Translucency perception emerges in deep generative representations for natural image synthesis

Chenxi Liao[1*], Masataka Sawayama[3], and Bei Xiao[2]

[1]American University, Department of Neuroscience, Washington DC, USA
[2]American University, Department of Computer Science, Washington DC, USA
[3]Inria, Bordeaux, France
[*]cl6070a@student.american.edu

ABSTRACT

Material perception is essential in planning interactions with the environment. The visual system relies on diagnostic image features to achieve material perception efficiently. However, discovering the features, especially for translucent materials, has been challenging due to the high variability of material appearances under interactions of shape, lighting, and intrinsic materials. Here, we learn a latent space informative of human translucency perception by developing a deep generative network trained to synthesize images of perceptually persuasive material appearances. Without supervision, human-interpretable scene attributes, including object's shape, material, and body color, spontaneously emerge in the latent space in a scale-specific manner. Critically, the middle-layers of the latent space selectively encode the translucency features correlating with perception, suggesting that translucent impressions are established in the mid-to-low spatial scale features. Our findings illustrate the promising capability of unsupervised learning in finding representative dimensions for materials and discovering perceptually relevant features for visual inference.

Introduction

Humans assess the appearance of materials every day to recognize objects and plan actions, such as evaluating the ripeness of the fruits or preparing to pick up a croissant without crushing it. Visually perceiving materials is a first step for anticipating multi-sensory experiences[1–3]. Yet, despite its biological significance and importance in connecting perception to action[4,5], material perception is still poorly understood in human cognition and artificial intelligence. The essential challenge of material perception is that materials can be made into objects with any color and shape, and their appearances can be profoundly changed under the joint effect of lighting, view point, and other external factors[6–10]. Nevertheless, humans can still effortlessly recognize and discriminate materials under diverse contexts[11,12]. How humans extract intrinsic material properties across the enormous range of different contexts remains unsolved.

The challenge of material perception especially stands out for translucent materials such as wax, fruit, and skin. Nearly all materials we encounter permit light into the surface to some degree, which involves a physical process of light transport, namely subsurface scattering[13,14]. When light hits a translucent object, some of it penetrates the object, refracts, and scatters multiple times throughout the body of the medium before exiting from a different location on the surface (see Supplementary Figure S.2 for an illustration). This gives rise to the essential "translucent" appearance, such as the aliveness of skin. Perceiving translucency not only plays a critical role in material discrimination and identification, such as telling the difference between raw and readily cooked food, but also allows us to appreciate the beauty of aesthetic objects such as jewelry, sculptures, and still life paintings[15,16]. Intrinsically, translucency is impacted by the material's optical properties, including absorption and scattering coefficients, phase function, and index of refraction[17–19]. Extrinsically, the object's shape, the surface geometry, and the illumination direction also have striking effects[7,20–26]. The generative process of translucency involves complex interactions among various intrinsic and extrinsic factors, leading to a wide variety of appearances under different contexts. There are two difficulties in studying translucency. First, given the large variation of translucent appearances across materials and scene factors, it has been difficult for humans to provide explicit labels to describe material properties. For instance, the label "soap" can refer to a variety of translucent appearances and in the mean time, humans may lack precise descriptions for the subtle visual differences between two translucent objects. This makes it difficult to measure translucency using real-world stimuli and obtain image datasets based on human perception, unlike objects and scenes[27]. The currently available translucent image datasets mostly used rendered images labeled by physical parameters instead of human perception[28]. Second, since many scene factors affect translucent appearance, how the visual system disentangles the contribution of these factors and achieves a compact representation of materials from the retinal image remains unanswered.

For the second difficulty, many previous studies in material perception sought to find diagnostic image features for perceived materials using analytical methods. For example, researchers have used well-controlled photorealistic images to analyze the physics-image relationships of a target material, extract the essential image features, and test if they are diagnostic for human perception[9,29–37]. Such an approach has been used to study various material qualities, including surface gloss[26,29,30,34,38–51], surface roughness[52,53], liquid viscosity[54–57], stiffness of objects[58–61] and cloths[62], surface wetness[63], transparency[64,65], and translucency[7,19,21,23,66–71]. However, finding image features from the physics-image analysis can be challenging when a material appears differently across scenes, causing the discovered features to be idiosyncratic to particular scene factors. This problem is especially amplified in translucency (see [9,37] for reviews). Recently, data-driven approaches have attempted to learn material representations by capturing the statistical structure of material appearance across many image samples[10,32,35,72–77]. This approach has been successfully used to model human perception. For example, Storrs et al. (2021) rendered opaque gloss and matte images under various illuminations and geometries, trained a variational autoencoder (VAE) model by the images without the supervision of physical attributes, and elucidated the latent image features correlated with human gloss perception[35,78,79]. Their work shows the capability of unsupervised learning to disentangle scene factors without physics-image analyses. Many recent works in high-level vision also utilize this unsupervised approach[76,80–85]. However, decoding translucency is still challenging because a simple encoder-decoder network used in VAEs cannot disentangle the contributing factors of translucent appearances due to material complexity without the supervision of physical parameters [28].

Here, we aim to learn, unsupervised, a compact latent representation containing the structural information of translucent materials and to explore whether such a latent representation informs perception. We propose a Translucent Appearance

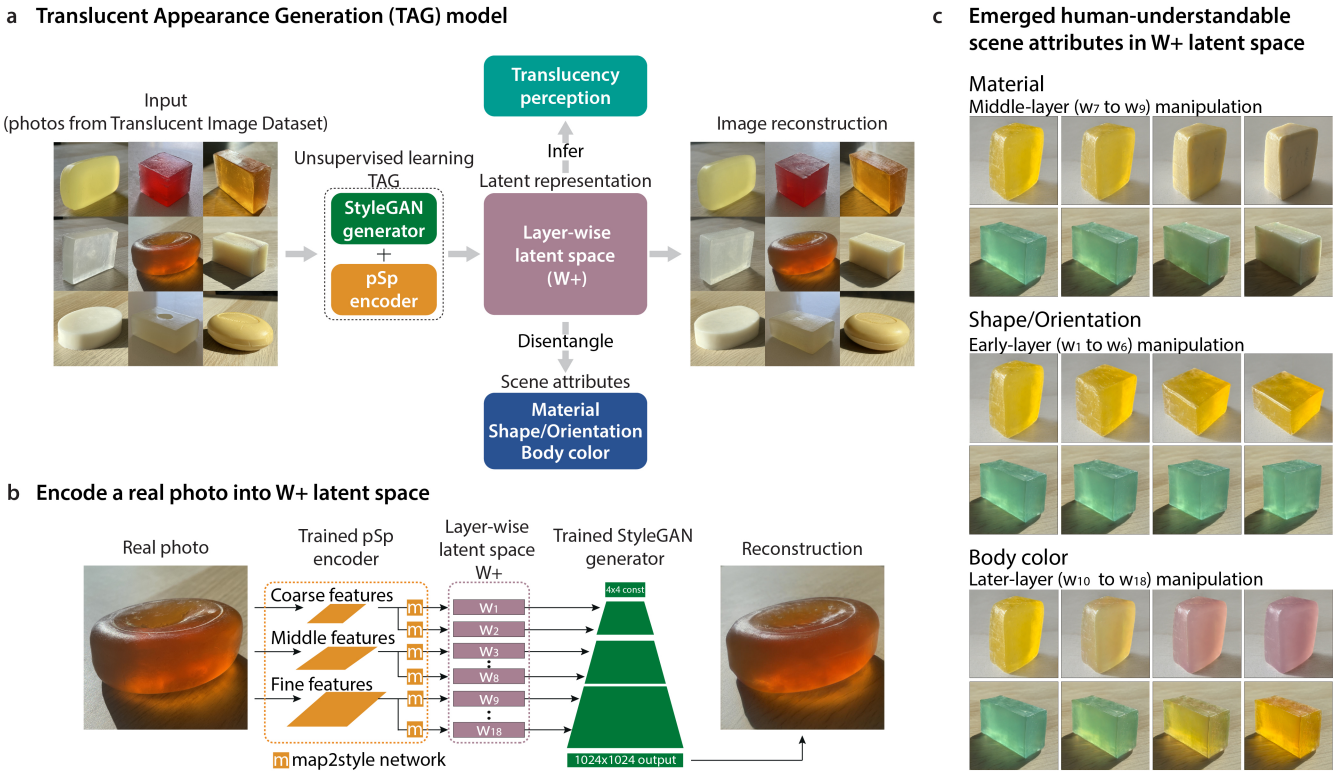

Figure 1. The Translucent Appearance Generation (TAG) model. a, The TAG framework, which is based on the StyleGAN2-ADA generator and pSp encoder architectures, learns to synthesize perceptually convincing images of translucent objects. The model maps photos of translucent objects into the $W+$ latent space. The $W+$ can disentangle the effects of scene attributes (e.g., shape, material, and body color) and predict human perception of translucency. **b,** The detailed process of embedding a photo into StyleGAN's $W+$ latent space. This allows us to generate image at a particular location in the latent space. **c,** Emergent human-understandable scene attributes in the layer-wise latent space. Without supervision, the $W+$ spontaneously disentangles three salient scene attributes: material, shape/orientation, and body color. In each row, an original generated image (left) is gradually manipulated by modifying its latent vectors at specific layers. Early-layer (w_1 to w_6) manipulation of $W+$ transforms the shape and orientation of the object. Middle-layer (w_7 to w_9) manipulation modifies the material appearance. Later-layer (w_{10} to w_{18}) manipulation changes the body color (color of the diffuse component of the surface reflection).

Generation (TAG) model trained on our own large-scale dataset of natural photographs of translucent objects (TID). We focus on a typical translucent object category commonly seen in daily life, soaps. The raw materials of soaps can be manufactured in varying shapes and colors, serving as a great medium to investigate the variety of translucent appearances. TAG contains two modules: a style-based generative adversarial network (StyleGAN)[86–88] and a pixel2style2pixel (pSp) encoder[89] (Figure 1(a)). StyleGAN learns to synthesize images of perceptually convincing translucent materials using its latent space. Unlike the traditional deep generative models (e.g., GAN[90] and DCGAN[91]), StyleGAN utilizes a layer-wise latent space to model high-dimensional distributions of data, leading to an unsupervised separation of visual attributes at different abstraction levels presented in the image domains[86,92,93]. We use the pSp encoder to navigate in the learned StyleGAN's latent space and efficiently explore its representative meaning in the expressiveness of translucency (Figure 1(b)). Our framework provides a pathway to alleviate two difficulties of studying translucency perception: the lack of an explicitly labeled image dataset, and obtaining a compact representation of material properties from high-dimensional image data. First, without explicit labels, our model learns to represent materials by finding a candidate distribution of features that is similar to the distribution corresponding to real photos of translucent objects. The learning process is based on a straightforward goal of generating samples that are indistinguishable from the real ones. Second, taking advantage of StyleGAN's representational power, we discover a layer-wise space that spontaneously disentangles translucency-relevant attributes and captures the internal dimensions characterizing the variation of translucent appearances.

