Unsupervised Learning Predicts Human Perception and Misperception of Specular Surface Reflectance

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

Summary

Gloss perception is a challenging visual inference that requires disentangling the contributions of reflectance, lighting, and shape to the retinal image [1-3]. Learning to see gloss must somehow proceed without labelled training data as no other sensory signals can provide the 'ground truth' required for supervised learning [4-6]. We reasoned that paradoxically, we may learn to infer distal scene properties, like gloss, by learning to compress and predict spatial structure in proximal image data. We hypothesised that such unsupervised learning might explain both successes and failures of human gloss perception, where classical 'inverse optics' cannot. To test this, we trained unsupervised neural networks to model the pixel statistics of renderings of glossy surfaces and compared the resulting representations with human gloss judgments. The trained networks spontaneously cluster images according to underlying scene properties such as specular reflectance, shape and illumination, despite receiving no explicit information about them. More importantly, we find that linearly decoding specular reflectance from the model's internal code predicts human perception and misperception of glossiness on an image-by-image basis better than the true physical reflectance does, better than supervised networks explicitly trained to estimate specular reflectance, and better than alternative image statistic and dimensionality-reduction models. Indeed, the unsupervised networks correctly predict well-known illusions of gloss perception caused by interactions between surface relief and lighting [7,8] which the supervised models totally fail to predict. Our findings suggest that unsupervised learning may explain otherwise inexplicable errors in surface perception, with broader implications for how biological brains learn to see the outside world.

Highlights

- We trained unsupervised neural networks to synthesise images of glossy surfaces
- They spontaneously learned to encode gloss, lighting and other scene factors
- The networks correctly predict both errors and successes of human gloss perception
- The findings provide new insights into how the brain likely learns to see

Results and Discussion

The central intuition behind our findings is that learning to compress the complex image structure created by reflections from glossy surfaces into a highly compact code forces the brain to discover representations that partially—but imperfectly—disentangle the distal physical factors responsible for variations within and between images. This potentially explains both the broad successes and specific pattern of errors known to occur in gloss perception [2,3,7-12]. To test this, we rendered 10,000 images from a virtual world of bumpy frontoparallel surfaces with either high or low specular reflectance, random colour and depth of surface relief, illuminated by six natural light fields (**Figure 1A-B**). We trained ten instances of an unsupervised PixelVAE network [13,14] on this dataset. 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. Importantly, this objective refers only to the proximal stimulus (i.e., pixel values) and involves no labels about the distal scene variables (gloss, bump height, illumination, etc.). Despite this, we find the networks not only discover representations that disentangle distal causes that are impossible to tease apart in the raw input (**Figure 2**), but also, more importantly, the internal representations predict human gloss perception and misperception better than the true physical reflectance properties of the surface.

The trained networks can generate wholly novel images, which often appear near-indistinguishable from the training data (**Figure 1D**), by sampling pixel-by-pixel from a learned distribution, conditioned on previous pixels and the values in a highly-compressed 10-dimensional latent code (**Figure 1C**). The representations within this

latent code are the focus of subsequent analyses. As a control model, we also trained ten instances of a supervised ResNet [15] network to classify the same images as high or low gloss, using the ground truth reflectance as training labels. Mean classification accuracy was 99.4% +/-SD 0.001. The supervised network included a 10-dimensional fully-connected layer prior to output, which we treated as its high-level latent code for subsequent comparisons with the unsupervised model (**Figure 1C**).

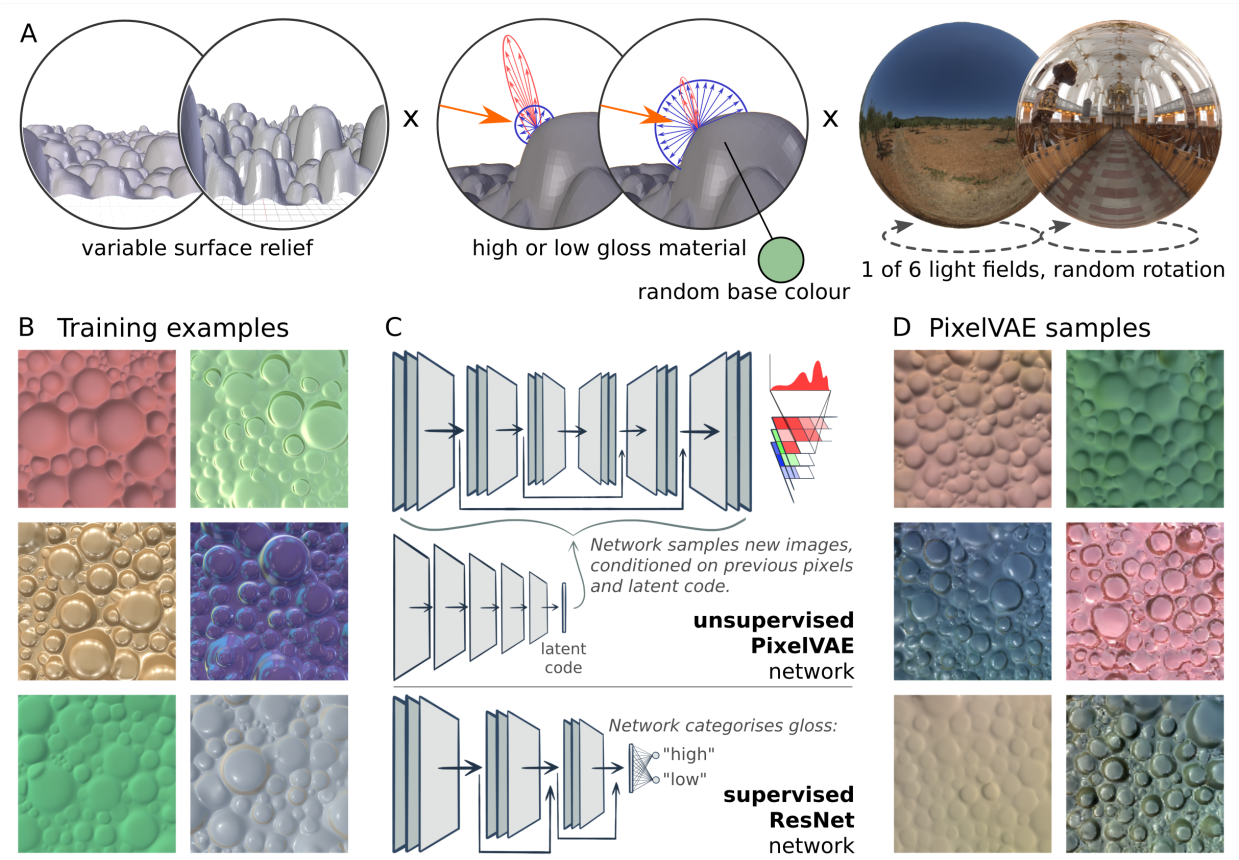


Figure 1: An unsupervised neural network learns to generate plausible novel images from a simulated visual world.
(A) The simulated physical world consisted of a bumpy surface viewed from above. The scene varied in depth of surface relief (left), surface base colour and whether the surface material had high or low specular reflectance (middle), and illumination environment (right).
(B) Six examples from 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 substream 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 also used to train a supervised ResNet network, which outputs 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, to create a representation of equivalent dimensionality to compare 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 a diversity of 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.

The unsupervised PixelVAE model's training objective deals only with proximal image data. Yet we found that distal scene properties spontaneously clustered within the models' 10-dimensional latent codes (cf. [14,16,17]). tSNE visualisation [18] 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 were arranged by the illumination angle, as well as by surface relief, with flatter surfaces occupying nearby points, and bumpier surfaces more distant points.

This rich clustering of images by scene properties was not caused by trivial 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 [19] also failed to separate low- from high-gloss images (**Figure S2D**). Nor is their disentanglement a necessary property of any high-level feature space capable of subserving gloss classification. The same visualisation applied to the penultimate 10D layer of the gloss-supervised models revealed neat

separation of high and low gloss images, but other world factors are highly intermixed (**Figure 2C**). Similar qualitative patterns held for all ten instances of both unsupervised and supervised models.

