in Cognitive Science. 2015;7(2):274–286.
13. Hoffman DD, Singh M, Prakash C. The interface theory of perception. Psychonomic Bulletin & Review. 2015;22(6):1480–1506.
14. Fleming RW, Storrs KR. Learning to see stuff. Current Opinion in Behavioral Sciences. 2019;30:100–108.

15. Ho Y-X, Landy MS, Maloney LT. Conjoint measurement of gloss and surface texture. Psychological Science. 2008;19(2):196–204.
16. Marlow PJ, Kim J, Anderson BL. The Perception and Misperception of Specular Surface Reflectance. Current Biology. 2012 Oct 23;22(20):1909–13.
17. Doerschner K, Fleming RW, Yilmaz O, Schrater PR, Hartung B, Kersten D. Visual motion and the perception of surface material. Current Biology. 2011;21(23):2010–2016.
18. Wendt G, Faul F, Ekroll V, Mausfeld R. Disparity, motion, and color information improve gloss constancy performance. Journal of Vision. 2010;10(9):7–7.
19. Chadwick AC, Kentridge R. The perception of gloss: A review. Vision research. 2015;109:221–235.
20. Beck J, Prazdny S. Highlights and the perception of glossiness. Perception & Psychophysics. 1981;
21. Anderson BL, Kim J. Image statistics do not explain the perception of gloss and lightness. Journal of vision. 2009;9(11):10–10.
22. Toscani M, Guarnera D, Guarnera C, Hardeberg JY, Gegenfurtner K. Three perceptual dimensions for specular and diffuse reflection. ACM Transactions on Applied Perception. 2020;
23. Ferwerda JA, Pellacini F, Greenberg DP. Psychophysically based model of surface gloss perception. In: Human Vision and Electronic Imaging VI. International Society for Optics and Photonics; 2001. p. 291–301.
24. Lagunas M, Malpica S, Serrano A, Garces E, Gutierrez D, Masia B. A Similarity Measure for Material Appearance. ACM Transactions on Graphics (SIGGRAPH 2019). 2019;38(4).
25. Ingersoll LR. The Glarimeter: an Instrument for Measuring the Gloss of Paper. Journal of the Optical Society of America. 1921;5(3):213–217.

26. Ward GJ. Measuring and modeling anisotropic reflection. In: Proceedings of the 19th annual conference on Computer graphics and interactive techniques. 1992. p. 265–272.
27. Wills J, Agarwal S, Kriegman D, Belongie S. Toward a perceptual space for gloss. ACM Transactions on Graphics (TOG). 2009;28(4):1–15.
28. Serrano A, Gutierrez D, Myszkowski K, Seidel H-P, Masia B. An intuitive control space for material appearance. ACM Transactions on Graphics (SIGGRAPH ASIA 2016). 2016;35(6).
29. Obein G, Knoblauch K, Viéot F. Difference scaling of gloss: Nonlinearity, binocularity, and constancy. Journal of Vision. 2004;4(9):4–4.
30. Fleming RW, Dror RO, Adelson EH. Real-world illumination and the perception of surface reflectance properties. Journal of Vision. 2003;3(5):3–3.
31. Vangorp P, Laurijssen J, Dutré P. The influence of shape on the perception of material reflectance. In: ACM SIGGRAPH 2007 papers. 2007. p. 77–es.
32. Olkkonen M, Brainard DH. Perceived glossiness and lightness under real-world illumination. Journal of Vision. 2010;10(9):5–5.
33. Attneave F. Some informational aspects of visual perception. Psychological Review. 1954;61(3):183.
34. Barlow HB. Possible principles underlying the transformation of sensory messages. Sensory Communication. 1961;1:217–234.
35. Simoncelli EP, Olshausen BA. Natural image statistics and neural representation. Annual Review of Neuroscience. 2001;24(1):1193–1216.
36. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 1996;381(6583):607–609.
37. Grossberg S. Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. Biological Cybernetics. 1976;23(3):121–134.