We demonstrate that TAG can create perceptually persuasive and diverse translucent appearances (Figure 1(a)). Crucially, we show that human-understandable scene attributes emerge in our model's learned latent space (Figure 1(c)). Without supervision of physical factors, scale-specific scene attributes related to translucency perception can be separately represented in the layer-wise latent space: material, shape/orientation, and body color. We find that the middle-layers of the latent space selectively encode the translucency features correlated with human perception. By leveraging the representational properties of the learned latent space, we identify critical image features diagnostic of translucency such as scale-specific color basis functions. Our results suggest the unsupervised generative framework may discover an efficient representational space of materials and reveal image regularities potentially used by the visual system to estimate material properties.

Results

Translucent Appearance Generation (TAG) model

Our main goal is exploring the learned latent space of our model. TAG consists two parts, illustrated by Figure 1(a) and (b): a StyleGAN2-ADA generator[88] and a pSp encoder network[89]. We began by training a StyleGAN2-ADA generator, a variant of StyleGAN2[87] with adaptive discriminator augmentation (ADA), with unlabeled images from our customized Translucent Image Dataset (TID). TAG's generator network aims to synthesize novel images that are indistinguishable from the real photographs of soaps, without having any knowledge about the physical process of translucency. After training the generator, we could use it to synthesize numerous novel images of translucent objects by sampling from the learned StyleGAN's latent space.

Instead of generating soaps randomly, we wanted to reconstruct a real photo by mapping it into the StyleGAN's latent space so that we could explore how various visual attributes of a material are represented. After obtaining the trained StyleGAN2-ADA generator, we separately trained a pSp encoder network, which could embed a real photograph of soap into the StyleGAN's extended intermediate latent space $W+$. Mapping the real photo into the layer-wise latent space $W+$ leads to accurate reconstruction quality and expressiveness of the input[94–96]. Given a real image, the pSp encoder extracts the 18 latent vectors of $W+$ (w_1 to w_{18}), which are then inserted into the trained StyleGAN2-ADA generator's convolution layers corresponding to their spatial scales in order to reconstruct the input (Figure 1(b)). Figure 1(a) shows examples of the model-generated images of soaps using these methods. The above steps allowed us to effectively examine whether the layer-wise latent space can disentangle the effects of scene attributes on the image appearance and further explore whether such latent representation informs human perception (Figure 1(c)).

TAG synthesizes perceptually convincing and diverse translucent appearances

Before looking into the learned latent space, we first evaluated the perceptual quality of the generated images from two aspects. In Experiment 1, we evaluated the overall image quality and realism of the generated images. In Experiment 2, we further investigated whether the material properties of the generated objects were perceptually convincing and could convey material attributes in the same way as the real images.

In Experiment 1, twenty observers completed a real-versus-generated discrimination task wherein they were instructed to discriminate whether an image is a photograph of a soap or was generated from the TAG model. We presented the observers with 300 images of soaps, half of which were real photographs and the other half were generated images. Figure 2(a) shows examples of the stimuli. Each stimulus was presented for one second, then the observers made the real-versus-generated judgment (Figure 2(b)). The 300 stimuli were pre-randomized, and each stimulus was judged twice.

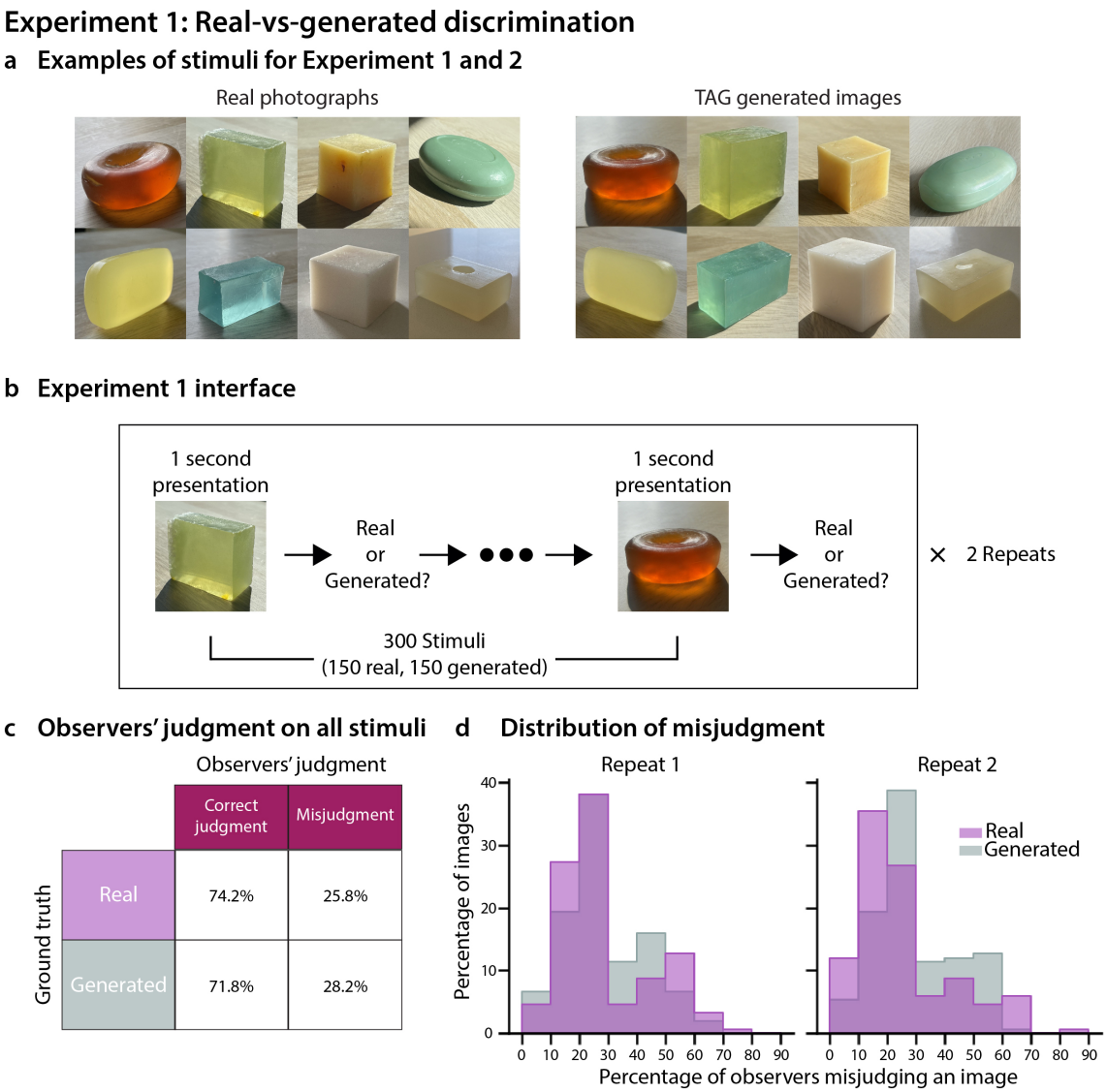

Figure 2. Experiment 1: Real-versus-generated discrimination. a, Examples of real photographs and model-synthesized images of soaps. The "generated" soaps were synthesized by embedding a real photograph into the $W+$ latent space of the trained StyleGAN2-ADA using the pSp encoder. We used 150 real photographs and 150 generated images as stimuli for Experiment 1 and 2. **b,** The procedure of Experiment 1. **c,** Overall correct and error rates of judging real and generated images. The error rate of 50% indicates pure guessing. **d,** Distribution of the percentage of real and generated images misjudged by the observers. The x-axis represents the percentage of observers misjudging an image and the y-axis is the percentage of images being misjudged. Purple color represents data of real images and gray represents data of generated images.

If observers could perfectly tell the generated image from the real, they would have 0% misjudgment. On the other hand, if they failed to distinguish between real and generated images, they would be purely guessing and misjudging at a 50% chance. Our results show that across all observers and trials, the observers misjudged 28% of generated images and 25% of real photos (Figure 2(c)). Meanwhile, Figure 2(d) shows that the distributions of observers' misjudgments were very similar for both real and generated conditions in both repeats. Specifically, approximately 40% of generated images were erroneously judged as "real" by at least 30% of observers in both repeats. Only 10% of the generated images were correctly identified by all observers. For a substantial number of images, observers could not discriminate the generated images from the real ones. Our results are on par with the recent findings of human evaluation of StyleGAN-generated high-resolution human face images, where the error rate of judging generated images was also 28%[97]. Overall, the results indicate that our model can successfully generate a large number of perceptually convincing images that fool observers into judging them as real.

In Experiment 2, we evaluated whether the generated images of soaps could convey perceptually persuasive material

qualities. Specifically, the same twenty observers from Experiment 1 rated three translucency-related attributes on a seven-point scale (1 means low, 7 means high): translucency, see-throughness, and glow (Figure 3(a)), which were found in a previous study to be descriptive in semantic judgments of translucent objects[12]. Material attribute ratings were normalized to the range 0 to 1 for each observer. For each image, the normalized attribute ratings were averaged across observers, and subsequent data analysis was based on these values. Figure 3(b) shows that observers perceived different degrees of translucency, see-throughness, and glow from the generated images, with ratings distributed similarly to those of real photos. This shows that observers could perceive a wide range of translucent material attributes from the generated images. Meanwhile, the material attributes perceived by the observers are highly positively correlated with one another for both real photographs and generated images of soaps (Figure 3(c)). The correlation among the attributes are in agreement with our previous empirical findings[12]. Figure 3(d) shows examples of real and generated images judged to have various degrees of translucency similar to that of real photographs. Together, our results suggest that TAG learns to synthesize diverse perceptually convincing translucent appearances and conveys material attributes similarly to the real photographs.