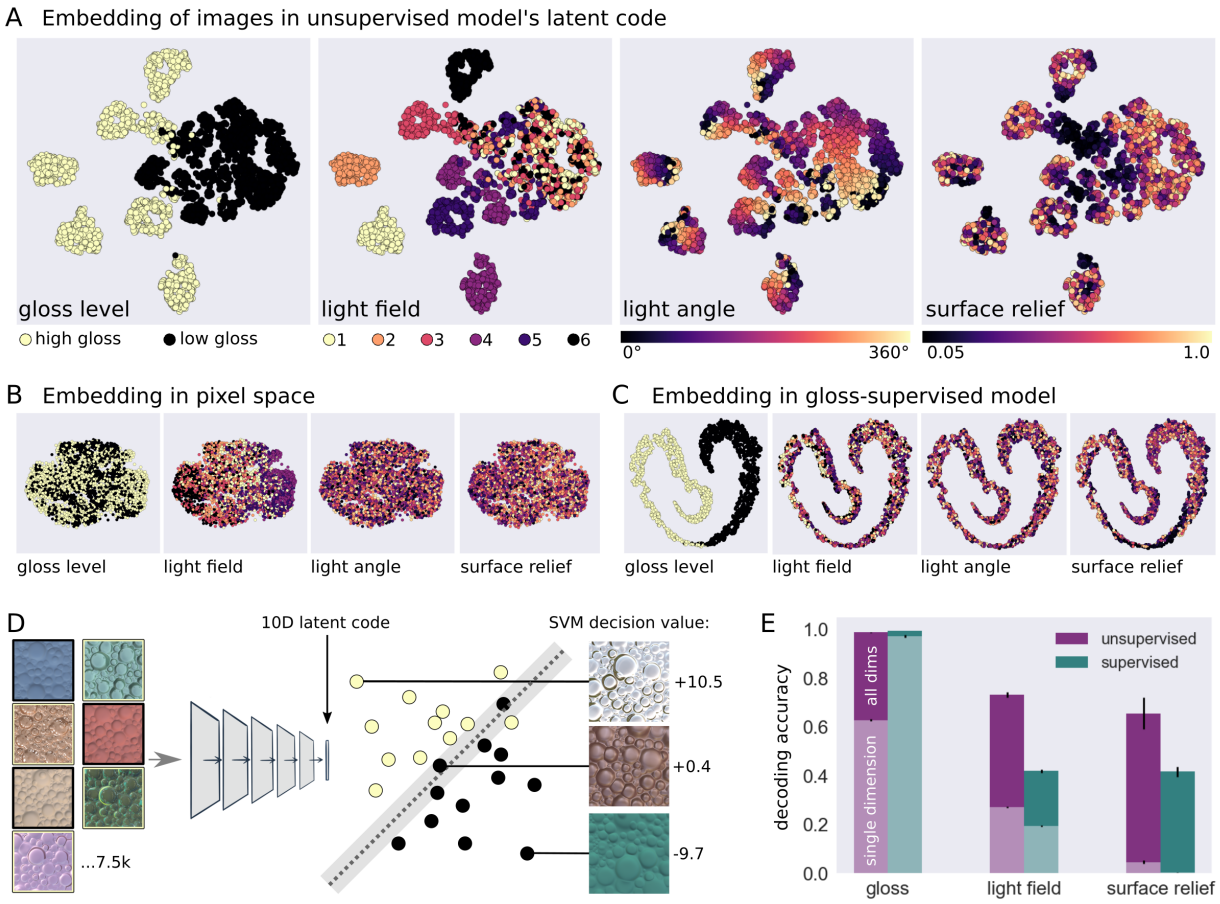


Figure 2: World factors are disentangled in the unsupervised model's latent code.
(A) Visualisations of distances between 4,000 images, in the 10-D latent code of one unsupervised PixelVAE network, projected onto 2D using tSNE [18]. Unsupervised learning spontaneously disentangles the underlying world factors, arriving at a nested hierarchical representation wherein images of low-gloss surfaces form one large cluster (left panel), while images of high-gloss surfaces form multiple small subclusters 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 more similar to one another than those with high relief (right panel).
(B) tSNE visualisation of the same images raw pixel space, where world factors are thoroughly entangled.

(C) The gloss-supervised ResNet model has a far less rich representation in its 10-D latent code. High- and low-gloss images form two clearly-separated clusters, but other world factors are entangled.
(D) To quantify the degree to which each world factor is represented in each model, we trained linear SVM discriminants to classify gloss and light field, and a linear regression to predict surface relief, from the representations of 7,500 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.
(E) 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.

Our central question was whether the unsupervised model could predict human gloss perception. 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. [20-23]). Thus, to derive quantitative gloss predictions from the models, we trained a linear SVM to find the hyperplane in the 10D latent code of each network that best separates high- from low-gloss images (**Figure 2D**). Gloss-classification accuracy for novel images over the ten unsupervised models was 99.3% +/-SD 0.002, and from the 10D latent code of the supervised models 99.4% +/- 0.002 (**Figure 2E**). Light field and surface relief could also be decoded substantially above chance from our unsupervised models (**Figure 2E**), and significantly better than from the supervised models (two-tailed independent-samples t-test comparing light field decoding: $t_{10} = 23.25$, $p = 7\text{x}10^{-15}$; surface relief: $t_{10} = 3.30$, $p = 0.004$). Representations in the unsupervised network were distributed rather than sparse, as a weighted linear combination of all latent dimensions predicted scene properties much better than any individual dimension could (**Figures 2E** and **S1C**). Crucially, 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 2D**). This allowed us to compare the model with human gloss perception.

For Experiment 1 we rendered 50 new images of surfaces with random magnitudes of specular reflectance, drawn 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 3B** 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 = 8.17, p = $1\text{x}10^{-7}$). We considered a number of simple alternative models, the best of which was a linear gloss classifier trained on a multi-scale texture description comprising 1,350 feature dimensions [24], which performed similarly to ground-truth, and significantly worse than the unsupervised model (t = 8.00, p = $2\text{x}10^{-7}$). Others performed poorly, including a gloss classifier trained on a 10-dimensional embedding of images via tSNE, MDS or LLE, and luminance histogram skewness, a measure previously proposed to predict human gloss perception [25] (**Figure 3B** bargraph).

Since PixelVAE networks are generative models, we can also create novel images by pixelwise sampling from them (**Figure 1D**). In Experiment 2 we tested whether it was possible to manipulate the perceived gloss of generated images by shifting the location in the latent space on which their sampling was conditioned, along the axis orthogonal to the gloss-separating hyperplane in 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, and another four sequences moving from high to lower gloss points (**Figure 3A**).The same 20 observers sorted the images within each sequence from least to most glossy. **Figure 3C** 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}).

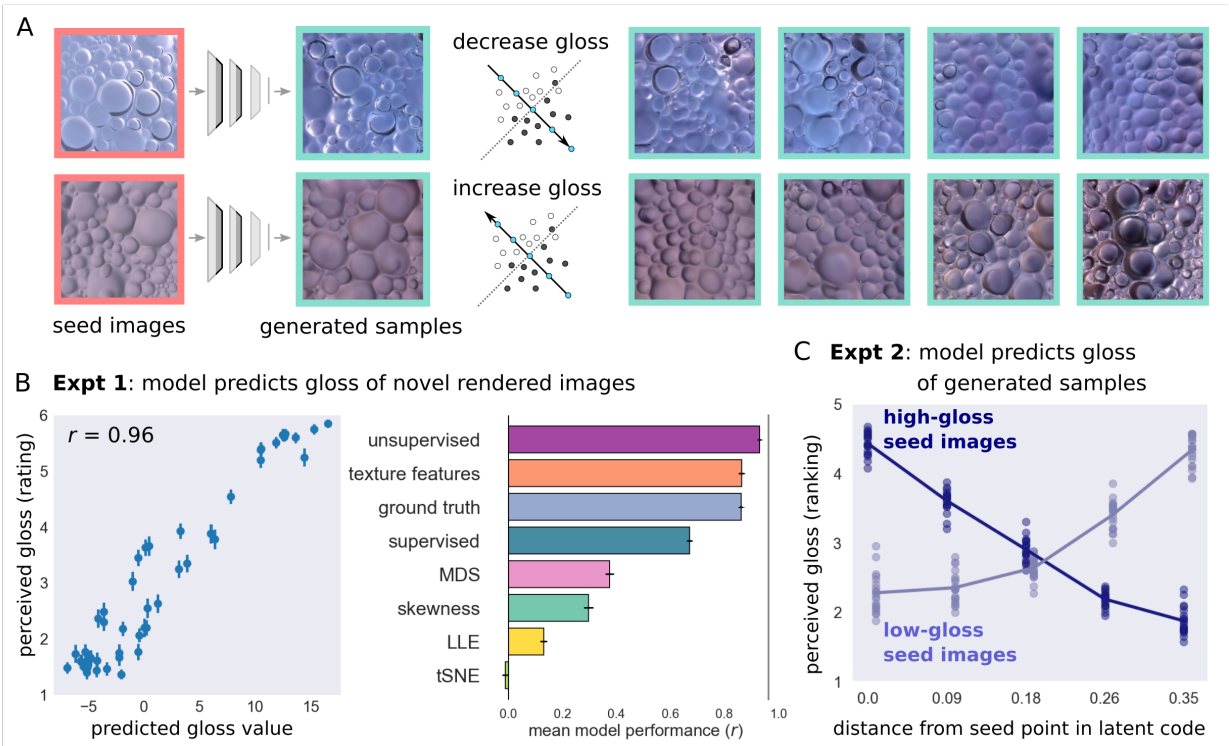


Figure 3: The unsupervised model predicts perceived gloss of new surfaces, and can be used to modulate perceived gloss of generated samples.
(A) Two example sequences of images sampled from a PixelVAE network. A glossy or matte rendered image (left) is input to the network and its latent representation is recorded. This seed point in latent space is resampled after stepping by small amounts along the model’s gloss-discriminating axis (i.e. the axis orthogonal to the decision plane of an SVM trained to classify gloss from the latent code), either towards the “matte” or “glossy” directions. Modulations affect apparent gloss while keeping other factors, such surface colour, approximately fixed.

(B) Gloss values in the unsupervised networks well predict human gloss ratings for 50 novel rendered surfaces of random gloss levels. Scatterplot shows gloss decision value for each test image in the PixelVAE networks on the x-axis (averaged over the ten training instances of the model), and human gloss ratings for the same images on the y-axis (1-6 scale). Barplot shows average correlation to each human participant's ratings for the unsupervised model and for diverse alternative predictors (from best to worst): linear gloss classifier based on texture features [24]; ground truth (i.e. the rendering setting of specular magnitude); the gloss-supervised networks; gloss classifier based on MDS embedding of images; skewness of the image luminance histogram [25]; and gloss classifiers based on LLE and tSNE embeddings of images. Vertical grey line indicates how well individual human ratings can be predicted from the data of other participants, giving the maximum possible model performance. Error bars: standard error of the mean over participants.
(C) Moving along the gloss-discriminating axis in a network's latent code as shown in (A) 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 gloss ratings given to each modulation step by each individual participant, averaged over four test sequences using low-gloss initial seed images, and four using high-gloss seed images. Lines indicate mean over all participants.

Importantly, while human gloss perception is generally good, it also exhibits some well-documented systematic errors. For example, bumpier surfaces tend to look glossier than flatter surfaces, and specific combinations of lighting and surface relief reliably yield specific patterns of misperception (**Figure 4A**; [7,8]). A key test of any computational model of biological vision is that it mimics perceptual errors as well as successes. We therefore used systematic failures of gloss constancy to decouple human perception from true scene properties, and rigorously test competing models.

To do this, we measured perceived gloss within rendered sequences in which reflectance and other scene properties were fixed, while surface relief increased in seven steps (**Figure 4A**). If observers correctly estimated reflectance, all surfaces should appear equally glossy, yet they do not. Across Experiments 3a and 3b, two groups of 20 observers provided gloss judgements for two sets of ten such sequences via a paired-comparison task (**STAR Methods**). To maximally distinguish between the

models, we selected sequences for which the unsupervised and supervised models made strongly divergent predictions.