38. Földiak P. Forming sparse representations by local anti-Hebbian learning. Biological Cybernetics. 1990;64(2):165–170.
39. Phong BT. Illumination for computer generated pictures. Communications of the ACM. 1975;18(6):311–317.
40. Salakhutdinov R. Learning deep generative models. Annual Review of Statistics and Its Application. 2015;2:361–385.
41. Zhao S, Song J, Ermon S. Towards deeper understanding of variational autoencoding models. arXiv preprint arXiv:170208658. 2017;
42. Gulrajani I, Kumar K, Ahmed F, Taiga AA, Visin F, Vazquez D, et al. Pixelvae: A latent variable model for natural images. arXiv preprint arXiv:161105013. 2016;
43. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. International Conference on Learning Representations, 2016. arXiv preprint arXiv:151106434. 2016;
44. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. International Conference on Learning Representations. 2017;2(5):6.
45. Lindsay G. Convolutional neural networks as a model of the visual system: past, present, and future. Journal of Cognitive Neuroscience. 2020;1–15.
46. Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience. 2016;19(3):356.
47. Storrs KR, Kriegeskorte N. Deep learning for cognitive neuroscience. In: The Cognitive Neurosciences. 6th ed. MIT Press; 2020.
48. Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, et al. A deep learning framework for neuroscience. Nature Neuroscience. 2019;22(11):1761–1770.

49. Kriegeskorte N. Deep neural networks: a new framework for modeling biological vision and brain information processing. Annual Review of Vision Science. 2015;1:417–446.
50. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May;521(7553):436–44.
51. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–778.
52. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision. 2015;115(3):211–252.
53. Taigman Y, Yang M, Ranzato M, Wolf L. Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 1701–1708.
54. Cadieu CF, Hong H, Yamins DL, Pinto N, Ardila D, Solomon EA, et al. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Computational Biology. 2014;10(12):e1003963.
55. Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, et al. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? bioRxiv preprint, bioRxiv:10.1101/407007v1, 2018;
56. Storrs KR, Kietzmann TC, Walther A, Mehrer J, Kriegeskorte N. Diverse deep neural networks all predict human IT well, after training and fitting. bioRxiv preprint, bioRxiv:2020.05.07.082743. 2020;
57. Khaligh-Razavi S-M, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Computational Biology. 2014;10(11).

58. Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 427–436.
59. Geirhos R, Temme CR, Rauber J, Schütt HH, Bethge M, Wichmann FA. Generalisation in humans and deep neural networks. In: Advances in Neural Information Processing Systems. 2018. p. 7538–7550.
60. Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:181112231. 2018;
61. Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut Learning in Deep Neural Networks. arXiv preprint arXiv:200407780. 2020;
62. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:13126114. 2013;
63. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006;313(5786):504–507.
64. Mehrer J, Spoerer CJ, Kriegeskorte N, Kietzmann TC. Individual differences among deep neural network models. bioRxiv preprint, bioRxiv:10.1101/2020.01.08.898288v1. 2020;
65. DiCarlo JJ, Cox DD. Untangling invariant object recognition. Trends in Cognitive Sciences. 2007;11(8):333–341.
66. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence. 2013;35(8):1798–1828.
67. Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. A toolbox for representational similarity analysis. PLoS Computational Biology. 2014;10(4).

68. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. 2015. p. 1026–1034.
69. Testolin A, Stoianov I, Zorzi M. Letter perception emerges from unsupervised deep learning and recycling of natural image features. Nature Human Behaviour. 2017;1(9):657–664.
70. Maaten L van der, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008;9(Nov):2579–2605.
71. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000;290(5500):2323–2326.
72. Kriegeskorte N, Kievit RA. Representational geometry: integrating cognition, computation, and the brain. Trends in Cognitive Sciences. 2013;17(8):401–412.
73. Kriegeskorte N, Diedrichsen J. Inferring brain-computational mechanisms with models of activity measurements. Philosophical Transactions of the Royal Society B: Biological Sciences. 2016;371(1705):20160278.
74. Testolin A, Zorzi M. Probabilistic models and generative neural networks: Towards an unified framework for modeling normal and impaired neurocognitive functions. Frontiers in Computational Neuroscience. 2016;10:73.
75. Hong H, Yamins DL, Majaj NJ, DiCarlo JJ. Explicit information for category-orthogonal object properties increases along the ventral stream. Nature Neuroscience. 2016;19(4):613.
76. Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. Neuroimage. 2011;56(2):400–410.
77. Gatys L, Ecker AS, Bethge M. Texture synthesis using convolutional neural networks. In: Advances in neural information processing systems. 2015. p. 262–270.

78. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. p. 586–595.
79. Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. Journal of Neuroscience. 2018;38(33):7255–7269.
80. Portilla J, Simoncelli EP. A parametric texture model based on joint statistics of complex wavelet coefficients. International Journal of Computer Vision. 2000;40(1):49–70.
81. Motoyoshi I, Nishida S, Sharan L, Adelson EH. Image statistics and the perception of surface qualities. Nature. 2007;447(7141):206–209.
82. Funke CM, Borowski J, Stosio K, Brendel W, Wallis TS, Bethge M. The Notorious Difficulty of Comparing Human and Machine Perception. arXiv preprint arXiv:200409406. 2020;
83. Golan T, Raju PC, Kriegeskorte N. Controversial stimuli: pitting neural networks against each other as models of human recognition. arXiv preprint arXiv:191109288. 2019;
84. Wang Z, Simoncelli EP. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. Journal of Vision. 2008;8(12):8–8.
85. Marlow PJ, Todorović D, Anderson BL. Coupled computations of three-dimensional shape and material. Current Biology. 2015;25(6):R221–R222.
86. Marlow PJ, Anderson BL. Material properties derived from three-dimensional shape representations. Vision Research. 2015;115:199–208.
87. Marlow PJ, Anderson BL. Generative constraints on image cues for perceived gloss. Journal of Vision. 2013;13(14):2–2.

88. Barlow H. The exploitation of regularities in the environment by the brain. Behavioral and Brain Sciences. 2001;24(4):602–607.
89. Simoncelli EP. Vision and the statistics of the visual environment. Current Opinion in Neurobiology. 2003;13(2):144–149.
90. Nishida S. Image statistics for material perception. Current Opinion in Behavioral Sciences. 2019;30:94–99.
91. Fleming RW. Visual perception of materials and their properties. Vision Research. 2014 Jan 1;94:62–75.
92. Anderson BL. Is “Inverse Optics” Actually Ill-Posed? [Internet]. International Symposium on the Future of Shitsukan Research; 2014 Jul 16; University of Tokyo. Available from: https://www.youtube.com/watch?v=Y0CMD70yk2k
93. Marlow PJ, Anderson BL. Motion and texture shape cues modulate perceived material properties. Journal of Vision. 2016;16(1):5–5.
94. Leibo JZ, Liao Q, Anselmi F, Freiwald WA, Poggio T. View-tolerant face recognition and Hebbian learning imply mirror-symmetric neural tuning to head orientation. Current Biology. 2017;27(1):62–67.
95. Wallis G, Baddeley R. Optimal, unsupervised learning in invariant object recognition. Neural Computation. 1997;9(4):883–894.
96. Wiskott L, Sejnowski TJ. Slow feature analysis: Unsupervised learning of invariances. Neural Computation. 2002;14(4):715–770.
97. Wiesel TN, Hubel DH. Ordered arrangement of orientation columns in monkeys lacking visual experience. Journal of Comparative Neurology. 1974;158(3):307–318.
98. Yang J, Otsuka Y, Kanazawa S, Yamaguchi MK, Motoyoshi I. Perception of surface glossiness by infants aged 5 to 8 months. Perception. 2011;40(12):1491–1502.

99. Balas B. Children's use of visual summary statistics for material categorization. Journal of Vision. 2017;17(12):22–22.

100. Balas B, Auen A, Thrash J, Lammers S. Children's use of local and global visual features for material perception. Journal of Vision. 2020;20(2):10–10.

101. Smith LB, Slone LK. A developmental approach to machine learning? Frontiers in Psychology. 2017;8:2124.

102. Pouget A, Beck JM, Ma WJ, Latham PE. Probabilistic brains: knowns and unknowns. Nature Neuroscience. 2013;16(9):1170.

103. Friston K. The free-energy principle: a rough guide to the brain? Trends in Cognitive Sciences. 2009;13(7):293–301.

104. Deneve S. Bayesian spiking neurons I: inference. Neural Computation. 2008;20(1):91–117.

105. Brainard DH, Roorda A, Yamauchi Y, Calderone JB, Metha A, Neitz M, et al. Functional consequences of the relative numbers of L and M cones. Journal of the Optical Society of America A. 2000;17(3):607–614.

106. Smirnakis SM, Berry MJ, Warland DK, Bialek W, Meister M. Adaptation of retinal processing to image contrast and spatial scale. Nature. 1997 Mar;386(6620):69–73.