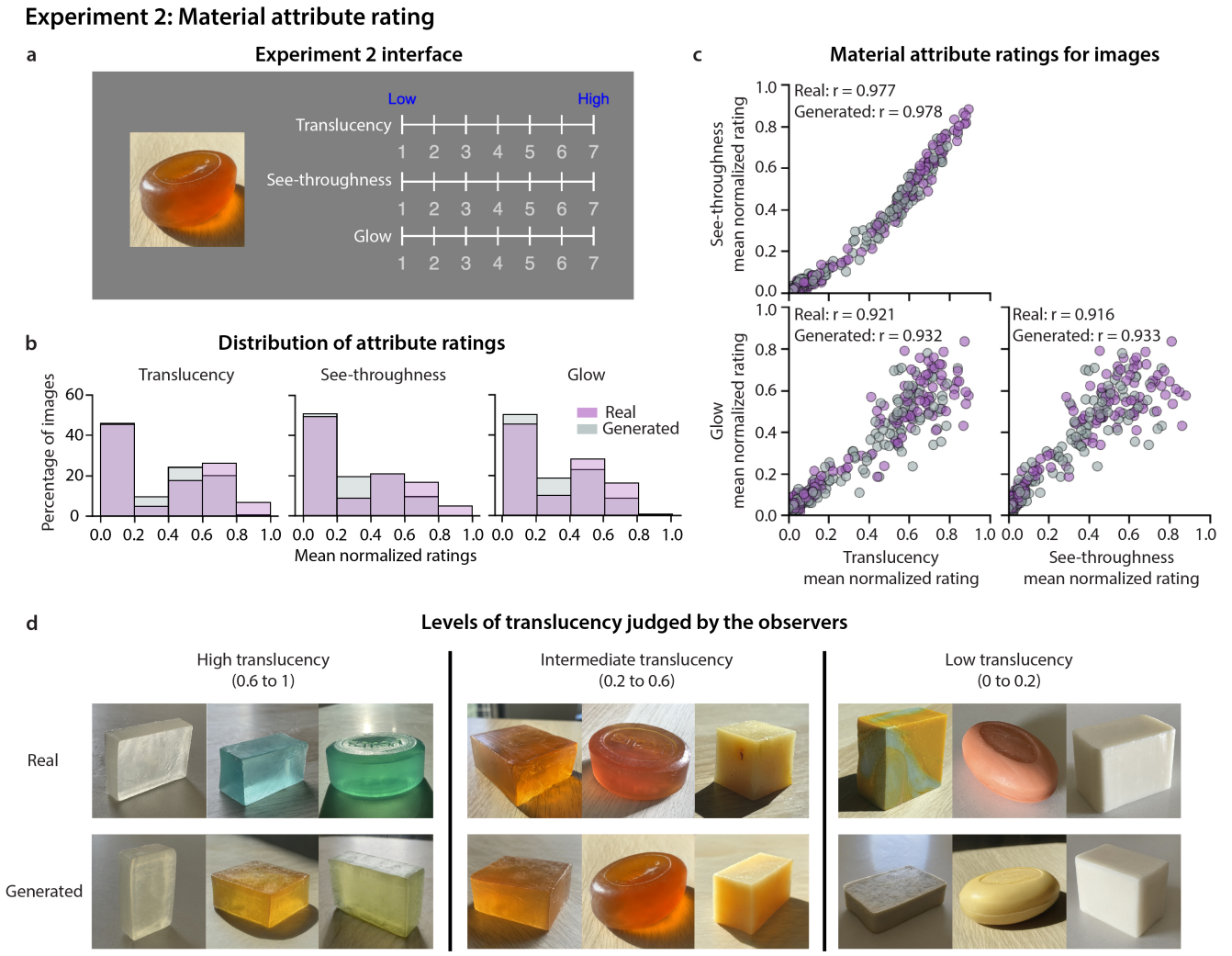

Figure 3. Experiment 2: Material attribute rating. **a,** The user interface of Experiment 2. **b,** The distribution of the mean normalized attribute ratings across observers. For each observer, we normalize their attribute ratings to 0 and 1. The x-axis represents the normalized ratings of an attribute averaged over 20 observers, and the y-axis shows the percentage of images. **c,** The scatter plots of ratings between a pair of material attributes, with the Pearson correlations shown at the top. All correlation coefficients are statistically significant at confidence level of 95%. In both (b) and (c), purple and gray colors represent results for real and generated images, respectively. **d,** Examples of real and generated images judged to have different levels of translucency. We grouped the images based on the mean normalized translucency rating: high (0.6 to 1), intermediate (0.2 to 0.6), and low (0 to 0.2).

Perceptually meaningful scene attributes emerge in the learned latent space

What makes the generated images convey perceptually persuasive material appearance? We hypothesize that TAG's $W+$ latent space is obliged to learn the explanatory factors underlying the structure of observed data. To test our hypothesis, we systematically manipulated different layers of the latent code and inspected how these manipulations affect the visual attributes of the output image. Specifically, we applied morphing between the latent codes of a pair of images (a source and a target), which differ in their shapes, intrinsic materials, lighting environments, and body colors (Figure 4(a)).

Given two generated images A (source) and B (target) with their corresponding $W+$ latent codes, w_A and w_B (18×512-dimensional latent codes), morphing can be applied on particular layers of their latent codes to create a sequence of generated images with visual appearances lying between the source and the target. The morphed latent vectors are generated by a linear interpolation of a particular set of layers (s) between the source ($w_A^{\{s\}}$) and target ($w_B^{\{s\}}$) while keeping the other layers from source image unchanged:

$$w_\lambda^{\{s\}} = (1-\lambda)(w_A^{\{s\}}) + \lambda(w_B^{\{s\}}), \lambda \in [0,1] \tag{1}$$

where λ is the interpolation step and $w_\lambda^{\{s\}}$ is the resultant latent vectors of the set of layers. The generator then uses the combination of $w_\lambda^{\{s\}}$ and the remaining unchanged latent vectors from the source image to produce a new image (e.g., one of the intermediate images in image sequence shown in Figure 4(a)). For example, when we apply morphing between the source and the target on their latent vectors of layers 7, 8, and 9, the resultant latent vectors follow $w_\lambda^{\{7,8,9\}} = (1-\lambda)(w_A^{\{7,8,9\}}) + \lambda(w_B^{\{7,8,9\}})$. When $\lambda = 0$, the output is the original latent vectors of the source image. When $\lambda = 1$, the latent vectors on layers 7, 8, and 9 of the source image are replaced by those of the target image (Figure 4(a) middle panel).

Figure 1(c) shows examples of layer-wise manipulation in the $W+$ latent space. We observed the emergence of three salient attributes when we performed image morphing at different layers: early-layers (layers 1 to 6) determined the shape and orientation of the soap, middle-layers (layers 7 to 9) effectively changed the material (e.g., transformed from glycerin to milky soap, and vice versa), and later-layers (layers 10 to 18) primarily changed the body color of the object. This shows that StyleGAN's deep generative representation mechanistically disentangles the scene attributes without external supervision.

Perceptual evaluation of emerged scene attributes

To examine how naive observers interpret the scene attributes that emerged in layer-wise representation of $W+$ space, we created image sequences by morphing between selected pairs of images on three different sets of layers (early, middle, later). For each layer manipulation (equation 1), we selected source-target image pairs under three material conditions: opaque-translucent (OT), opaque-opaque (OO), and translucent-translucent (TT). Together, we sampled 450 image sequences (see Methods). Figure 4(a) shows examples of image sequences generated from the three layer-manipulation methods (top panel for early-layer manipulation, middle panel for middle-layer manipulation, bottom panel for later-layer manipulation) under three source-target material conditions. For each image sequence, observers were asked to select the "ONE MOST prominent visual attribute changed from left to right" (Figure 4(b)). Figure 4(c) illustrates which attribute observers chose as the dominant change for each layer manipulation. The heat maps show that observers unanimously agreed that manipulation on early-layers changed the shape of the objects (approximately 97%), independently of the material condition of the target and source images. Observers also agreed that manipulation on middle-layers mainly changed the translucent appearance of the objects (approximately 75%) for opaque-translucent pairs. When the source and the target have similar materials (OO and TT pairs), the middle-layer manipulation led to less obvious change of material appearance (approximately 35%), and observers also selected lighting or color as the main variation factor depending on the scenes. For example, when we morphed two translucent soaps, either material or lighting could be viewed by observers as the dominant change (Figure 4(a), middle panel, third row). Lastly, observers mostly agreed that manipulation on later-layers changed the body color of the objects across material conditions (approximately 70%). We conducted a Bayesian multilevel multinomial logistic regression on the behavioral data, and analysis results coincided with our observations[98]. All three layer-manipulation methods are credible parameters for the estimation of the most prominent scene attribute. We also examined the conditional effects of layer manipulations. For the early-layer manipulation, the estimated probability of selecting "shape/orientation" was close to 1 across all three types of source-target pairs. For the middle-layer manipulation applied on OT pairs, the estimated probability of selecting "material" was 77.9% (95% highest density interval, [69.5%, 84.5%]) (Supplementary Figure S.3). These results show that the scene attributes disentangled in the latent space are perceptually meaningful and each attribute can be separately controlled in different layers' latent vectors.