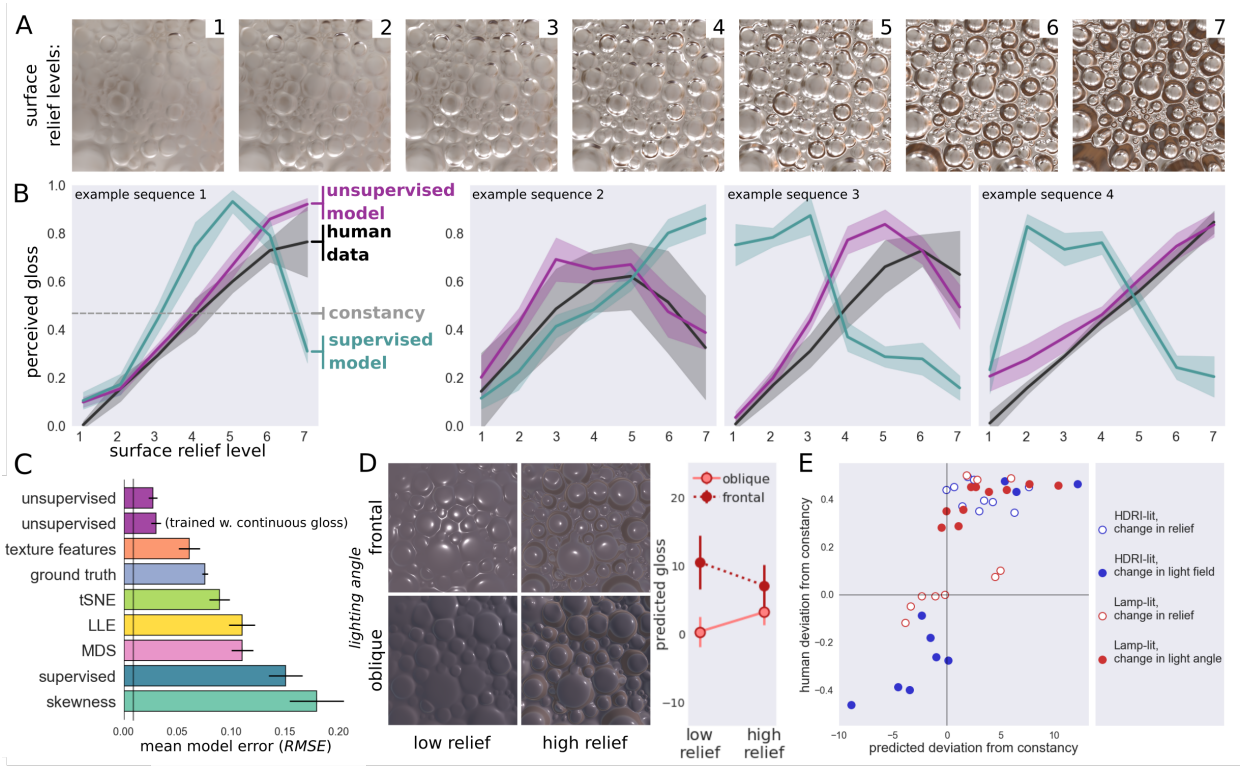


Figure 4: Unsupervised learning predicts human perception and misperception of gloss on an image-by-image basis.
(A) Material properties of surface are identical in all seven images, but surface relief increases. People often see bumpier surfaces as glossier than flatter ones, constituting a failure of gloss constancy [7,8].
(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 each image step reported glossier than others in sequence); Purple: prediction of unsupervised model; Teal: predicted of supervised model. Veridical gloss perception would be flat horizontal line. Shading indicates standard deviation over 20 human participants or 10 model instances.
(C) Summary of average goodness of fit (RMSE) of each model to each participant's data, averaged across all twenty image sequences From best to worst: unsupervised model; unsupervised model retrained on a new dataset of 10,000 images within which gloss is sampled continuously (also see **Figure S3**); texture feature model; ground truth (i.e. perfect constancy); tSNE image embedding; LLE image embedding; MDS image embedding; supervised model; luminance histogram skewness. Vertical grey line indicates how well individual human ratings can be predicted from the data of other

participants, giving the minimum possible model RMSE. Error bars: standard error of the mean over participants.
(D) Changes in lighting also often change apparent material properties for human observers [7,8]. The quartet of images shows same surface material rendered on 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 B of a pair seen as glossier than image A). 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 averaged over 20 observers. X-axis: difference in gloss values between the images in each pair, averaged over 10 instances of PixelVAE. Performances of all alternative models is shown in **Figure S3B**.

For almost all sequences, 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 bump heights (four examples shown in **Figure 4B;** complete data: **Figure S3A**). The unsupervised model, despite never being explicitly trained to represent gloss, and never being fit to any human data, predicted patterns of failures of gloss constancy remarkably well (**Figure 4C**, 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, and for almost all sequences made predictions that were anti-correlated with human patterns (median R^2 = -1.36). Predictors based on luminance histogram skewness [25], ground truth, and tSNE, MDS or LLE image embeddings all performed poorly (R^2 values between -0.07 and -1.58). A gloss classifier based on texture features [24] provided the next best performance to the unsupervised model (median R^2 = 0.54),

but was significantly worse (two-tailed repeated-measured t_{19} = 5.66, p = 4x10^{-6}; see **Figure S3B**).

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 4D**; **STAR 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; **STAR Methods**). Twenty observers indicated which surface in each pair appeared glossier, with each pair repeated eight times. **Figure 4E** shows that, for these image with diverse scene properties, the unsupervised model predicted the degree and direction of human (failures of) constancy (R^2 = 0.55, p < 5x10^{-8}), and outperforms all alternative models (**Figure S3B**). Fitting a simple logistic function to relate model and human gloss values further improves the prediction (R^2 = 0.73).

Together, our findings show that human perception and misperception of gloss can be predicted on an image-by-image basis from the rich and flexible latent code discovered by an unsupervised neural network trained to model the pixel statistics of images. This suggest that rather than framing perception as the optimal *estimation* of physical quantities via 'inverse optics' [26-29], it may be more useful to formulate vision as a data-driven process of *disentanglement* in a multidimensional feature space, in which non-linear transformations reorganise inputs into new configurations that make reading

out distal properties easier [6,20]. It is the imperfect nature of the disentanglement that yields the characteristic errors humans make.

Previous research [8] 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 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 (**Figure S4A**). A linear combination of the three cues in the layer immediately preceding the latent layer correlated with gloss predicted from the latent layer (r = 0.80; **Figure S4B**). Manipulating the highlights in images to weaken each cue also reduced predicted gloss (**Figure S4C**). 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 S4C)**. This manipulation is known to decrease perceived gloss [8,30], yet it preserves coverage, contrast and sharpness. This suggests that failures of gloss constancy may be largely explainable in terms of mid-level highlight features, despite our sensitivity to richer photogeometric constraints.

The composition of the training dataset has profound effects on all machine learning models. To test the robustness of our model's predictions to different training datasets, we trained five new PixelVAEs on 10,000 additional renderings in which gloss was sampled continuously rather than bimodally. These new models yielded excellent gloss classification on the original bimodal gloss dataset (mean accuracy = 96.7% +/- SD

1.3). More importantly, the gloss constancy predictions of the new networks were statistically indistinguishable from those of the original models (**Figure S3A-B**; two-way repeated-measures ANOVA $F_{(6,114)}$ *interaction* = 0.81, p = 0.56, *ns*), and predicted human data equally well (Figure 4C; repeated-measures t-test t_{19} = -1.76, p = 0.10, *ns*). This indicates the results are not an artefact of the distribution of gloss values in the training set.

Despite the constrained nature of the training set, our models generalise well across changes in simulated scene properties such as lighting and 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 S2B**). We could well predict, using regression, the level of gloss in the continuously-varying images (mean R^2 = 0.79 +/- SD 0.03), and prediction for the four new geometries was at least as good as for the familiar geometry (see **Figure S2B**). A tSNE visualisation of the 2,500 images' embedding in latent space indicated a strong organisation by surface specularity (**Figure S2C**). Most remarkably, we found that the models can even categorise close-up photographs of diverse real-world surfaces as being of high or low gloss (**Figure S2A**).

Our tight control over the statistics of the training environment allowed us to guarantee that surfaces could not be disentangled in the raw input space, and that variations in distal properties were strictly independent of one another. Our model's success at replicating human failures of gloss constancy therefore suggests that these do not arise from statistical correlations between distal factors, as has been proposed to explain certain illusions [31].

It is important to emphasize that the PixelVAE model is not intended as a physiological simulation of visual cortex, but as a proof of principle that unsupervised objectives can discover distinctively human-like visual representations. The computer vision literature boasts many recent impressive results using diverse unsupervised and self-supervised learning objectives. For example, rich image structure can be learned by predicting novel viewpoints [32]; by “curiosity-driven” active learning [33,34]; by reinforcement learning [35,36]; and by adversarial learning [37]. However, the brain cannot learn raw pixel statistics, as cortical inputs are already significantly filtered and processed. There remain many intriguing questions: how does optical and retinal processing of images [38,39] influence the types of unsupervised objectives used by brains? How is this learning shared across neural sites and sensory modalities, and how do the statistics of our infant environments [40] shape learning?

Unsupervised learning objectives such as data compression and spatial prediction are ecologically plausible ways for biological visual systems to learn, and offer a solution to the problem of how brains seem to represent properties of the distal world without access to ground truth [3,5,6,41,42]. Our work shows for the first time that perceptual dimensions, like gloss, that appear to (imperfectly) estimate properties of the physical world can emerge spontaneously by learning to efficiently encode sensory data. This suggests a novel interpretation of failures of perceptual constancy: they are misalignments between the dimensions of image variation that are learned by unsupervised processes, and the dimensions of variation in underlying physical properties. Failures of constancy, which are rife in vision, therefore offer tell-tale clues to how we learn to see. Indeed, we believe previous research emphasis on individual

cues can be greatly supplemented by understanding the learning processes that discover the relevant regularities in images in the first place.