107. Lagarde S, Lachambre S, Jover S. An artist-friendly workflow for panoramic HDRI. SIGGRAPH. 2016;

108. Salimans T, Karpathy A, Chen X, Kingma DP. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. International Conference on Learning Representations. 2017;2.

109. Van den Oord A, Kalchbrenner N, Espeholt L, Vinyals O, Graves A, others. Conditional image generation with pixelcnn decoders. In: Advances in neural information processing systems. 2016. p. 4790–4798.

110. Schmid AC, Doerschner K. Chocolate, chrome, or cloth? The appearance of specular highlights determines perceived material category. Journal of Vision. 2019;19(10):252–252.

111. Hassen R, Wang Z, Salama MMA. Image Sharpness Assessment Based on Local Phase Coherence. IEEE Transactions on Image Processing. 2013 Jul;22(7):2798–810.

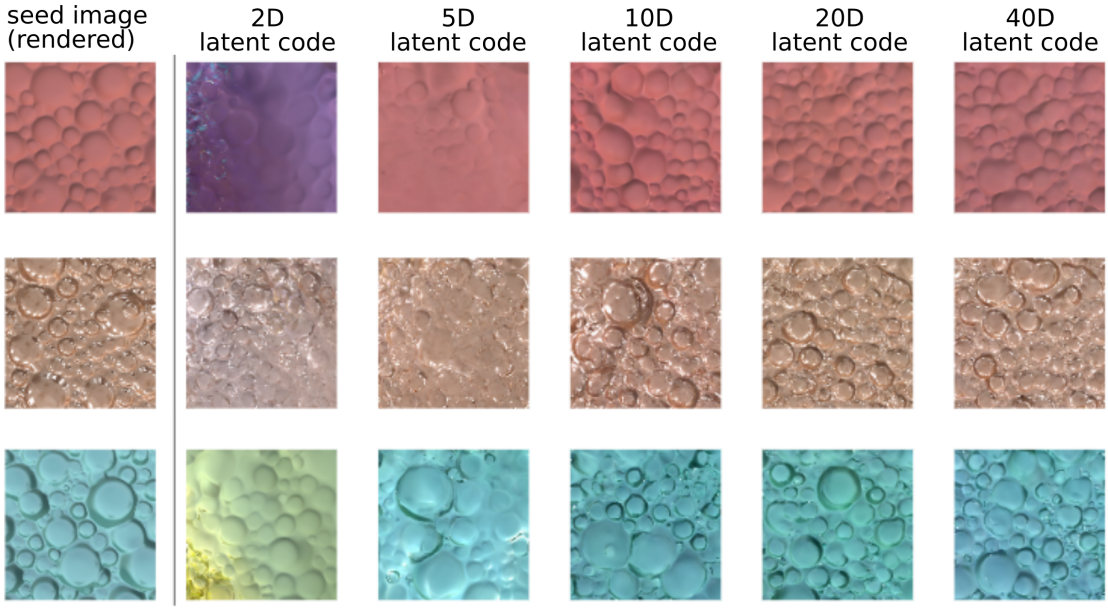
Acknowledgements: This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–project number 222641018–SFB/TRR 135 TP C1, by the European Research Council (ERC) Consolidator Award 'SHAPE'–project number ERC-CoG-2015-682859 and by an Alexander von Humboldt Postdoctoral Research Fellowship to K.R.S.. We would like to thank Alexandra Schmid and Katja Doerschner for sharing code to implement the highlight feature measurement model, and Karl Gegenfurtner, Bart Anderson and James Todd for comments on earlier versions of this manuscript.

Author contributions: Conceptualization, K.R.S. and R.W.F.; Methodology, K.R.S. and R.W.F.; Software, K.R.S.; Formal Analysis, K.R.S.; Investigation, K.R.S.; Resources, R.W.F.; Writing – Original Draft, K.R.S. and R.W.F.; Writing – Review & Editing, K.R.S. and R.W.F.; Visualization, K.R.S; Supervision, R.W.F.; Funding Acquisition, R.W.F. and K.R.S..