We also observed some participants chose lighting as the dominant change resulting from the middle-layer manipulation for similar target and source materials, suggesting that the middle-layers of the latent space can also represent lighting to some degree. The effect of lighting may have two aspects: the direction and the environment of lighting. The direction of lighting, expressed in the images through the position and shape of the cast shadow, was captured in a subset of earlier layers

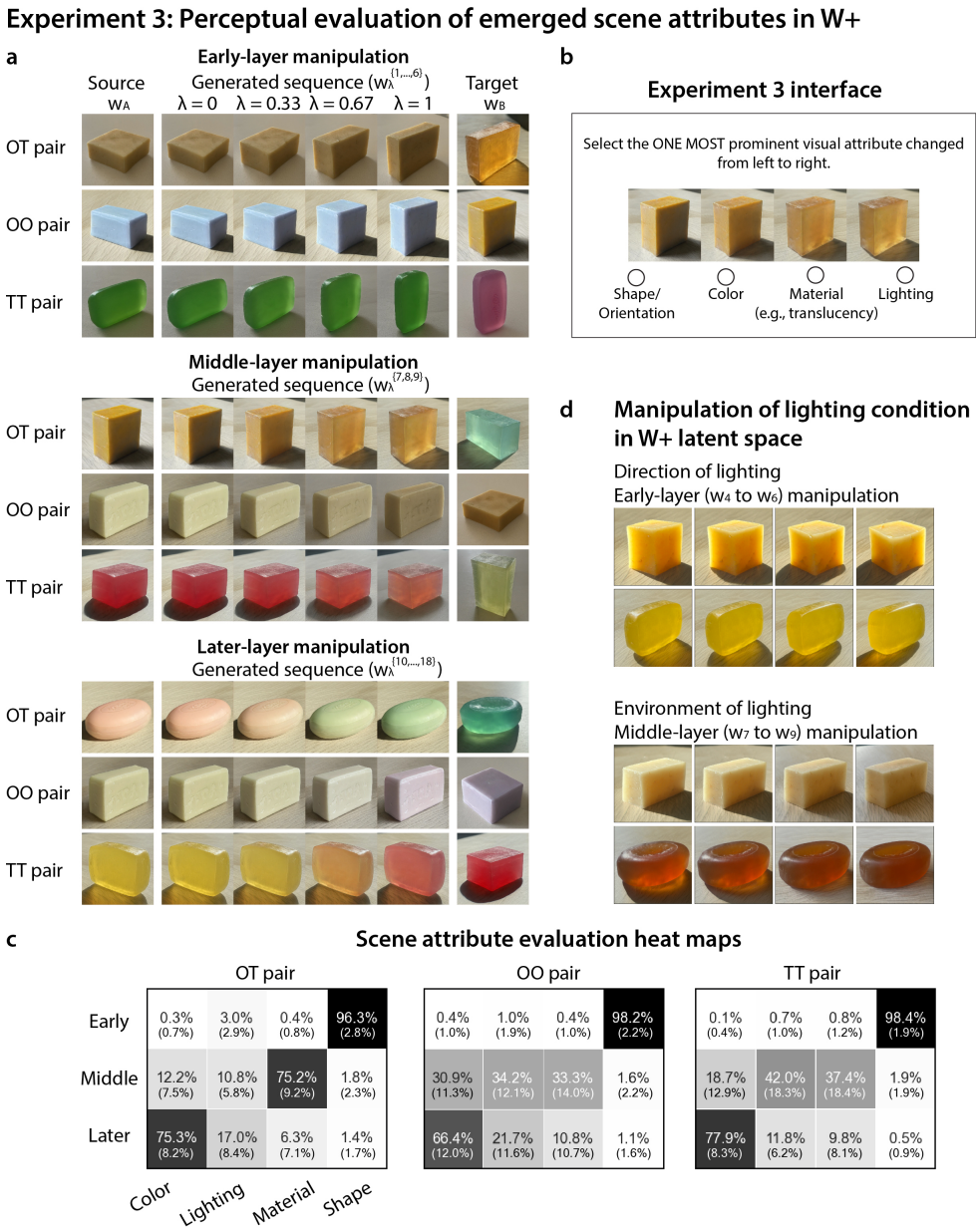

Figure 4. Experiment 3: Perceptual evaluation of emerged scene attributes in the layer-wise latent space. a, Examples of morphed image sequences used in Experiment 3, generated by linearly interpolating between the latent codes of source (w_A) and target (w_B) separately at the early (w_1 to w_6), middle (w_7 to w_9), and later-layers (w_{10} to w_{18}). The λ is the interpolation step from the source image in the linear interpolation. Source-target pairs were picked under three conditions based on soap's material properties: opaque-translucent (OT), opaque-opaque (OO), and translucent-translucent (TT). **b,** The user interface of Experiment 3. **c,** The perceptual results on how different layers correspond to scene attributes. The number in each cell represents the average percentage of times observers chose a visual attribute as the most prominent one that changes in the image sequence generated by the corresponding layer manipulation. The standard deviation across observers is shown in parentheses. Each row of the heat map accounts for 50 image sequences. **d,** The representation of lighting in the latent space. Top panel: manipulation of early-layers (layers 4 to 6) also changes the direction of lighting. From left to right, the lighting direction rotates clockwise. Bottom panel: manipulation of middle-layers (layers 7 to 9) alters the environment of lighting. From left to right, the strength of backlighting gradually decreases.

(layers 4, 5, and 6). The top panel in Figure 4(d) shows manipulating such layers conveyed the impression of rotating the light source clockwise. On the other hand, the environment of lighting (e.g., sunny versus overcast) affects the color distribution of objects in an image. This effect is manifested in the middle-layers (layers 7 to 9). The bottom panel in Figure 4(d) shows that manipulating these layers yielded the impression of varying the strength of backlighting. This observation is consistent with previous findings that lighting environment affects translucent material perception in that objects under strong backlighting tend to appear more translucent[7,21].

The middle-layers of the model's latent space capture human translucency perception

Our next goal was to examine whether the middle-layers of the latent space could capture human translucency perception. To derive quantitative translucency prediction from the model, we trained a linear support vector machine (SVM) classifier to find the decision boundary with each layer of the images' latent codes that best distinguishes translucent soaps from opaque ones. We manually labeled the soaps into two categories based on their listed ingredients: milky and glycerin. We sampled 1000 real photos from the TID dataset. Half were of glycerin soaps, and the other half were of milky soaps. The shape, lighting condition, and body color largely varied across instances. After obtaining the corresponding $W+$ latent codes of the embedding of real photos through the pSp encoder, we extracted their latent vector at each of the 18 layers to train a linear SVM classifier. Therefore, we had 18 distinct decision boundaries. Figure 5(a) illustrates the trained decision boundary (d_i) using the i-th layer of $W+$.

Next, we computed the SVM model predictions and compared them with the human attribute ratings measured in Experiment 2. Specifically, we obtained 18 distinct model prediction values from each layer's latent vector for the 150 generated images used in Experiment 2. For a given image with its i-th layer latent vector, we measured its distance from the learned decision boundary d_i (normalized to 0 to 1 range). For example, as shown in the middle columns in Figure 5(b), using an image's layer-9 latent vector, we could plot its model prediction value against the mean normalized human attribute rating for translucency, see-throughness, and glow, respectively. The Pearson correlation between the model prediction and perceptual rating (r_{hc}) is calculated for each attribute. The data shows that predictions from a middle layer (e.g., layer-9) strongly correlate with human material attribute ratings while the predictions from an earlier layer (e.g., layer-6) and a later layer (e.g., layer-12) have relatively weak correlation with perception. By repeating this step for each layer, we obtained the correlation coefficients between each layer's model prediction and the perceptual ratings. Figure 5(c) shows the tuning curve of correlation coefficients over the layers. The correlations r_{hc} peaked at the middle-layers (layers 7, 8, and 9), implying that these layers may most effectively encode visual information that observers also utilized for translucency perception.

The trained SVM serves as a general guidance for material appearance editing. The decision boundaries from middle-layers reflect the linear separability of intrinsic material in the latent space. The normal to the decision boundary becomes an interpretable direction that captures the variation of material appearance. As shown in Figure 5(d)'s top row, manipulating the 9th layer's latent vector of a given image (left end) along the positive direction of the normal to d_9 persuasively made the material more milky and opaque, without changing the object's shape. Conversely, moving to the negative direction made an opaque soap more translucent. In contrast, manipulating a single latent vector from early or later layers (e.g., layer-12) along the found decision boundary's normal did not lead to effective modification of the material appearance (Figure 5(d) bottom). The manipulation on all layers can be found in the Supplementary Figure S.4 and Figure S.5.

Visualization of intermediate outputs of the generator network

To break down how translucent appearance is created in the final output, we examined feature maps generated in the intermediate stages of the synthesis network of StyleGAN2-ADA[86,99]. The generator starts from a learned constant input of size $4 \times 4 \times 512$ and gradually expands the spatial resolution via affine transformation layers. At each resolution, from 8×8 to 1024×1024, an additional single convolution layer (tRGB layer) transforms the feature maps into the RGB image. As shown in Figure 6, we visualized the intermediate steps to generate the images of soaps with their corresponding $W+$ latent codes.

Consistent with the discovery of emerged scene attributes in the latent space, the early-layers (w_1 to w_6), spanning from 8×8 to 16×16 resolutions, formed the general shape and contour of the object. The middle-layers, with layer 7 and 8 (w_7 and w_8) at 32×32 resolution and layer 9 (w_9) at 64×64 resolution, established the critical features of translucency. The image contrast and color variation across the volume of the soap in the 64×64 resolution images gave the impression of "glow", which is useful to distinguish translucent materials from opaque ones. At 128×128 resolution (layers 11 and 12), surface reflectance properties such as specular highlights and caustics were further specified. The later layers (w_{13} to w_{18}), from 256×256 to 1024×1024 resolutions, enriched the details of lighting environment and color scheme, delivering more appealing material appearance. This suggests that latent image features at relatively coarse spatial scale are sufficient to capture the visual impression of translucent materials.

Diagnostic image features for translucency

To understand what information the intermediate generative representation encodes, we explored the image descriptors for the tRGB layer's representation with the middle spatial scale, which is sensitive to translucency (Figure 7). Inspired by sparse