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 for comments on an earlier version 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

STAR Methods

KEY RESOURCES TABLE

RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Raw data	This paper	tbd
Stimuli	This paper	tbd
Software and Algorithms		
Unity 3D	Unity Technologies	https://unity.com/
Python	Python Software Foundation	https://www.python.org/
PsychoPy	Jonathan Peirce	https://www.psychopy.org/
Tensorflow	Google Inc.	https://www.tensorflow.org/
MATLAB	Mathworks Inc.	https://www.mathworks.com/
PixelVAE network implementation	Zhao, Song & Ermon (2017)	https://github.com/ermongroup/Generalized-PixelVAE
ResNet network implementation	Alexia Xu	https://github.com/wenxinxu/resnet-in-tensorflow
MDS, tSNE and LLE algorithms	Scikit-learn	https://scikit-learn.org/
Highlight feature measurement model	Schmid & Doerschner (2019) [43]	doi:10.1167/19.10.252
Texture analysis algorithm	Portilla & Simoncelli (2000) [24]	https://www.cns.nyu.edu/~lcv/texture/

CONTACT FOR RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Katherine Storrs (katherine.storrs@gmail.com).

EXPERIMENTAL SUBJECT DETAILS

Three groups of 20 human naïve observers reported perceived gloss across the five psychophysical experiments. In Experiments 1 and 2, participants were 16 women and 4 men with a mean age of 23.5. In Experiment 3a, participants were a different group of 16 women and 4 men, also with a mean age of 23.5 In Experiments 3 and 4, participants were 14 women and 6 men with a mean age of 24.3. 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 design or hypothesis of the experiments after the first session, so remained naïve. Experiments were conducted in accordance with the Declaration of Helsinki and prior approval was received from the local ethics committee of Justus-Liebig University Giessen. Volunteers were paid €8 per hour in compensation for their time.

METHOD DETAILS

Stimuli

Stimuli were images of bumpy surfaces rendered at a resolution of 800x800 pixels using the Unity3D graphics engine (version 2018.2.3). A virtual 40cm x 40cm sheet with irregularly positioned bumps was placed in a scene and illuminated by one of six HDRI environmental light fields. Light fields were white-balanced light-probe images of four exterior (beach, bay, woodland, savanah) and two interior (church, conference hall) environments at a resolution of 8192x4096 pixels, captured by Unity Technologies [44]. A camera was located 15cm above the sheet, viewing directly downwards onto the surface. By randomly varying the location of the camera in the horizontal plane, and the rotation of the camera around the vertical axis, an extremely large number of unique images could be generated from this simple scene.

Main training dataset

A set of 10,000 images was rendered to create the training, validation, and test datasets for neural network models. On each rendered image, one of the six HDRIs was randomly selected to illuminate the scene. Surface relief was scaled multiplicatively in depth with a scaling factor that varied randomly uniformly, so that the

highest point on the surface was between 0.1cm and 2.5cm high. Camera location and rotation was varied as described above. Material properties of the surface were controlled via the Unity3D standard shader, using the “specular setup,” in which appearance is determined by setting a base albedo, a magnitude between 0-1 for the specular reflectance component, and a smoothness/concentration value between 0-1 for the specular component. The base colour of the surface material was chosen by randomly selecting R, G, and B 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, the magnitude of the specular reflection component was selected randomly uniformly in the interval 0.1 to 0.3, and the concentration of the specular component was drawn randomly uniformly between 0.2 and 0.4. For high gloss surfaces, specular magnitude was drawn randomly uniformly between 0.3 and 0.5, and specular concentration was drawn randomly uniformly between 0.75 and 0.95. The training dataset therefore consisted of a bimodal distribution of half low- and half high-gloss surfaces, with some variation in reflectance properties within each of those groups. The same dataset was also used when training classifiers to decode world factors from latent representations in models.

Continuously-sampled gloss training dataset

To verify that the representations learned by our models were not an artefact of the bimodal gloss sampling (either high-gloss or near-matte) in the main dataset, we rendered a separate set of 10,000 images in which specular magnitude was sampled randomly uniformly from the interval 0.1 to 0.8, and specular concentration was set to the same value as magnitude. All other scene properties were varied as in the main dataset.

Geometry and lighting generalisation test set

To test the generalisation abilities of our models, we rendered test sets of 500 images each using the original surface sheet, and using four novel surface geometries. The novel geometries were virtual 40cm x 40cm sheets with different shapes and sizes of irregularly placed lines, bumps, and indentations (close-ups of the four geometries are shown in **Figure S3B**). Illumination was selected at random for each scene from a novel set of eight 4096x2048 pixel HDRI light fields (six exterior, two interior) from the Dutch Skies 360º HDRI Project (http://www.dutch360hdr.com/). As in the continuously-

sampled gloss training set, specular magnitude was sampled from the range 0.1 to 0.8, and specular magnitude was set to the same sampled value.

Experiment 1: Renderings with continuously sampled gloss levels

Stimuli were 50 new rendered images for which a specular magnitude was chosen randomly uniformly in the interval 0.2 to 1.0, and specular concentration was set to the same value as specular magnitude. All other attributes were sampled as in the training dataset.

Experiment 2: Gloss-modulated images generated from PixelVAE models

Four low-gloss and four high-gloss "seed images" from the training dataset were input to each of the ten trained PixelVAE models, after having determined the axis in that model's latent space along which gloss could be most strongly discriminated (see **Model analysis**). From each network, one new image was sampled conditioning on the point in latent space corresponding to the seed image. Another four images were then generated, each time conditioning on a point 0.07 units in latent space along that model's gloss axis, either in the matte (for high-gloss seed images) or glossy (for high-gloss seed images) direction. This resulted in 400 stimulus images (10 models x 5 steps x 8 seed images).

Experiments 3a and 3b: Sequences of renderings with increasing surface relief

For each experiment, we rendered a different set of 200 candidate sequences of 7 images with fixed random material, illumination, and camera position, but exponentially increasing surface relief, corresponding to maximum bump heights of approximately {0.10, 0.18, 0.31, 0.55, 0.97, 0.70, 3.0} cm. All surfaces had relatively high gloss. In Epxeriment 3a, specular magnitude was selected randomly uniformly from the interval 0.2 to 1.0, and specular concentration randomly uniformly from the interval 0.7 to 1.0. In Experiment 3b, magnitude was selected from the interval 0.1 to 0.3, and concentration from the interval 0.8 to 1.0. Other attributes were sampled randomly as in the training dataset. All 1400 images were then input to each of the ten trained PixelVAE (unsupervised) and ResNet (supervised) models, as well as calculating their representation within each of the alternative feature space models, and a predicted gloss value for each image for each model was obtained (see **Model analysis**). For each candidate sequence of 7 images, we then simulated the responses from each model under a two-alternative forced-choice (2AFC) experiment in which each possible

pair of images was compared, and the model selected the one with the highest predicted gloss value. This yielded a distribution of predicted 2AFC experimental results from each model.

For the psychophysical experiments, we sought to select from among the candidates sequences that had (a) clear differences between the constancy patterns predicted by the unsupervised vs supervised models, and (b) were likely to produce a variety of patterns of failures of constancy in human observers. First, we sorted sequences according to the strength of the interaction between unsupervised and supervised model predictions (as quantified by the *F* interaction statistic of a two-way ANOVA), to prioritise those with strong disagreements between model predictions. 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, from a different candidate set. After ranking by model disagreement, we classified sequences into four broad qualitative groups based on the pattern of data predicted, by fitting polynomial functions to the average unsupervised model predictions: (1) those for which the model predicted approximate constancy (sequences for which the range of "proportion more glossy" values predicted was < 0.25, and a linear function fit achieved $R^2 > 0.70$); (2) those for which the model predicted an approximately linear increase in gloss with bump height (range of data > 0.50, and a linear fit achieved $R^2 > 0.90$); (3) those for which the model predicted 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); and (4) those for which the model predicted an upward-swinging nonlinear failure of constancy (range > 0.50, and a quadratic fit of $R^2 > 0.90$ with a positive coefficient of the squared term). From each pattern group we then selected a number of sequences N roughly proportional to how common that pattern was among the candidate sequences (1 from the "constant" group, 4 from the "linear" group, 2 from the "non-monotonic" group, and 3 from the "nonlinear" group), by taking the top N sequences of that type with the highest interaction terms.

Experiment 4: Pairs of renderings with fixed surface materials

We rendered 800 candidate pairs of images with fixed material and camera position, but differing in either illumination or surface relief. For 400 pairs, illumination was

identical, but surface relief was low (approx. 0.40cm maximum bump height) in one image and high (approx. 1.5cm maximum bump height) in the other. Within these, 200 pairs were illuminated by one of the six HDRI light fields used for the training dataset, and 200 pairs were illuminated by a directional lamp at an angle of either 30° (oblique) or 90° (frontal). The other 400 pairs had fixed bump height (either both low or both high) but differed in lighting. Within these, 200 pairs differed in which of the six HDRI light fields they were illuminated by, and 200 pairs were lit by a directional lamp and differed in the angle of illumination (either 30° or 90°). All 800 images were input to each of the ten trained PixelVAE models, and an average predicted gloss value for each image was obtained (see **Model analysis**).

For the psychophysical experiment, we sought stimulus pairs (a) with a diverse variety of properties, i.e. changing either in surface relief or lighting, and lit by either natural or lamp illumination, and (b) for which the unsupervised PixelVAE model made a wide range of predictions, ranging from good constancy to strong failures of constancy. To achieve this balanced stimulus set, we divided the candidate pairs into four groups: (1) naturally lit, differing in surface relief; (2) lamp-lit, differing in surface relief; (3) naturally lit, differing in light field; and (4) lamp-lit, differing in illumination angle. Within each of these groups, we ranked pairs by the size of the absolute difference in gloss values predicted for the two images in the pair, and then selected the ten images from each group lying at each of the 10th percentiles. This yielded a set of 40 test pairs which were well balanced in scene properties, and which spanned a wide range of predicted constancies.

Real-world images

To test generalisation to real-world images (**Figure S2A**), we took 10 colour photographs of close-ups of common high-gloss surfaces (e.g. tomatoes, plastic, metallic foil) and 10 of matte surfaces (e.g. chalk, fleece, styrofoam).