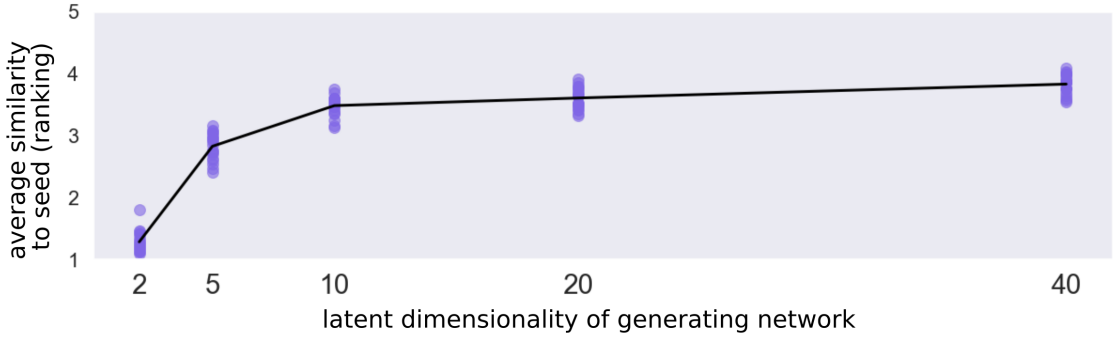
Declaration of interests: none

Supplementary Figures

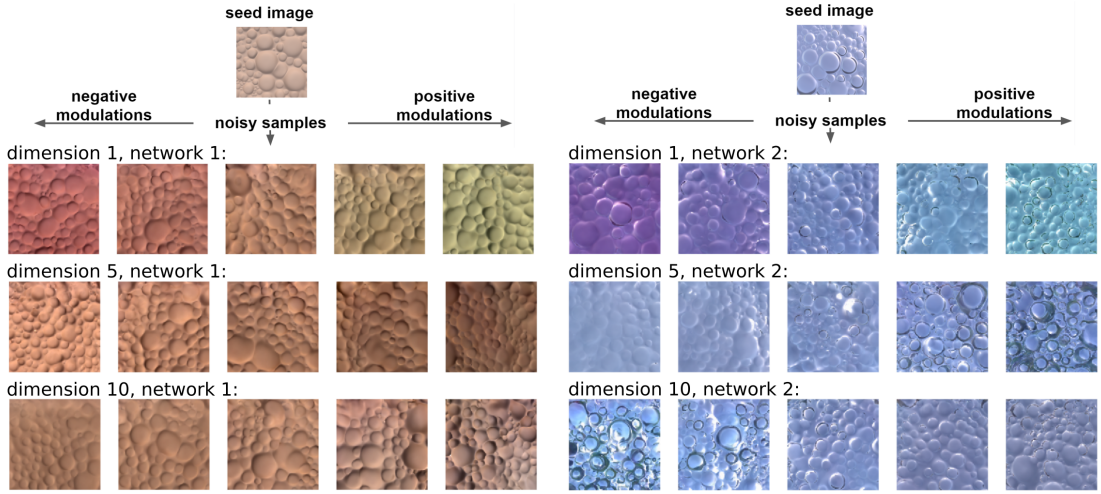
A Example images generated by PixelVAEs of different latent dimensionalities