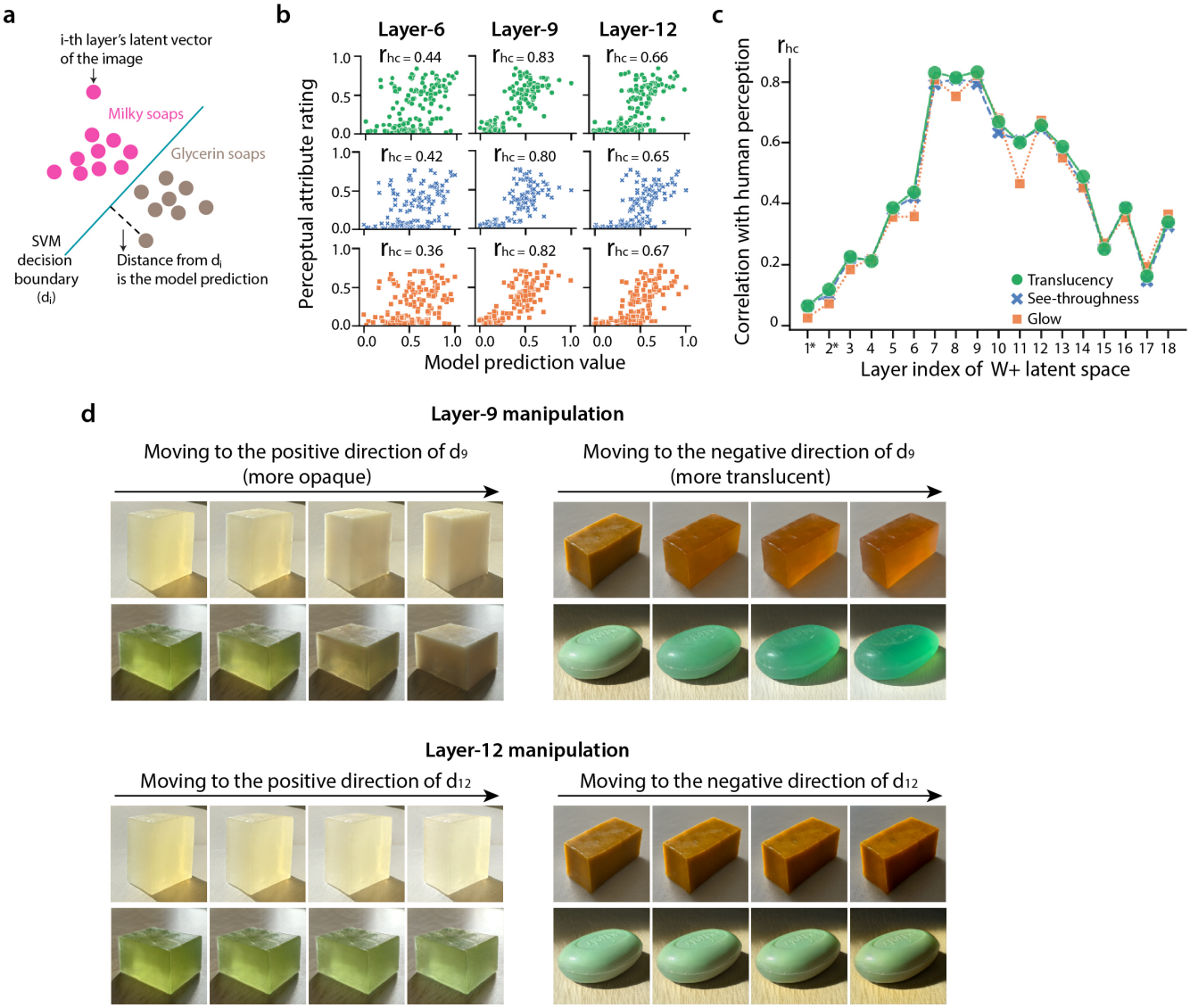

Figure 5. The middle-layers of $W+$ latent space can effectively modulate translucency of generated images and predict human perception. a, Illustration of a trained layer-specific supported vector machine (SVM) classifier for the milky-versus-glycerin soap discrimination. **b,** The scatter plots show the model prediction value computed and the human mean normalized attribute ratings for each generated image in Experiment 2. Green, blue, and orange colors represent the data for translucency, see-throughness, and glow, respectively. **c,** The tuning curve of correlation coefficients (correlation between model prediction and human perceptual rating, r_{hc}) over all layers in the $W+$ latent space. Model prediction values using the middle-layers' decision boundaries (d_7, d_8, and d_9) strongly correlate with human attribute ratings. "*" indicates the correlations at that layer are statistically insignificant at the 95% confidence level. **d,** Examples of translucency-modulated sequences. Top: Manipulating the layer-9 latent vector of the original image (left end) along the normal of the learned decision boundary has a coherent effect on the translucent material appearance of the object. Left: moving to the positive direction of the normal of the decision boundary makes the soap appear more opaque. Right: moving to the negative direction of the normal of the decision boundary makes the soap appear more translucent. Bottom: Manipulating the layer-12 latent vector of the original image along the normal of the learned decision boundary does not fundamentally change the translucent appearance.

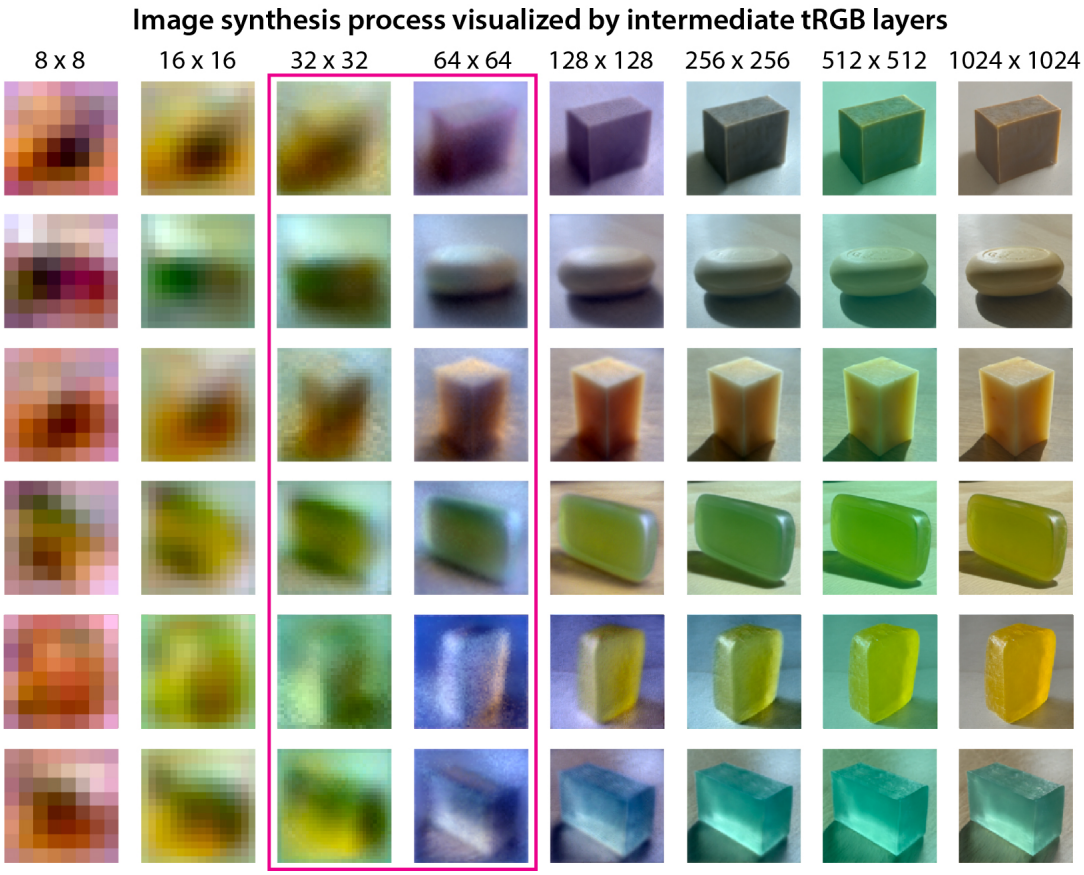

Figure 6. Visualization of the generative process of the network. Impression of translucency emerges at early stages of the image synthesis process while more details of the appearance are added in the later stages. Each row corresponds to the intermediate generative outputs from a sequence of tRGB layers at different spatial resolutions in the StyleGAN2-ADA's generative network. Translucency-related features are established as early as 32 pixels × 32 pixels (layers 7 and 8) and 64 pixels × 64 pixels (layer 9). The surface reflective properties such as specular highlights are only clearly visible at 128 pixels × 128 pixels (layers 11 to 12). The body color of the soap was finalized at the resolution of 1024 pixels × 1024 pixels (layers 17 to 18).

coding used in understanding natural images[100–103], we applied independent component analysis (ICA)[104] on local regions of the intermediate tRGB images to investigate the efficient representation of translucent appearances. Specifically, based on the results of Experiment 2, we created a new set of high-translucency generated images and extracted the intermediate tRGB images with 64 pixels × 64 pixels, whose layer is sensitive to translucency emergence. While keeping the relative kernel size constant with the StyleGAN's convolution process, we applied FastICA to learn 64 basis functions[104].

In the learned representation (Figure 7(b)), the activation features are chromatic or achromatic with a variety of orientations. Figure 7(c) demonstrates the results of applying three-dimensional convolution of each of the 64 kernels to a real photograph of translucent soap. While luminance kernels provide information of object contours, shadow boundaries, and specular reflectance, chromatic kernels reveal subtle image features indicating translucency, such as color gradients around the edges and corners. Figure 7(d) shows examples of filtering results on a few translucent and opaque objects. For example, applying the oriented chromatic kernels (rows 1–4 in the matrix of convolution results) on the transparent soaps (columns 1 and 3) activated patterns of color variations on the caustics, which are not present in the more opaque soap (rows 1 and 2, column 2). Next, the red-green chromatic kernels also detected the internal "glow" of the translucent soaps. For example, the convolution results on the yellow translucent soap (column 4) showed the spatial gradient of saturation near the edges (row 1, column 4). At the same time, the resulting image also revealed the "glowing edge" on the same soap (row 1, column 4). Notably, the orientation-free chromatic kernels revealed the color variation over a relatively coarse spatial scale across the object, which might be diagnostic of translucency (row 5, column 4). Furthermore, these translucency-related features could not be obtained by the basis functions extracted from the coarser intermediate representation (Supplementary Figure S.7). Together, our results indicate that the oriented chromatic kernels with mid-to-low spatial frequency can be diagnostic for translucent appearance.