Highlight-feature manipulated images

To manipulate highlight features (**Figure S4C**), we rendered images of ten scenes with high-gloss surfaces, once using the full material reflectance description, and once after setting the specular magnitude and concentration parameters to zero, to obtain only the diffuse component. We subtracted the second image from the first to obtain the specular component, which we manipulated in four different ways: (1) To create

translated highlight stimuli, we simply shifted the specular component rightwards in ten steps of 5 pixels, before recombining the specular and diffuse components by addition. To avoid edge artefacts after translation, the original images were rendered at 1024x1024 resolution, and all final stimuli consisted of the lower right 800x800 pixel section. (2) To reduce the coverage of highlights, we used the image erosion algorithm in the OpenCV package for Python. Erosion with a kernel size of 2x2 was applied to the specular component for ten iterations to create ten progressively reduced highlight maps, before recombining with the diffuse component. (3) To reduce the sharpness of highlights, we convolved the specular image with a gaussian filter of ten increasing step sizes, from 1x1 to 11x11, before recombining with the diffuse component. (4) Finally, to reduce the contrast of highlights, we first created a “highlight mask” by identifying those pixels which had values of 4 or higher (in a 0-255 range) in a greyscale version of the specular-only image (arbitrary threshold chosen by visual inspection). We then reduced the standard deviation of brightness values of pixels within the highlight map in ten steps, from 1 to 0.1 times their original SD, while retaining their original mean, before recombining with the diffuse component.

Neural network architectures and training

All neural network implementation and training was performed using Python 2.7 with Tensorflow 1.14 on a GPU-accelerated machine using 1-4 GTX1080 graphics cards. All networks were trained on the first 9,000 images of the main dataset, using the next 500 for validation during training, and the final 500 for testing accuracy of the trained networks (in the case of the supervised networks). In order to bring training time and memory resources into a reasonable range, all training and test images were downsampled to a resolution of 128x128 pixels. For both unsupervised and supervised model types, ten independent instances of the same network architecture were trained, using identical hyperparameters but different random network weight initialisations and different random sampling of batches during training. In addition, we trained another five instances of the PixelVAE network on the continuously-sampled gloss dataset, again using the first 9,000 for training, and the next 500 for validation. No data augmentation was used. Both network architectures used rectified linear activation functions.

Unsupervised PixelVAE model

We used the implementation from [14] of the PixelVAE architecture [13], 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. The other stream is an autoregressive PixelCNN++ model that learns the structure of the training data in terms of a polynomial probability distribution over pixel level 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 row by row from the top left to the bottom right of the image, by sampling from its learned probability distribution, conditioning on both the values of pixels generated so far (which constrain the local structure of the part of the image currently being sampled), and on the values provided in the latent code (which constrain overall image properties).

In our models, the conditioning network consisted of three convolutional layers of 64, 128, and 256 feature maps, respectively, followed by a fully connected layer of 512 units and a 10-dimensional latent code layer. The autoregressive network consisted of six "residual blocks" of three convolutional layers of 64 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 began to plateau. No regularisation was used during training, to encourage the autoregressive image-generating network to depend on information in the latent code [14].

Supervised ResNet model

We used the Tensorflow implementation of the ResNet architecture [15] from https://github.com/wenxinxu/resnet-in-tensorflow. Our networks consisted of three "residual blocks" each consisting of three layers, with one layer comprising batch normalisation, followed by a nonlinearity, and a convolution using 56 feature maps. Skip connections in the ResNet architecture allow information to circumvent whole blocks if useful to the training objective. The output of the final convolutional layer was passed to a 10-dimensional fully-connected layer, which we treated as the "latent code"

of the model, and this passed, via one more nonlinearity to a two-unit softmax output layer. The network was trained to classify images as high or low gloss, based on the rendering parameters of the specular component (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 true labels. Networks were trained for 75 epochs, by which point validation accuracy plateaued above 99%.

Procedure

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 2560x1440 and a refresh rate of 60Hz. Python 3.6 and PsychoPy 3.1 were used to present stimuli and record responses. At the beginning of each experiment, the observer was shown a screen with 4-12 examples of the experimental stimuli, in random arrangement, to familiarise them with the appearance and range of material properties they would be asked to judge during the task. In all experiments, participants could take as long on each trial as they wished, and stimuli continued to be displayed until a response was recorded.

Experiment 1: Gloss ratings for novel rendered images

Fifty images were presented singly in the centre of the screen at a resolution of 800x800 pixels (18.6cm x 18.6cm on screen). Each image was repeated three times, for a total of 150 trials, which were presented in a different random order for each participant. Participants were asked to rate the apparent gloss of each image using a 6-point scale with the endpoints labelled “1 = completely matte” and “6 = extremely glossy”.

Experiment 2: Gloss rankings for gloss-modulated network samples

Eighty sets of five PixelVAE-generated images were presented, with each image shown at the network-output resolution of 128x128 pixels (2.9cm x 2.9cm on screen). 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 centre row of 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 stimuli

Experiments 3a and 3b differed only in the specific sequences shown to participants. Both experiments consisted of 10 sets of 7 image steps, from all possible pairs were shown twice each, for a total of 420 trials. Images appeared at a resolution of 800x800 pixels (18.6 x 18.6cm), side by side on screen. Side of screen was randomised on each trial, and trial order was 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 stimuli

Each of 40 pairs of images were shown 8 times, for a total of 360 trials. Images appeared at a resolution of 128x128 pixels (2.9 x 2.9cm), side by side, with side of screen randomised on each trial, and order of pairs randomised. As in Experiments 3a and 3b, participants reported whether the surface depicted on the left or right appeared more glossy.

QUANTIFICATION AND STATISTICAL ANALYSIS

All analyses of data from models and humans were performed in either Python 2.7 or 3.6, using the numpy v1.16.5 and/or scikit-learn v0.21.3 packages.

Model analysis

tSNE

The clustering of images in the latent spaces of trained models and pixel space (as shown in **Figure 2A**, **B** and **C** and **Figure S2D**) was visualised using t-weighted Stochastic Neighbourhood Embedding (tSNE; [18]) to create a 2D embedding of the locations of 4,000 raining dataset images in the respective latent spaces. Default tSNE parameters, as implemented in the scikit-learn package, were used for all visualisations.

In order to quantify how well a tSNE-embedding of raw images could subserve gloss classification, and to obtain tSNE-based gloss predictions for the experimental images, a 10-dimensional tSNE embedding was performed on a set of images consisting of

8,000 of the training images, as well as all test images from the four experiments (770 images in total). A linear Support Vector Machine (SVM, as implemented in the scikit-learn package) was trained on the resultant 10D representations of 6,000 of the training dataset images, and tested on the remaining 2,000 to obtain gloss classification accuracy (barchart in **Figure S2D**). The decision values of the 10D representations of the experimental test images, using the trained SVM, were taken as predicted gloss values for model comparison purposes (**Figure 3B** and **4C**, and **Figure S3B** and **C**).

MDS and LLE

As for tSNE, the visualisations of pixel embeddings in **Figure S2D** were created by embedding 4,000 training images in two dimensions, using either the metric Multidimensional Scaling (MDS) or Locally Linear Embedding (LLE; [19]) algorithms as implemented in scikit-learn. As for tSNE, model predictions of perceived gloss (**Figure 3B** and **4C**, and **Figure S3B** and **C**) were obtained by training a linear SVM to classify gloss based on a 10D embedding of 8,000 training images, and taking the signed decision values for test images embedded in the same space.

Histogram skewness

Skewness was defined simply as the skew (i.e. third moment) of the distribution of pixel intensities in a greyscale version of an image. Raw skew values were used as gloss predictors for the purposes of model comparison (**Figure 3B** and **4C**, and **Figure S3B** and **C**). Ground-truth gloss classification accuracy (barchart in **Figure S2D**) was defined as the accuracy using an optimal threshold to separate high- from low-gloss images in our dataset from one another on the basis of their skewness, fitting the threshold on 7,500 images and testing on the remaining 2,500.

Texture features

For each image we calculated a multi-scale texture feature description [24] using the TextureSynth package for MATLAB (www.cns.nyu.edu/~lcv/texture/). Images were rescaled to 256x256 pixels, and analysed at 3 scales, 4 orientations, with a spatial neighbourhood of 7, which produced a description of each image in terms of 1,350 feature dimensions. Gloss classification accuracy (barchart in **Figure S2D**) was calculated by training a linear SVM to separate high- from low-gloss images based on 7,500 of the training dataset images, and testing on the remaining 2,500. Gloss

predictions (**Figure 3B** and **4C**, and **Figure S3B** and **C**) were calculated by taking the signed decision value of this classifier, when applied to the texture feature description of test images.

Decoding world factors from neural network models

For unsupervised and supervised models (bars shown in **Figure 2E**), we used a random sample of 7,500 of the training dataset images to train linear SVMs to classify gloss level (2-way classification) or light field (6-way classification), and the remaining 2,500 images to evaluate classification accuracy. For the continuous factor of surface relief, we fitted a linear regression model using 7,500 images and measured R^2 for the remaining 2,500. All classifications were performed using the full 10-dimensional latent space for each model training instance, and also using each of the single individual latent dimensions for each model, to assess how strongly information about world factors was contained within individual dimensions of the latent codes (see **Figure S1C** for visualisation of single dimensions).

Measuring highlight features

The coverage, sharpness and contrast of highlights (**Figure S4A**) were measured using a MATLAB package developed by [43]. Briefly, the features are quantified as follows: (1) Coverage is defined as the proportion of image pixels which have higher intensities in the full image than in a corresponding image rendered using only the diffuse component of the surface reflectance function. (2) Sharpness is defined as the Local Phase Coherence [45] within only the specular-component image (i.e. full image minus diffuse-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 gloss ratings were simply averaged across participants for each stimulus. In Experiment 2, gloss rankings were converted to the average ranking position of each image step for each participant, averaging over the four sequences using matte seed images, and the four sequences using glossy seed images. In Experiments 3a and 3b, pairwise gloss comparisons were converted to the proportion of times each image of each sequence was judged as having higher gloss, across pairings and repetitions. In Experiment 4,

pairwise gloss comparisons were converted to the proportion of times “image B” (arbitrarily labelled, but consistent across participant and model analyses) was judged as having higher gloss than “image A”, and this proportion was subtracted from 0.5, so that a value of 0 indicates equal apparent gloss (good constancy), and deviations in either direction indicate deviations from constancy.

DATA AND SOFTWARE AVAILABILITY

All data, analysis code, and stimuli will be made available at Zenodo upon acceptance.