B Perceived similarity to seed image, as a function of latent dimensionality



C Example visualisations of individual dimensions in 10D latent codes



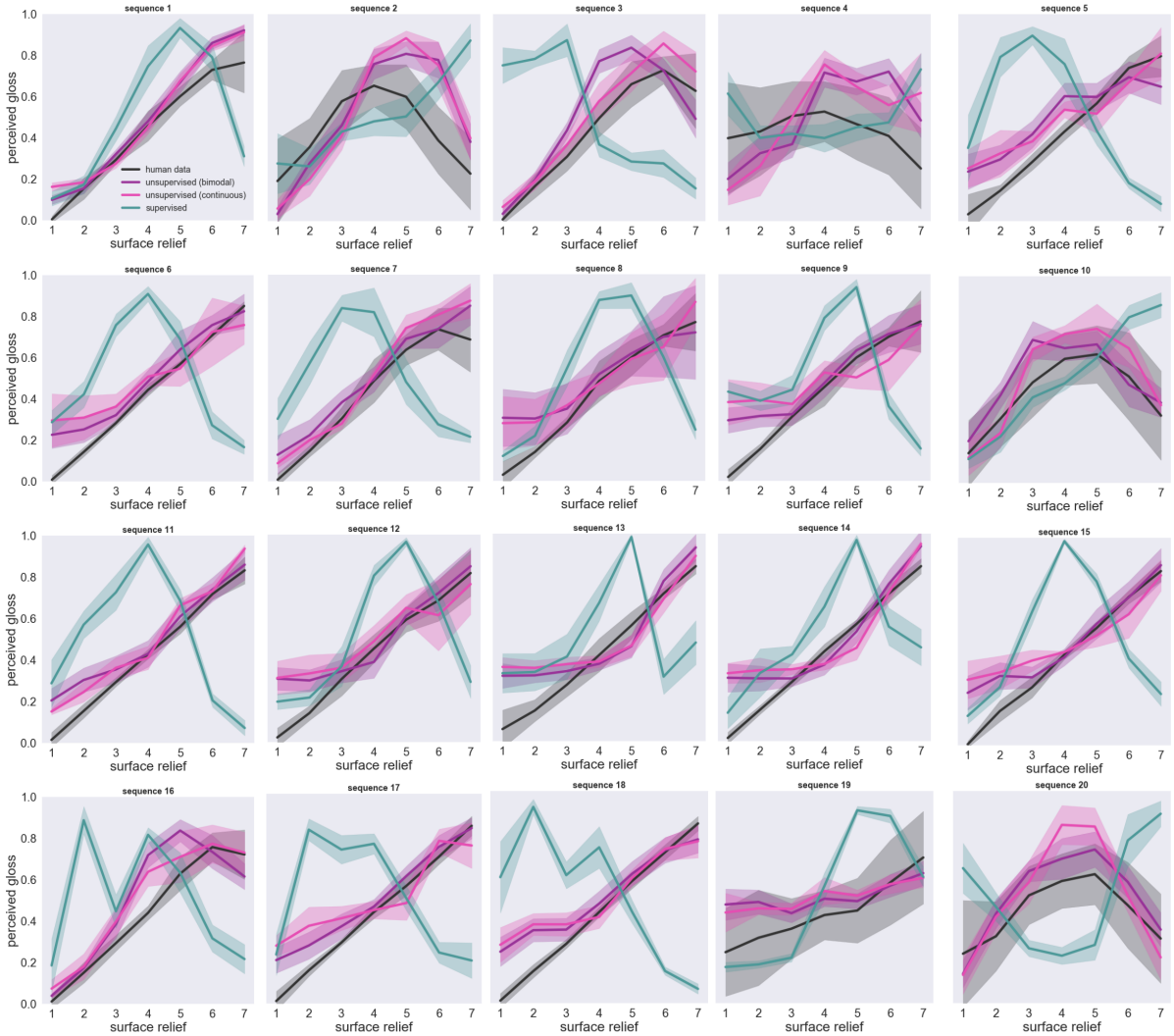
Supplementary Figure 1 [previous page]: Choice of latent dimensionality, and visualisation of single dimensions.

(A) Examples of images generated by networks with latent layers containing 2, 5, 10, 20, and 40 dimensions, using each of the three images in the leftmost column as "seeds" to select a point to sample in the latent space. Sample quality improves as latent dimensionality increases, up to a point.