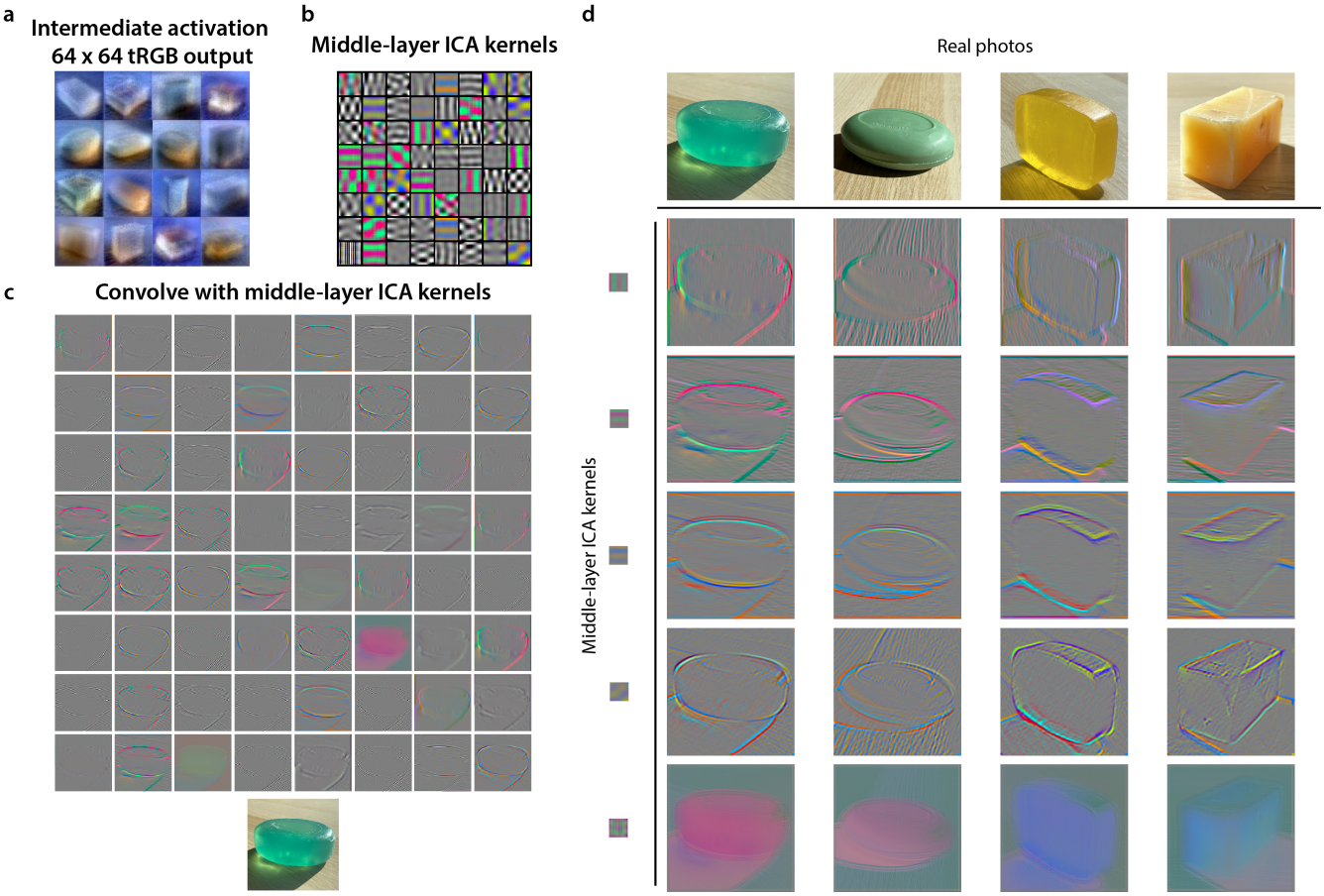

Figure 7. Visualization of features for translucency. a, The intermediate generative results (tRGB layer output at 64 pixels $\times$ 64 pixels resolution) of the images from the high-translucency dataset. The images are resized to 128 pixels $\times$ 128 pixels for display. **b,** Middle-layer ICA kernels obtained by training a system of 64 basis functions on 24 $\times$ 24 image patches extracted from images in (a). The kernels are of size 24 $\times$ 24. **c,** Visualization of applying three-dimensional convolution of the individual ICA kernels in (b) on a real photograph of translucent soap. **d,** The resulting filtered images of four different soaps with selected chromatic kernels. The mid-to-low spatial frequency chromatic kernels can capture features of translucency such as "chromatic caustics" (row 2, column 1), "glowing edges" (row 1, column 4), and "glow from inside" (row 1, column 1, 3, and 4). The orientation-free kernel in the last row reveals the variation of color over a relatively coarse spatial scale, which is also diagnostic of translucency.

Discussion

We presented a deep image generation model trained with natural photographs to obtain a compact layer-wise latent space that can capture human perception of translucency. Our study demonstrates that the learned latent space spontaneously disentangles salient visual attributes and captures the latent dimensions of translucent appearances. Notably, we find the represented scene attributes are scale-specific, where early-layers represent shape, middle-layers represent translucency, and later-layers represent body color. The middle-layers of the latent space can successfully predict human perception of translucency in various generated images. Our findings suggest that humans might use a scale-specific structure to characterize visual information from retinal images, facilitating the representation of materials for robustly estimating their attributes under various contexts. Our framework could serve as an effective method for discovering generalized image features across a high degree of perceptual variability of materials.

The image generation process of our model (Figure 6) resembles the strategy an artist uses to paint a translucent object by structurally depicting the observed visual attributes. Therefore, the representational system the model learns might be similar to those of the mental process of painting. The Dutch artists in 17th century were capable of painting vivid translucent objects on canvas by depicting the critical image features that trigger a visual impression of translucent properties, without strictly conforming to physical laws[16,105]. Imagine an artist painting a grape on a dining table (e.g., *Still Life with Oysters*

and Grapes, Jan Davidsz. de Heem, 1653). As a first step, before exquisitely adding any details, the artist often starts with carving out the object's contour. After the general shape is set, colors are gradually filled in to mimic shadow and shading to yield a first impression of the 3D shape and reflect the lighting condition in the scene. More details are added to depict surface reflections and caustics. The artist can continually perfect the painting by adding fine details to deliver a more convincing material quality. The combination of multiple levels of detail in the painting contributes to the formation of translucency appearance. Our model generates translucency appearance in a similar scale-specific manner. Given that previous studies have also shown that the scale-specific process has a role in material perception[106–108], discovering and formulating the structure of visual information in a scale-specific manner might be helpful for recoding the complexity of material appearances to obtain an efficient representation.

In our study, we use soap as a medium to illustrate the possibility of learning a semantically meaningful representation of natural images of materials. Although it is possible that the exact meaning represented by the latent space differs across the training datasets, the corresponding latent space will still disentangle scene attributes with various abstract levels arising from the scale-specific image features. We expect the middle-layers (medium coarseness) to represent volumetric material appearance even if the model is trained with image datasets of other materials. For our dataset, we demonstrated that scale-specific image features can be separately controlled. The translucent appearance (i.e., associated with middle resolution features) and the body color (i.e., associated with fine resolution features) can be directly manipulated without changing the shape of the object (i.e., associated with coarse resolution features). The observations from our finding also align with previous investigations of the StyleGAN's representative power of its layer-wise latent space. In the generation of images of human faces, "styles" of coarse spatial resolutions (2 to 4 cycles/image) correspond to high-level aspects such as pose and face shape, "styles" of middle resolutions (8 to 16 cycles/image) control smaller scales of facial features and hairstyle, and "styles" of fine resolutions (32 to 512 cycles/image) contribute to microstructures and color scheme[86]. Likewise, in indoor scene synthesis, the latent space could separately control spatial layout of the room (coarse), categorical objects in the scene (middle), and color scheme (fine)[92]. Widening the scope beyond soaps, it is feasible to use TAG to model the image data of broad categories of materials.

Our study constitutes a break from the long history of studying material perception using well-controlled computer-rendered images. We discovered critical image features that are diagnostic of translucency across diverse geometries and lightings by applying unsupervised learning schemes on a large-scale dataset of natural photographs of translucent objects without specifically constrained physical environments. Some of our found image features could be robust indicators for translucent materials, and they confirm previous empirical findings. For instance, the edge intensity profile on translucent objects has been found to be different from those of opaque ones[109]. Our ICA analysis shows that oriented chromatic kernels can detect complex patterns along the translucent edges (see Figure 7(d) row 2, column 4). In addition, such chromatic kernels also capture the effect of "glow", an important feature characterizing the spatial distribution of color of translucent materials[12,21]. Furthermore, our results indicate that the presence of caustic patterns can be an important cue for translucency perception[110].

We also discover the intricate role of color in translucent appearance. Most of the previous works explored the effect of color on material perception and recognition by manipulating the color/luminance distribution of material images[12,21,63,70,111,112]. For example, converting color translucent images to grayscale ones decreases perceived translucency[12,21,63,71]. However, it is still unclear how the visual system functionally processes color information for material perception. Our findings, based on a data-driven approach, suggest that two functional aspects lie in color translucency processing: body color and spatial color processes. The body color represents the color of the matte component of surface reflectance, which is usually determined by the color of the dye used to make the soap. The latent space in our model can represent the body color of soap separately from material appearance. By manipulating the middle-layers of the latent code, we can create images of objects with different types of translucent appearances but similar body color. This suggests the model can establish a translucency impression without varying the body color. The other aspect is the spatial variation of color over the volume and surface of an object (e.g., the color gradient within an object due to light scattering and absorption). This "spatial color" is crucial for providing the translucent appearance in the middle-layers (Figure 1(c), top panel) and can be detected by the chromatic kernels with the mid-to-low scale (Figure 7(d)). Notably, this color process is scale-specific, i.e., a coarser kernel cannot detect the spatial color variation of translucency (Supplementary Material Figure S.7). Furthermore, the spatial color can be independent of the white-balancing process because the middle-layers in our model do not fix the white point in the scene (Figure 6). The finding suggests that the processing of saturation and hue based on a white point may not be necessary for this spatial color process. As such a spatial color process has been little understood in color vision literature [113,114], our work might provide novel directions for probing the role of color in material perception and other high-level visual processing in the brain.

The deep generative network (StyleGAN) is not designed to emulate biological vision systems, even though the elementary functional mechanisms (e.g., convolution, nonlinearity) are inspired by biological brains[115–118]. Therefore, we do not assume the learning process of StyleGAN is necessarily the mechanism of human material perception. Here, we take StyleGAN's representative power to model the feature space of diverse material appearances and discover the latent image features that humans might have used to estimate material properties in the natural scenes by comparing with psychophysical results. We

also acknowledge that the image features we learned in our model are still considered mid- to low-level sensory information. Future models need to be developed to address the role of top-down influence, such as context, object identity, and individual experience, on material perception.

One extension of the current work is to use our stimuli to measure brain responses to translucent material properties. One large obstacle to probing the neural correlates of material perception has been the lack of an effective way to manipulate the stimuli that isolate the effects of various external factors on material appearance while keeping the image's appearance natural and realistic. Our material manipulation through the latent space illustrates a novel and efficient approach for conditionally creating stimuli with translucent appearance resulting from a specific combination of scene attributes. Moreover, the discovered latent representation can be valuable for encoding/decoding investigations in brain-imaging studies to probe the interaction between neural representations of 3D shape, color, and materials, thus providing an efficient tool to discover the neural correlates of material perception. More generally, the approach we take here—using StyleGAN to derive a latent representation for translucency perception—is widely applicable to discover perceptually relevant features for a variety of visual inference tasks that deal with complex physical stimuli.