References

1. Anderson, B. L. (2011). Visual perception of materials and surfaces. *Current Biology*, *21*(24), R978-R983.
2. Chadwick, A. C., & Kentridge, R. W. (2015). The perception of gloss: A review. *Vision Research*, *109*, 221-235.
3. Fleming, R. W. (2017). Material perception. *Annual Reviews of Vision Science* 3(1). 365-388.
4. Anderson, B. L. (2015). Can computational goals inform theories of vision?. *Topics in Cognitive Science*, *7*(2), 274-286.
5. Hoffman, D. D., Singh, M., & Prakash, C. (2015). The interface theory of perception. *Psychonomic Bulletin & Review*, *22*(6), 1480-1506.
6. Fleming, R. W., & Storrs, K. R. (2019). Learning to see stuff. *Current Opinion in Behavioral Sciences*, *30*, 100-108.
7. Ho, Y. X., Landy, M. S., & Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture. *Psychological Science, 19*(2), 196-204.
8. Marlow, P. J., Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology, 22*(20), 1909-1913.
9. Doerschner, K., Fleming, R. W., Yilmaz, O., Schrater, P. R., Hartung, B., & Kersten, D. (2011). Visual motion and the perception of surface material. *Current Biology*, *21*(23), 2010-2016.
10. Nishida, S. Y. (2019). Image statistics for material perception. *Current Opinion in Behavioral Sciences*, *30*, 94-99.
11. Sawayama, M., & Nishida, S. Y. (2018). Material and shape perception based on two types of intensity gradient information. *PLoS Computational Biology*, *14*(4), e1006061.
12. Wijntjes, M. W., & Pont, S. C. (2010). Illusory gloss on Lambertian surfaces. *Journal of Vision*, *10*(9), 13-13.

13. Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A. A., Visin, F., Vazquez, D., & Courville, A. (2016). PixelVAE: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013*.
14. Zhao, S., Song, J., & Ermon, S. (2017). Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658.*
15. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1026-1034).
16. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... & Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. *ICLR, 2*(5), 6.
17. Testolin, A., Stoianov, I., & Zorzi, M. (2017). Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nature Human Behaviour*, *1*(9), 657.
18. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(Nov), 2579-2605.
19. Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323-2326.
20. DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333-341.
21. Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613.
22. Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401-412.
23. Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage, 56*(2), 400-410.
24. Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49-70.

25. Motoyoshi, I., Nishida, S. Y., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature*, *447*(7141), 206.
26. Kersten, D., Mamassian, P. & Yuille (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271–304.
27. Marr, D. (1982). *Vision*. San Francisco: Freeman.
28. Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research*, *41*(24), 3145-3161.
29. Poggio, T., Torre, V., & Koch, C. (1987). Computational vision and regularization theory. In *Readings in Computer Vision* (pp. 638-643). Morgan Kaufmann.
30. Anderson, B. L., & Kim, J. (2009). Image statistics do not explain the perception of gloss and lightness. *Journal of Vision, 9*(11), 1-17.
31. Geisler, W. S., & Kersten, D. (2002). Illusions, perception and Bayes. *Nature Neuroscience*, *5*(6), 508.
32. Eslami, S. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., ... & Reichert, D. P. (2018). Neural scene representation and rendering. *Science*, *360*(6394), 1204-1210.
33. Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., & Efros, A. A. (2018). Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*.
34. Haber, N., Mrowca, D., Fei-Fei, L., & Yamins, D. L. (2018). Emergence of structured behaviors from curiosity-based intrinsic motivation. *arXiv preprint arXiv:1802.07461.*
35. Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castaneda, A. G., ... & Sonnerat, N. (2018). Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. *arXiv preprint. arXiv:1807.01281*
36. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529.
37. Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR, 2016*

38. Brainard, D. H., Roorda, A., Yamauchi, Y., Calderone, J. B., Metha, A., Neitz, M., ... & Jacobs, G. H. (2000). Functional consequences of the relative numbers of L and M cones. *JOSA A*, *17*(3), 607-614.

39. Smirnakis, S. M., Berry, M. J., Warland, D. K., Bialek, W., & Meister, M. (1997). Adaptation of retinal processing to image contrast and spatial scale. *Nature*, *386*(6620), 69.

40. Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning?. *Frontiers in Psychology*, *8*, 2124.

41. Anderson, B. L. (2015). Can computational goals inform theories of vision? *Topics in Cognitive Science,* 7, 274-286.

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

43. Schmid, A. C., & Doerschner, K. (2019). Chocolate, chrome, or cloth? The appearance of specular highlights determines perceived material category. *Journal of Vision, 19*(10), 252-252.

44. Lagarde, S., Lachambre, S., & Jover, C. (2016) An artist-friendly workflow for panoramic HDRI. *SIGGRAPH, 2016*

45. Hassen, R., Wang, Z., & Salama, M. M. (2013). Image sharpness assessment based on local phase coherence. *IEEE Transactions on Image Processing, 22*(7), 2798-2810.

Supplemental Information

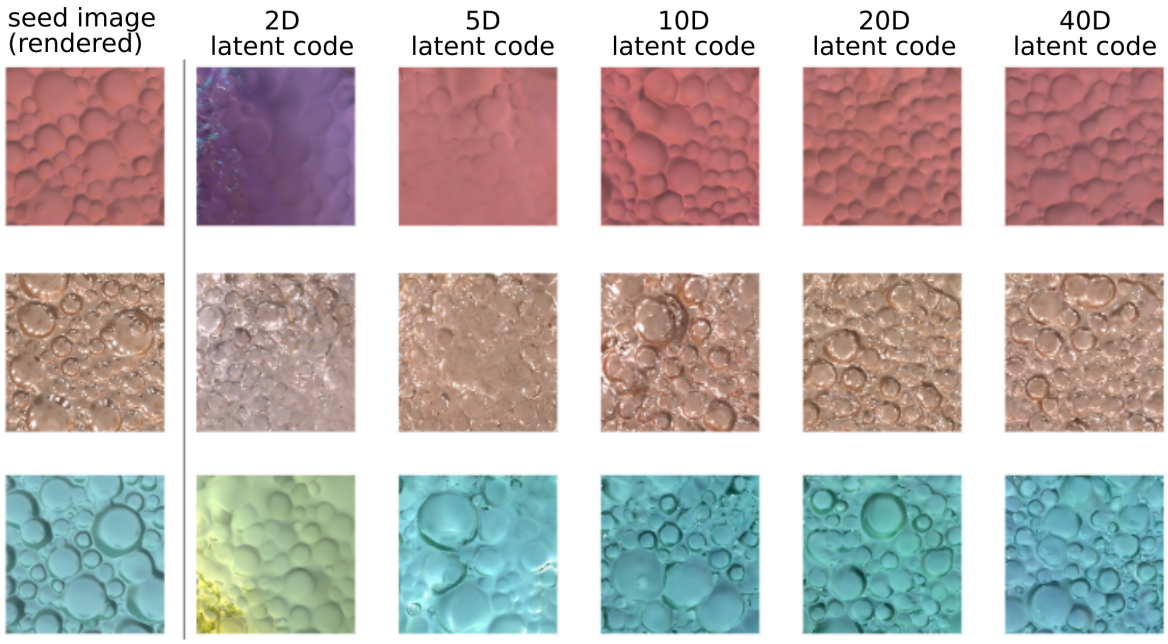
Figure S1. Related to Figure 1. (following page)

(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.

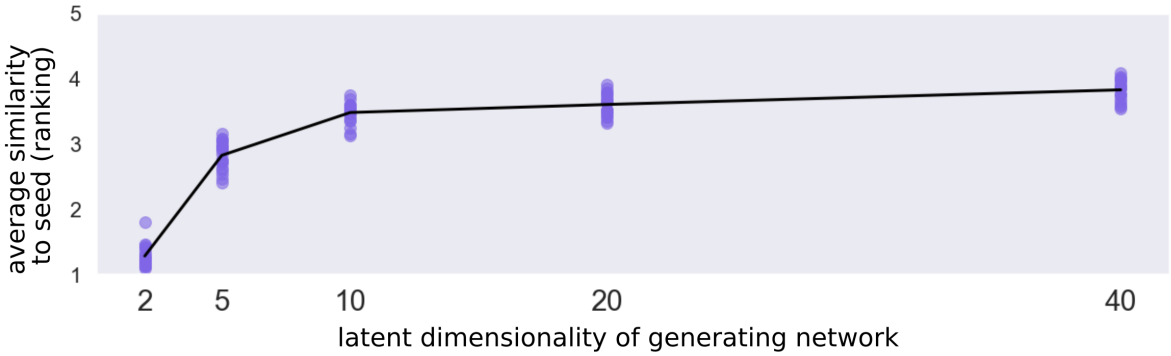
(B) We wanted to select a latent dimensionality for the PixelVAE model such that the network was forced to learn a highly compressed code, but still able to capture the majority of meaningful structure in the training images. To gauge where this dimensionality lay for our dataset, we showed twenty human 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) 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

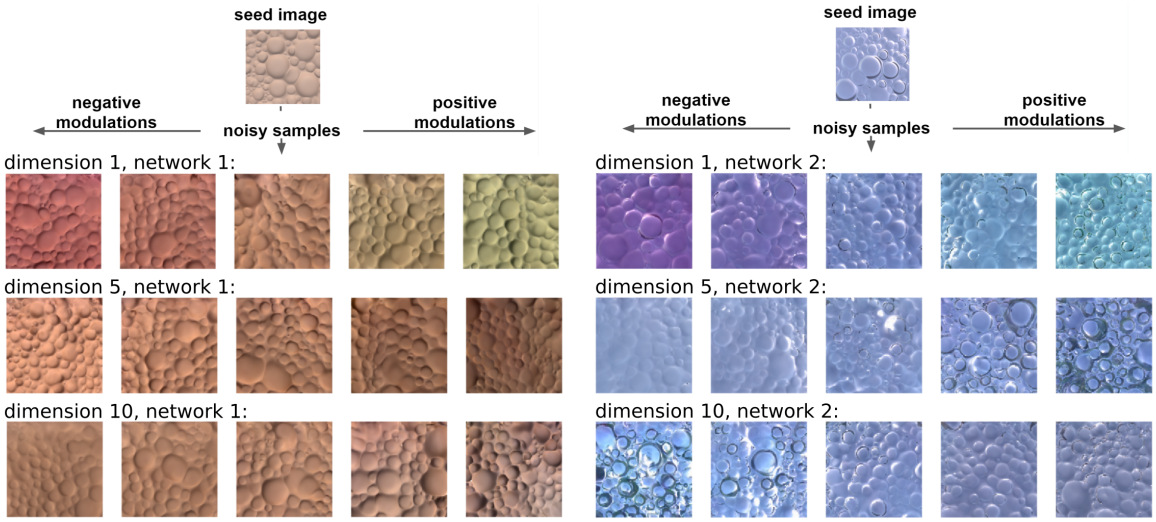


Figure S2. Related to Figure 2. (following page).