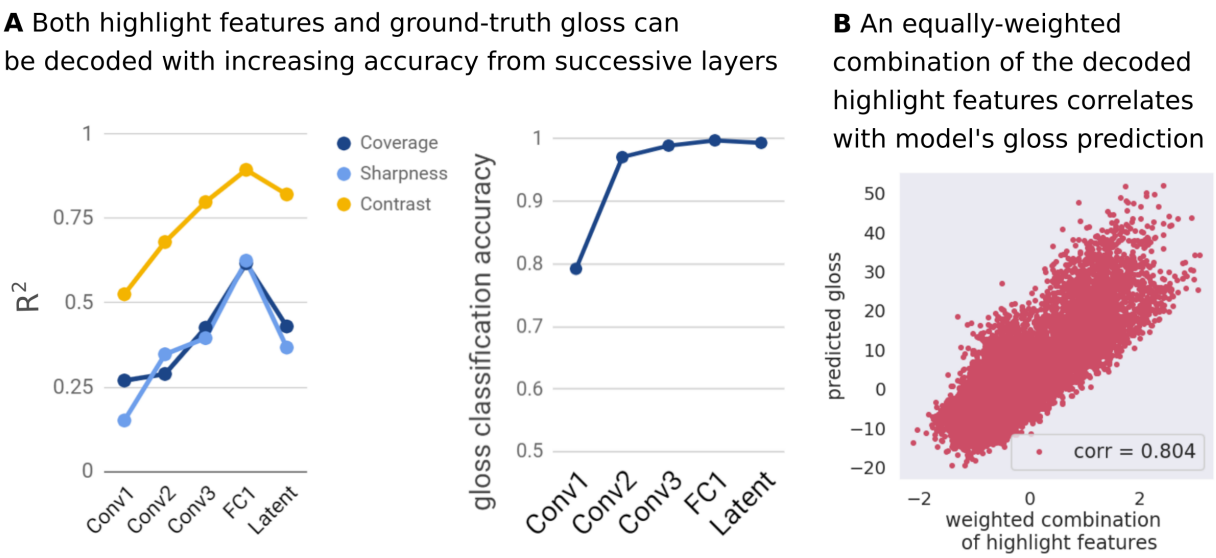
(B) We wanted to select a latent dimensionality for the PixelVAE model such that the network was forced to learn a highly compressed code, but still able to capture the majority of meaningful structure in the training images. To gauge where this dimensionality lay for our dataset, we showed twenty human observers 50 five-image sets like those above, using as seeds 50 random images the networks had not seen during training. For each set, participants were asked to arrange the five images in order of similarity to the original seed image. The plot shows the mean placement of samples (y-axis) from each of the networks (x-axis), for each observer (coloured dots). Line indicates mean over all observers. Above around ten latent dimensions, there are diminishing perceptual quality returns when adding further dimensions, for this image training set.

(C) Example visualisations of three individual dimensions within two different training instances of models with 10 latent dimensions, for two seed images. The central column beneath each seed image shows different samples generated by the network when sampled on the identical point in latent space. To generate the images in the two columns to the left of this, we conditioned on two points slightly negatively shifted, along only the stated dimension; to generate the images in the two right columns, we conditioned on two points slightly positively shifted along the same dimension. Individual dimensions capture variations in shape, colour, material, and lighting, and often combine multiple of these properties (i.e. the representation of world factors is distributed, not sparsely encoded by single dimensions).

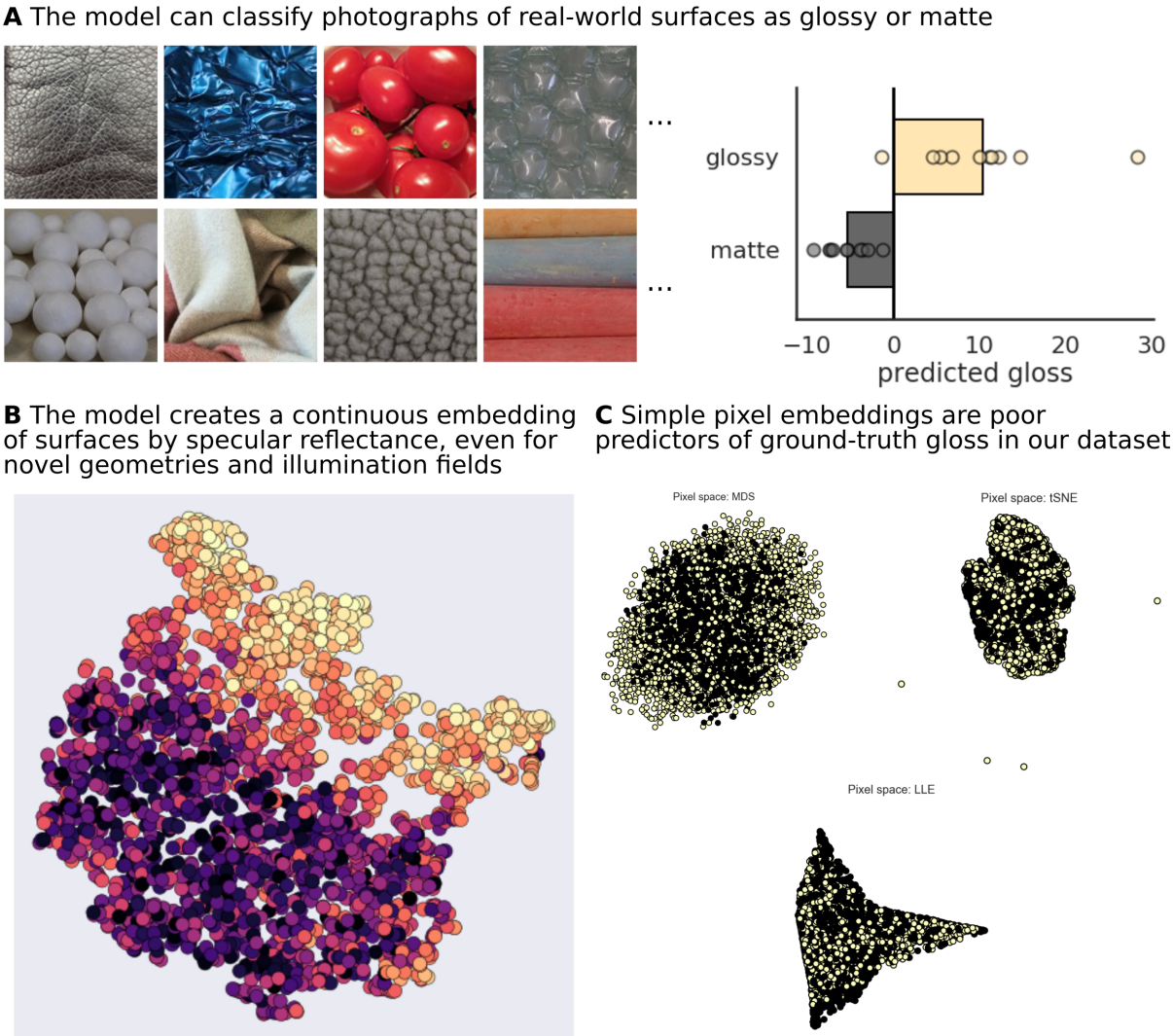
A Human data and model predictions for each test sequence