Methods

Translucent Image Datasets (TID)

Our customized image dataset of translucent objects has 8085 photographs of soaps. The dataset was created by photographing a variety of real-world soaps in natural backgrounds. We collected 60 unique soaps that included diverse materials, geometries, surface relief, and colors. We used an iPhone v12 mini smartphone to photograph our collection of soaps under various lighting environments and viewpoints at a relatively fixed distance, and built a dataset of high resolution images (1024 pixels $\times$ 1024 pixels JPEG images). In each photograph, the object was centered in the image. We did not intentionally balance the dataset on the distribution of shape, body color, lighting environment, and viewpoint. To our knowledge, this is the first large-scale natural image dataset of translucent materials and one of few image datasets of real-world materials.

Unsupervised learning framework: Translucent Appearance Generation (TAG) model

Deep generative network StyleGAN2-ADA We trained StyleGAN2-ADA, on the full TID dataset using the TensorFlow implementation of the model available at https://github.com/NVlabs/stylegan2-ada. StyleGAN2-ADA consists of two networks trained through a competitive process: a style-based generator, and a discriminator. The generator creates "fake" images, with the aim of synthesizing realistic images of soaps. The discriminator receives both "fake" and real images, and aims to distinguish them. As the training progresses, both the generator and the discriminator improve until the "fake" images are indistinguishable from the real ones. The training of the style-based generator involves two latent spaces. There is a input latent space Z that is normally distributed. Hence, a sequence of eight fully-connect layers transforms Z to an intermediate latent space W. The dimensions for both Z and W spaces are 512. With the 1024 pixels $\times$ 1024 pixels output, the generator starts with a constant input of size $4 \times 4 \times 512$ and gradually adjusts the "style" of the image at each of 18 convolution layers based on the latent vector[86]. For every major resolution (every resolution from 4 pixels $\times$ 4 pixels to 1024 pixels $\times$ 1024 pixels), there are two convolution layers for feature map synthesis and a single convolution layer (i.e. tRGB layer) that converts the output to a RGB image. Weight modulation and demodulation are applied in all convolution layers, except for the output tRGB layers[87]. At each convolution layer i, the generator receives the input through "style", which is a learned affine transformation from the 512-dimensional latent vector $w \in W$. More explicitly, when generating an image from W space, the same vector w is used for all convolution layers.

Using the network architecture of StyleGAN2, StyleGAN2-ADA inherently applies a wide range of augmentations on the input data to prevent the discriminator from overfitting, while ensuring that none of the augmentations leak to the generated images. During training, each image is processed by a series of transformation in a fixed order, and each transformation is randomly applied with probability $p \in [0, 1]$, which is adaptively adjusted to counter the effect of overfitting. This variant is named Adaptive Discriminator Augmentation (ADA)[88]. In practice, we allowed the following set of transformations: pixel blitting (x-flip, 90-degree rotatation, integer translation), general geometric transformation (isotropic scaling, anisotropic scaling, fractional translation), and color transformation (brightness, luma flip, hue, saturation). The total length of training of StyleGAN2-ADA is defined by the "the total number of real images", since the randomization of transformation is done separately for each image in a minibatch. We trained the model on one Tesla V100 GPU for a total length of 3,836,000 images, using the recommended learning rate of 0.002 and R_1 regularization of 10[88] for generating 1024 pixels $\times$ 1024 pixels resolution outputs. The FID (Fréchet Inception Distance), KID (Kernel Inception Distance), and recall for the trained model are 13.07, 0.0038 and 0.330 respectively.

pixel2style2pixel (pSp) encoder Upon training the StyleGAN2-ADA, we separately trained a pSp encoder on 80% of randomly sampled images from the TID dataset and validated on the rest of the images. We implemented the pSp encoder

based on the code released by https://github.com/eladrich/pixel2style2pixel[89]. The pSp encoder aims to efficiently embed a real photo into the StyleGAN's extended intermediate latent space $W+$[94]. Unlike W space, $W+$ is a concatenation of 18 different 512-dimensional vectors (w_1 to w_{18}), one for each convolution layer of StyleGAN2-ADA generator. Given a real image, we can map it to the latent space $W+$ and create its reconstruction image by feeding the obtained latent code into our pre-trained StyleGAN2-ADA generator.

The pSp encoder is built on a feature pyramid network[119] to generate three levels of feature maps (coarse, medium and fine)[86] from which 18 latent vectors of $W+$ were extracted using a small fully convolutional network (map2style). Latent vectors w_1 and w_2 are generated from the small feature map, w_3 to w_6 are generated from the medium feature map, and w_7 to w_{18} are generated from the large feature map. The latent vectors are then injected into the pre-trained StyleGAN2-ADA generator corresponding to their spatial scales to synthesize the reconstructed image. The feature pyramid network and the map2style networks are updated through backpropagation to learn to generate latent vectors which map to reconstructed images that are perceptually similar to the input real images. The architecture is illustrated in Figure 1(b).

The entire framework was trained on a set of loss function to encourage the accurate reconstruction of the real photos: pixel-wise loss (L_2), LPIPS loss (L_{LPIPS}), and regularization loss (L_{reg}). For an input image x, the total loss is defined as: $L(x) = \lambda_1 L_2(x) + \lambda_2 L_{LPIPS}(x) + \lambda_3 L_{reg}(x)$, where λ_1, λ_2, and λ_3 are constants defining the loss weights. Here, we set $\lambda_1 = 1$, $\lambda_2 = 0.8$, $\lambda_3 = 0.005$. The maximum number of training steps was set at 10000, and the model leading to the minimum total loss was consistently updated. We trained the model with one Tesla V100 GPU for 2 GPU days, and the model optimized at training step 9000 was used for the rest of the study. The total loss was 0.181.

Psychophysical experiment

Participants The same group of twenty participants completed Experiment 1 and 2 (N=20, median age, 20; age range, 18-27, 12 female, 8 male). They completed the experiments in one lab-based session. Another group of twenty participants completed Experiment 3 (N=20; median age, 21; age range, 18–27; 10 female, 10 male). Five individuals participated in all experiments. Observers received no information about hypotheses of the experiments. No statistical methods were used to predetermine sample sizes, but our sample sizes are consistent with those reported in previous publications of material perception measured in the laboratory[68,79,120]. All of the observers had normal or corrected-to-normal visual acuity and normal color vision. Participants were primarily undergraduate students from American University. The observers were given written informed consent and were compensated with either research course credits from American University or with $16 per hour. The experiments were conducted in accordance with the Declaration of Helsinki, with prior approval from the American University. All the experimental designs involving human participants were approved by the Institutional Review Board at American University.

Psychophysical procedures The psychophysical experiments were conducted in a dimly lit laboratory room. Observers sat approximately 7 inch away from the monitor and were given no fixation instructions. The stimuli were presented on an Apple iMac computer with a 27-inch Retina Display with a resolution of 5120 pixels $\times$ 2880 pixels and a refresh rate of 60 Hz. PsychoPy v.2021.1.2 was used to present the stimuli and collect the data[121]. At the beginning of each experiment, observers were given experiment-specific instructions and demos.

Experiment 1: Real-vs-generated discrimination

Stimuli To avoid using the same images as those in the model training process, we took 300 new photographs of our collection of soaps. We then split these photographs equally into two groups (A and B), which similarly capture the variety of materials, lighting fields, and view points. The 150 real photographs from Group A and the 150 generated images obtained from Group B were used as the stimuli for Experiment 1. Specifically, photographs from Group B were first encoded into the $W+$ latent space through the pSp encoder, and then were reconstructed through our trained StyleGAN2-ADA generator. In this way, we obtained generated images that cover the diverse samples of appearances of soaps in our dataset. Examples of stimuli are shown in Figure 2(a). All images were presented in size 1024 pixels $\times$ 1024 pixels.

Experimental procedure We first gave a brief introduction to each observer of how the real photographs and generated images of soaps were created. The observers were told that the "Real photographs of the soaps (Real) were taken using a smartphone camera, and the generated images were produced from a computer algorithm (Generated). The generated images would try to resemble the visual appearances of the object in the real photos." Afterwards, the observers were presented a series of images and were asked to judge whether the stimulus is Real or Generated. Each image was shortly displayed for one second, and then the observer made the judgement by a key press. Observers were given the prior knowledge that 50% of the stimuli were Real. We conducted the experiment with two repeats. In repeat 1, observers judged 300 images of real and generated images with a pre-randomized order in two blocks of 150 trials. They then completed another repeat of the same 300 images but with a different pre-randomized order. The experimental procedure is shown in Figure 2(b).

Experiment 2: Material attribute rating

Stimuli Experiment 2 stimuli were the same 300 images of real photographs and model generated images of soaps as in Experiment 1.

Experimental procedure Before the experiment started, we introduced the concept of translucency to the observers by showing them a simplified illustration of the subsurface scattering process (Supplementary Figure S.2). In Experiment 2, observers were asked to rate the material attributes of the images. On each trial, the observers rated each attribute using a seven-point scale (7 means high, 1 means low) by adjusting the slider (Figure 3(a)). They had unlimited time to make the judgements. The 300 images were equally split into two blocks, and presented in a pre-randomized order. This experiment was conducted with only one repeat.

Observers were provided with the definition of the material attributes as the following:

- Translucency: To what degree the object appears to be translucent.
- See-throughness: To what degree the object allows light to penetrate through.
- Glow: To what degree the object appears to glow light from inside.

Experiment 3: Perceptual evaluation of emerged scene attributes in the layer-wise latent space.

Stimuli We created image sequences by applying morphing between the source image A (w_A) and the target image B (w_B) using Equation 1. The morphing was separately applied on three sets of layers of the latent space: early-layers (layer 1 to 6), middle-layers (layer 7 to 9), and later-layers (layer 10 to 18) , with equal interpolation steps. To generate an image sequence, the interpolation step (λ), was set to have four values: 0, 0.33, 0.67, and 1 (see Figure 4(a)).

We picked 24 soaps from the TID dataset, half were opaque milky soaps (generally with low translucency, i.e., opaque) and half were translucent glycerin soaps (generally with high translucency, i.e., translucent). With these images of soaps, source-target image pairs were formed under three conditions: opaque-translucent (OT), opaque-opaque (OO), and translucent-translucent (TT). For each condition of source-target pairs, we created image sequences based on the morphing on the early-, middle-, and later-layers respectively, and then randomly sampled 50 sequences as stimuli. This led to 3 (condition of source-target pair) $\times$ 3 (layer-manipulation method) $\times$ 50 image sequences in total (Figure 4(a)). All individual images in the image sequences were resized to 256 pixels $\times$ 256 pixels for display.