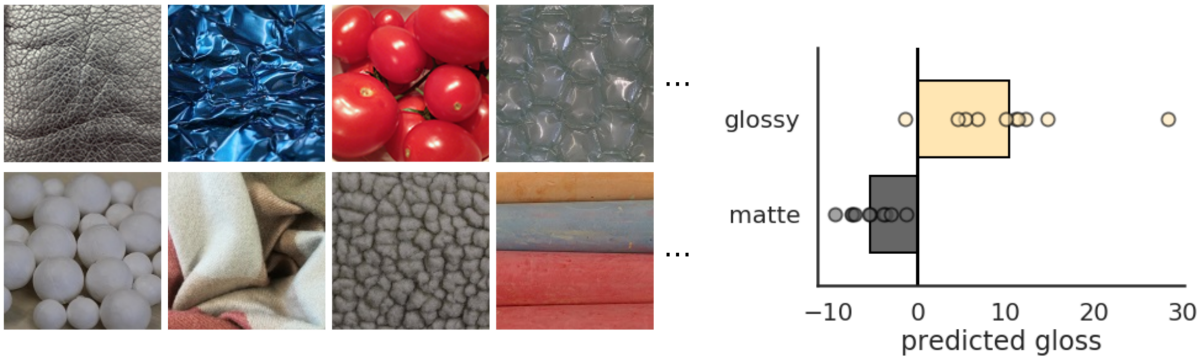
(A) When we input ten photographs of real-world glossy surfaces (four examples shown: leather, metallic foil, tomatoes, plastic), all but one were correctly categorised as high-gloss, based on their latent representations in the unsupervised models (using the gloss SVM classifier trained on the main 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) Although our unsupervised networks were trained on a simulated environment with a bimodal distribution of glosses (near-matte or high-gloss), their latent representations can nevertheless be used to smoothly regress gloss levels in new datasets containing continuously-varying gloss. The networks can also generalise across large changes in surface geometry. Scatterplots show predicted (x-axis) and ground-truth (y-axis) gloss. Gloss regression is equally good for the trained geometry (leftmost panel) as for four other arbitrary surface geometries never seen during training (right four panels).

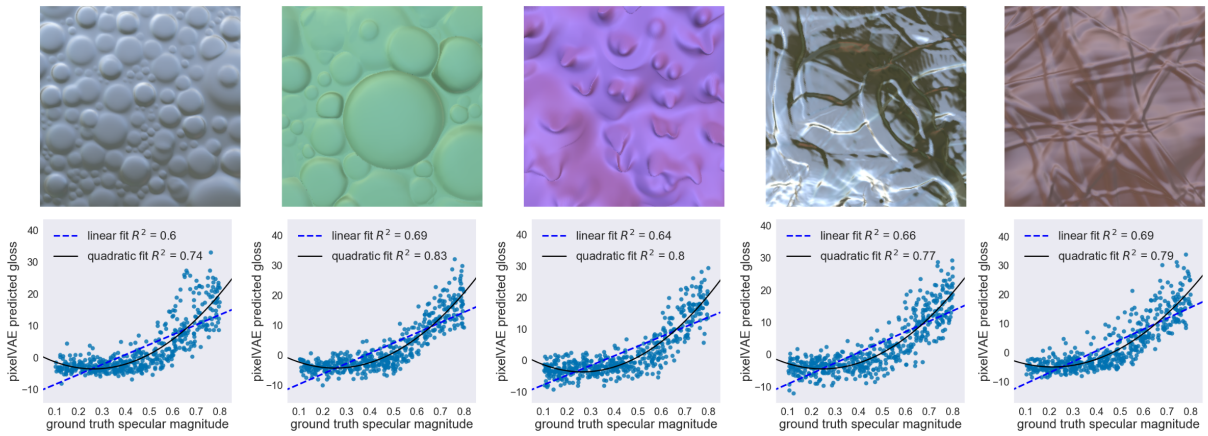
(C) The trained networks organise images by surface material even for unfamiliar geometries. Panels show tSNE embedding of 500 images from each of five surface geometries, colour coded by strength of specular reflectance.

(D) 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, (top right) t-weighted stochastic neighbourhood embedding, and (bottom left) locally-linear embedding. High gloss (pale dots) and low-gloss (dark dots) are thoroughly intermixed in all embeddings. Allowing an equal number of dimensions as the neural networks' latent codes does not dramatically improve gloss separation; bar chart shows gloss classification accuracy for (from left to right) the unsupervised PixelVAE models, the supervised ResNet mdoels, the skewness of image luminance histograms [25], a multi-scale texture description [24], and 10-dimensional embeddings of images via LLE, tSNE, and MDS.

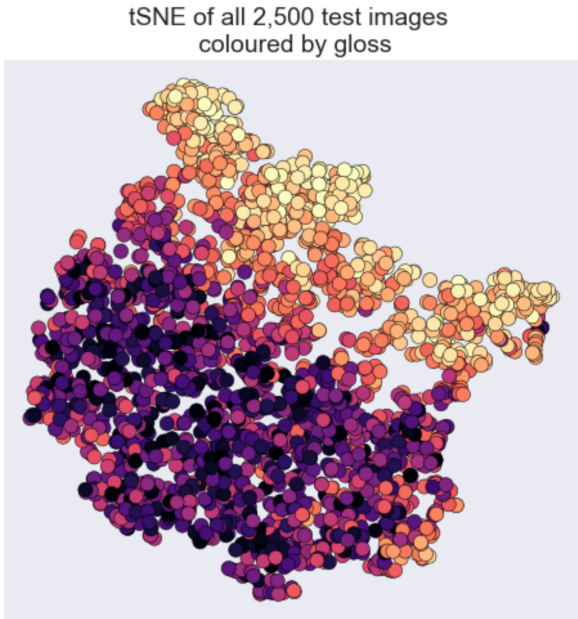
(a) The model can classify photographs of real-world surfaces as glossy or matte



(b) Continuous gloss levels can be decoded for original (leftmost) and four new geometries



(c) Gloss representation for all geometries



(d) Simple pixel embeddings are poor predictors of ground-truth gloss

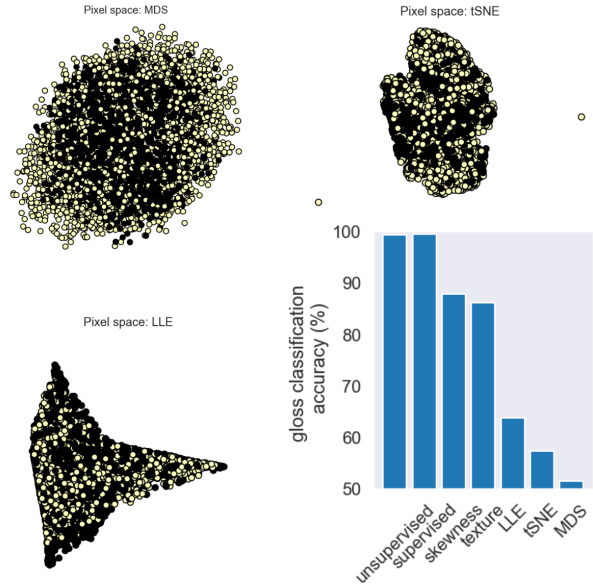


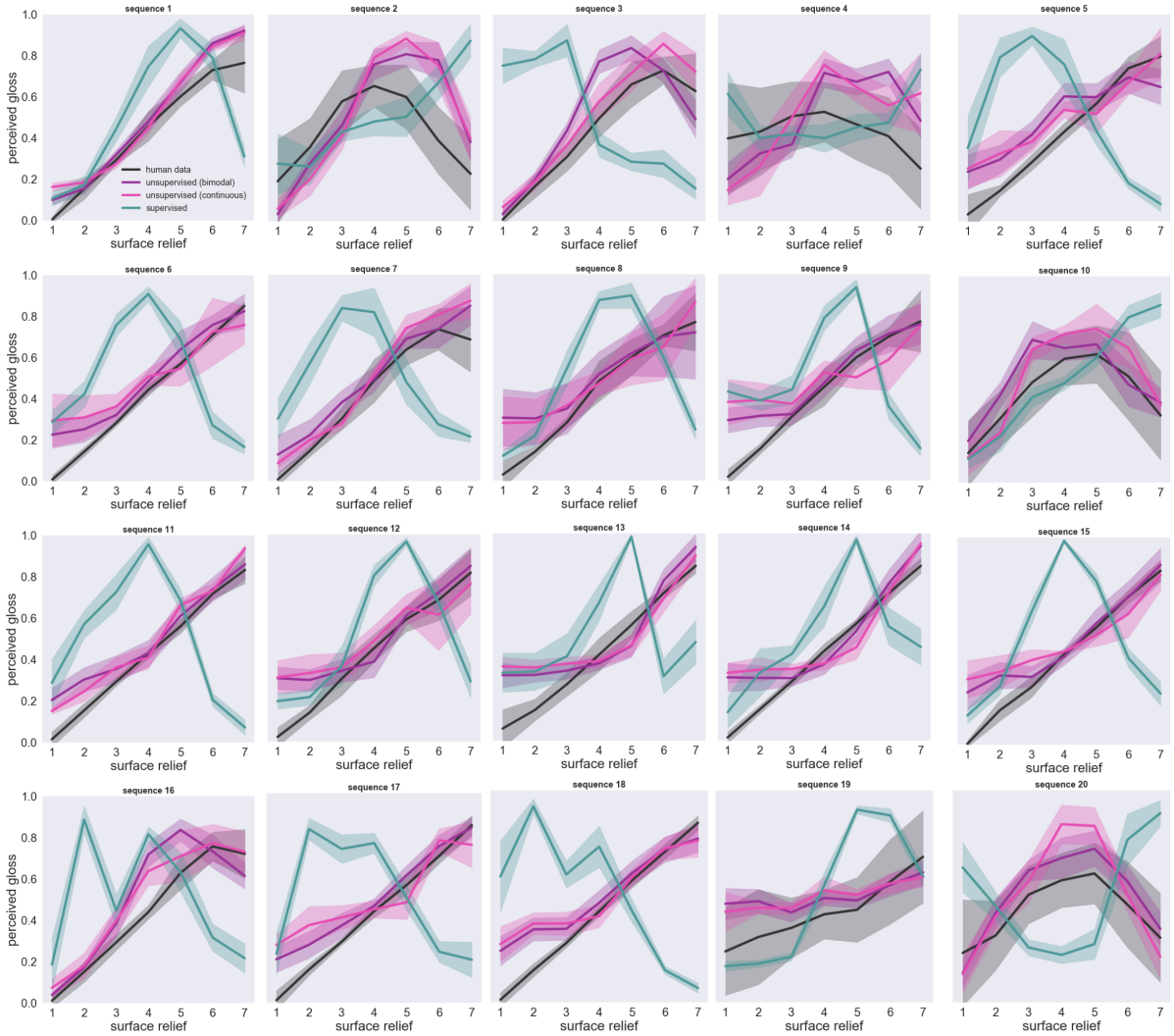
Figure S3. Related to Figure 4. (following page)

(A) Full data from each of the twenty constancy sequences summarised in Figure 4C 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.