Supplementary Figure 2: Full data for Experiment 3
(A) Full data from each of the twenty constancy sequences summarised in **Figure 5B-C** of the main text, including predictions from a new set of five PixelVAE models trained on an independent dataset in which gloss varies continuously rather than bimodally ("PixelVAE continuous-gloss" model referred to in main text and figures).



Supplementary Figure 3. Highlight coverage, sharpness and contrast correlate with model-predicted gloss level.
(A) The coverage, sharpness and contrast of highlights can be decoded from representations within one of the trained networks (here, we chose to investigate the model training instance with highest correlation with human-perceived gloss in Experiment 1). Both the decodability of mid-level highlight features (left) and the decodability of gloss category (right) increases throughout the network's layers, until the bottleneck latent code, in which there is a slight decline. (However, note that the dimensionality of the feature space reduces from 1000 units in the final fully connected layer to only 10 in the latent layer, while only minimally affecting gloss classification performance).
(B) Combining (via equal weighting) the estimates of the three highlight features, as decoded from the fully-connected layer immediately preceding the latent layer, yields a summary highlight measure which correlates well with the gloss level predicted from the latent layer.



Supplementary Figure 4: Our unsupervised networks can generalise outside of their training environment

(A) When we input ten photographs of real-world glossy surfaces (four examples shown: leather, metallic foil, tomatoes, plastic), all but one were correctly categorised as high-gloss, based on their latent representations in the unsupervised models (using the gloss SVM classifier trained on the main dataset of 10,000 renderings). Likewise, ten photographs of matte surfaces (examples shown: styrofoam, wool, fleece, chalk) were all correctly classified as low-gloss.

(B) The trained networks organise images lawfully by material even for renderings that fall outside the parameters of the training dataset. Plot shows tSNE embedding of 2500 images generated using five surface geometries and eight lighting environments never sampled from within the network's training dataset (see Methods: *Geometry and lighting generalisation test set)*. Points are colour coded by the strength of specular reflectance (cream = matte; dark purple = highest gloss).

(C) Gloss cannot be trivially decoded in our dataset: no simple non-linear pixel embedding achieves good separation between high-gloss and low-gloss images. Shown are visualisations of 4,000 random images from the training dataset, projected

into two dimensions via (top left) multi-dimensional scaling (MDS), (top right) t-weighted stochastic neighbourhood embedding (tSNE), and (bottom) locally-linear embedding (LLE). High gloss (pale dots) and low-gloss (dark dots) are thoroughly intermixed in all embeddings.