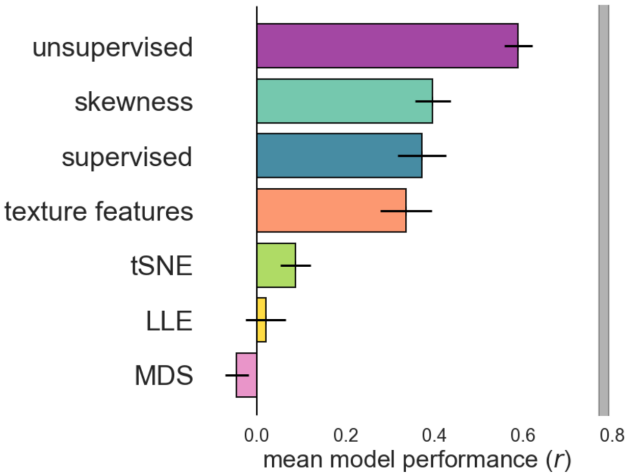
(B) Average correlation to each human participant's ratings for the unsupervised model and for alternative models. Vertical grey line indicates how well individual human ratings can be predicted from the data of other participants, giving the maximum possible model performance. Error bars: standard error of the mean over participants.

(C) Relationship between objective gloss-classification accuracy (x-axis) and average R^2 in predicting human psychophysical gloss judgements across Experiments 1, 3 and 4 (y-axis). The unsupervised deep network is the only model to both achieve extremely high gloss-classification, and to well predict the idiosyncratic errors of human gloss perception.

(a) Human data and model predictions for each test sequence



(b) Model comparison for Expt 4



(c) All-experiment summary

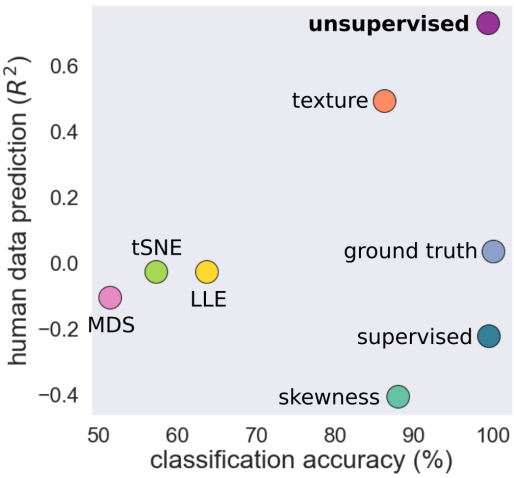


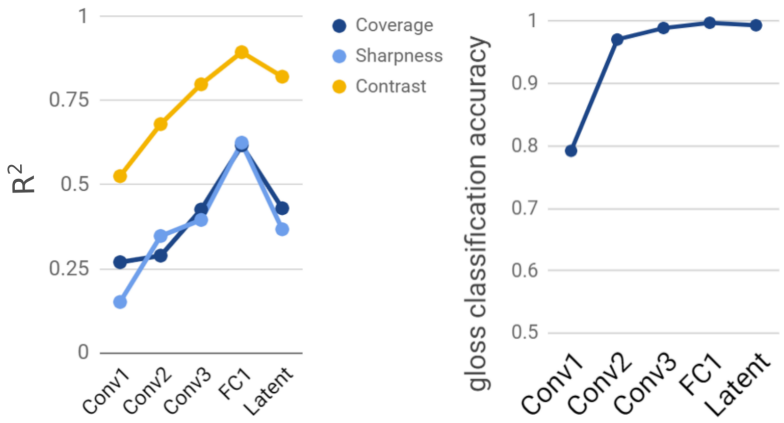
Figure S4. Related to Figure 4. (following page)

(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).

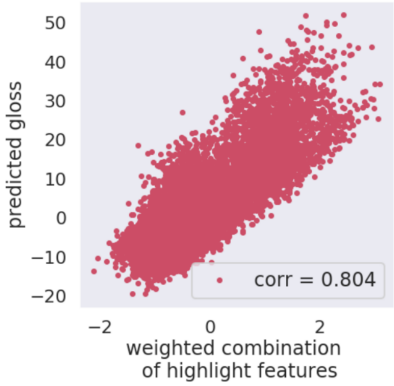
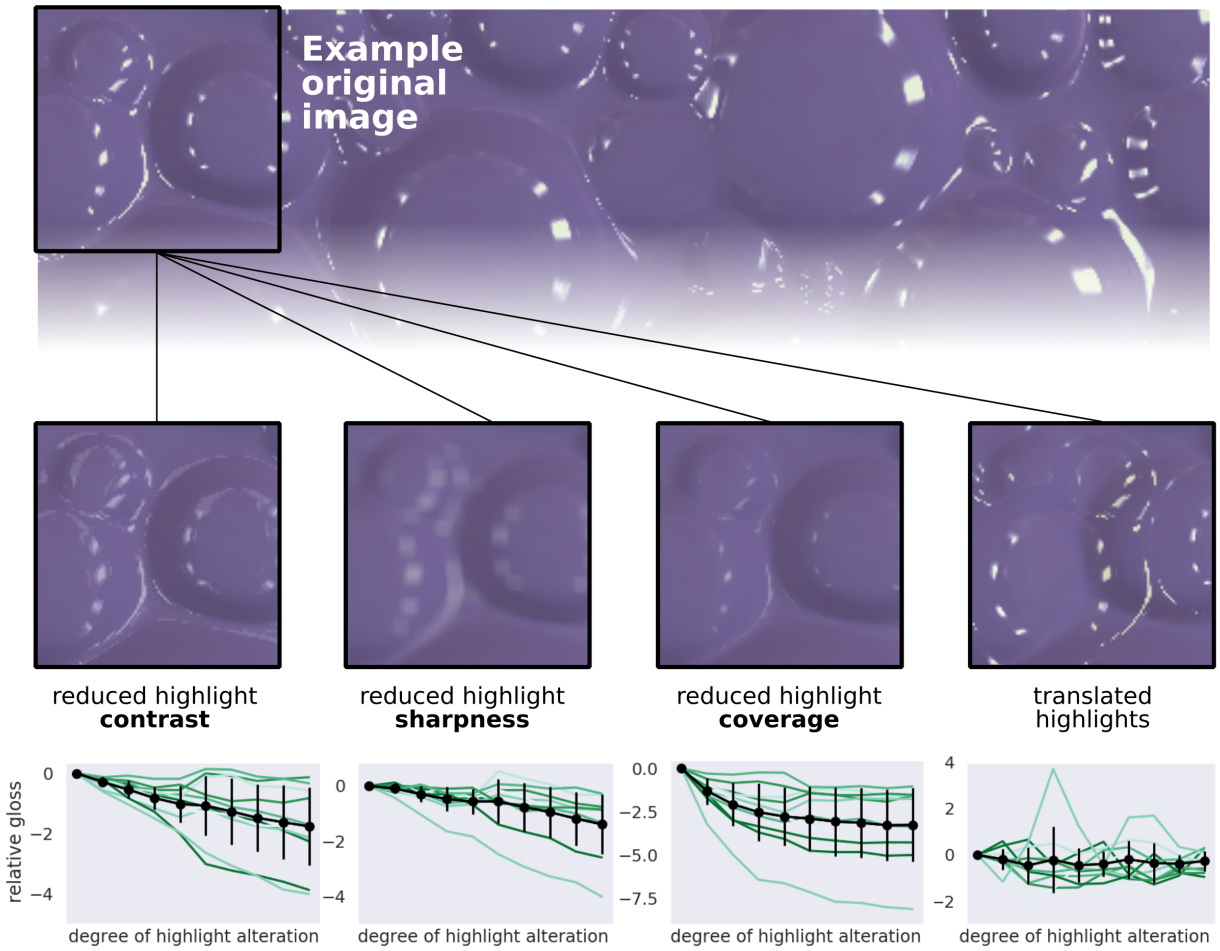
(B) Combining (via equal weighting) the estimates of the three highlight features, as decoded from the fully-connected layer immediately preceding the latent layer, yields a summary highlight measure which correlates well with the gloss level predicted from the latent layer.

(C) For a set of 10 probe images, we separately rendered the specular and matte components, and manipulated the specular component by either reducing contrast, blurring, shrinking coverage via an image erosion algorithm, or laterally translating, in ten progressive steps. We then recombined the two components to produce ten sets of ten full images, and presented them to the network. Model-predicted gloss systematically declined as any of the three highlight features was progressively reduced (left three plots). However, translating the specular components, which does not affect these highlight features, does not affect the model's predicted gloss (although it does reduce human-perceived gloss). Our unsupervised models appear to have learned a representation of gloss which extracts similar highlight features as human vision, but which is not yet sensitive to the more complex photogeometric material cues we are. Learning this may require richer experience with variations in surface shape and viewpoint.

(a) Both highlight features and ground-truth gloss can be decoded with increasing accuracy from successive layers



(b) An equally-weighted combination of the decoded highlight features correlates with model's gloss prediction



(c) Manipulations that diminish highlight features also reduce model's predicted gloss value