Experimental procedure At the beginning of the experiment, we showed observers a few samples of real soaps of different materials, shapes, and body colors with the goal of illustrating the effects of these scene attributes on material appearance.

The observers viewed 450 image sequences. For each image sequence, observers selected the "One most prominent visual attribute changed from left to right" from one of the following: shape/orientation, color, material (e.g., translucency), and lighting. The image sequences were equally split into three blocks and presented in a pre-randomized order. Observers had unlimited time to complete their judgement on each trial (Figure 4(b)). This experiment was conducted with one repeat.

Computing translucency decision boundaries from latent code

We trained binary SVM based on the latent vector of each layer of the latent space $W+$ to classify the material of the soap in the TAG generated image as either "milky" or "glycerin". The trained SVM classifiers were then used to generate model predictions on a continuous scale. We randomly sampled 500 real photographs of "milky" soaps and another 500 photos of "glycerin" soaps from the TID. We used 60% of images for training and the rest for validation. The 1000 photos were first embedded into the $W+$ latent space to obtain their corresponding 18×512 dimensional latent codes through our trained pSp encoder. Since the latent space contains 18 layers, we trained 18 SVM models based on each layer's latent vectors of the embedded images. In other words, there were 18 different feature matrices, each with dimension of $n \times 512$, where n is the number of training samples. We implemented LinearSVC from *scikit-learn* for model fitting, and used a relatively strong regularization ($C \in [0.001, 0.1]$) to reduce overfitting[122]. Hence, we obtained a linear decision boundary d_i for the i-th layer's latent vectors. We then computed the model prediction values of the 150 generated images used in Experiment 2. With the $w_+ \in W+$ latent code of a generated image, we extracted its i-th layer's latent vector and measured its distance from d_i. For each layer, the model prediction value, which is the normalized distance from d_i, was compared to human perceptual rating data from Experiment 2.

Independent Component Analysis (ICA) for the intermediate generative representation

Based on the results of Experiment 2, we selected 40 generated images with the highest translucency ratings (high-translucency). Meanwhile, we selected another 40 generated images of soaps with various shapes, orientations, and lighting environments. Then, we fully paired these 80 images (source) with the 40 high-translucency images (target). To create a new "high"

translucency image, we replaced the middle-layer (layer 7 to 9) latent vectors of the source image with those of the target and used the resultant latent code to generate the corresponding image through the generator. Then, we extracted the intermediate generated result from the tRGB layer corresponding to 64 pixels $\times$ 64 pixels spatial scale of which translucency is established. We repeated this step to obtain 3160 "high" translucency images at resolution of 64 pixels $\times$ 64 pixels (Figure 7(a)). For each image in the "high" translucency dataset, we first resized it to 512 pixels $\times$ 512 pixels resolution and sampled 10 image patches of 24 pixels $\times$ 24 pixels from random locations. FastICA was then applied on the 3160 $\times$ 10 image patches to learn 64 basis functions (i.e., middle-layer ICA kernels)[104]. For the learning of middle-layer ICA kernels, we also conducted the FastICA with different sampling of the image patches with 64 and 100 components (Supplementary Figure S.6.)

Statistical analysis

We used Bayesian multilevel multinomial logistic regression to model the psychophysical results from Experiment 3[98,123]. The goal is to examine whether the prominent scene attributes judged by the observers can be predicted by the layer-manipulation methods. We implemented the *brms* library supported in *R* for the analysis. The model's dependent variable is the scene attribute (i.e., shape/orientation, color, material, and lighting). The predictors include the layer-manipulation methods (i.e., early-layers manipulation, middle-layers manipulation, and later-layers manipulation), the type of source-target pair (i.e., opaque-opaque (OO), opaque-translucent (OT), and translucent-translucent (TT)), and the interaction between these two factors, while considering the individual observer as a grouping variable. Three Markov chains were used for the parameter posterior distribution estimation, with 8000 iterations for each chain of the Markov Chain Monte Carlo (MCMC) algorithm. We assumed a uniform distribution for the priors of the parameters. The complete results of the analysis can be found in the Supplementary Figure S.3 and Table 1.

References

1. Adelson, E. H. On seeing stuff: the perception of materials by humans and machines. In *Human vision and electronic imaging VI*, vol. 4299, 1–12 (SPIE, 2001).
2. Tiest, W. M. B. Tactual perception of material properties. *Vis. Res.* **50**, 2775–2782 (2010).
3. Xiao, B., Bi, W., Jia, X., Wei, H. & Adelson, E. H. Can you see what you feel? color and folding properties affect visual–tactile material discrimination of fabrics. *J. Vis.* **16**, 34–34 (2016).
4. Komatsu, H. & Goda, N. Neural mechanisms of material perception: Quest on shitsukan. *Neuroscience* **392**, 329–347 (2018).
5. Schmid, A. C. & Doerschner, K. Representing stuff in the human brain. *Curr. Opin. Behav. Sci.* **30**, 178–185 (2019).
6. Olkkonen, M. & Brainard, D. H. Joint effects of illumination geometry and object shape in the perception of surface reflectance. *i-Perception* **2**, 1014–1034 (2011).
7. Xiao, B. *et al.* Looking against the light: How perception of translucency depends on lighting direction. *J. Vis.* **14(3):**, 1–22 (2014).
8. Marlow, P., Kim, J. & Anderson, B. Coupled computations of 3d shape and translucency. *J. Vis.* **16**, 947–947 (2016).
9. Fleming, R. W. Material perception. *Annu. Rev. Vis. Sci.* **3**, 365–388 (2017).
10. Lagunas, M., Serrano, A., Gutierrez, D. & Masia, B. The joint role of geometry and illumination on material recognition. *J. Vis.* **21**, 2–2 (2021).
11. Sharan, L., Rosenholtz, R. & Adelson, E. Material perception: What can you see in a brief glance? *J. Vis.* **9**, 784–784 (2009).
12. Liao, C., Sawayama, M. & Xiao, B. Crystal or jelly? effect of color on the perception of translucent materials with photographs of real-world objects. *J. Vis.* **22**, 6–6 (2022).
13. Hanrahan, P. & Krueger, W. Reflection from layered surfaces due to subsurface scattering. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, 165–174 (1993).
14. Jensen, H. W., Marschner, S. R., Levoy, M. & Hanrahan, P. A practical model for subsurface light transport. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 511–518 (2001).
15. Beck, J. & Ivry, R. On the role of figural organization perceptual transparency. *Percept. & Psychophys.* **44**, 585–594 (1988).
16. Di Cicco, F., Wiersma, L., Wijntjes, M. & Pont, S. Material properties and image cues for convincing grapes: The know-how of the 17th-century pictorial recipe by willem beurs. *Art & Percept.* **8**, 337–362 (2020).

17. Gkioulekas, I. *et al.* Understanding the role of phase function in translucent appearance. *ACM Transactions on Graph. (TOG)* **32**, 1–19 (2013).

18. Chandrasekhar, S. *Radiative transfer* (Courier Corporation, 2013).

19. Gigilashvili, D. *et al.* The role of subsurface scattering in glossiness perception. *ACM Transactions on Appl. Percept. (TAP)* **18**, 1–26 (2021).

20. Marlow, P. J., Kim, J. & Anderson, B. L. Perception and misperception of surface opacity. *Proc. Natl. Acad. Sci.* **114**, 13840–13845 (2017).

21. Fleming, R. W. & Bülthoff, H. H. Low-level image cues in the perception of translucent materials. **2**, 346–382, DOI: 10.1145/1077399.1077409 (2005).

22. Chowdhury, N. S., Marlow, P. J. & Kim, J. Translucency and the perception of shape. *J. Vis.* **17**, 17–17 (2017).

23. Marlow, P. J. & Anderson, B. L. The cospecification of the shape and material properties of light permeable materials. *Proc. Natl. Acad. Sci.* **118**, e2024798118 (2021).

24. Marlow, P. J., Gegenfurtner, K. R. & Anderson, B. L. The role of color in the perception of three-dimensional shape. *Curr. Biol.* **32**, 1387–1394 (2022).

25. Gigilashvili, D., Urban, P., Thomas, J.-B., Hardeberg, J. Y. & Pedersen, M. Impact of shape on apparent translucency differences. In *Color and Imaging Conference*, vol. 2019, 132–137 (Society for Imaging Science and Technology, 2019).

26. Sawayama, M. *et al.* Visual discrimination of optical material properties: A large-scale study. *J. Vis.* **22**, 17–17 (2022).

27. Hebart, M. N. *et al.* Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS One* **14**, e0223792 (2019).

28. Che, C., Luan, F., Zhao, S., Bala, K. & Gkioulekas, I. Towards learning-based inverse subsurface scattering. In *2020 IEEE International Conference on Computational Photography (ICCP)*, 1–12 (IEEE, 2020).

29. Nishida, S. & Shinya, M. Use of image-based information in judgments of surface-reflectance properties. *JOSA A* **15**, 2951–2965 (1998).

30. Motoyoshi, I., Nishida, S., Sharan, L. & Adelson, E. H. Image statistics and the perception of surface qualities. *Nature* **447**, 206–209 (2007).

31. Doerschner, K. *et al.* Visual motion and the perception of surface material. *Curr. Biol.* **21**, 2010–2016 (2011).

32. Fleming, R. W. Visual perception of materials and their properties. *Vis. Res.* **94**, 62–75 (2014).

33. Brainard, D. H., Cottaris, N. P. & Radonjić, A. The perception of colour and material in naturalistic tasks. *Interface focus* **8**, 20180012 (2018).

34. Nishida, S. Image statistics for material perception. *Curr. Opin. Behav. Sci.* **30**, 94–99 (2019).

35. Fleming, R. W. & Storrs, K. R. Learning to see stuff. *Curr. Opin. Behav. Sci.* **30**, 100–108 (2019).

36. Anderson, B. L. Visual perception of materials and surfaces. *Curr. Biol.* **21**, R978–R983 (2011).

37. Gigilashvili, D., Thomas, J.-B., Hardeberg, J. Y. & Pedersen, M. Translucency perception: A review. *J. Vis.* **21(8):**, 1–41 (2021).

38. Fleming, R. W., Dror, R. O. & Adelson, E. H. Real-world illumination and the perception of surface reflectance properties. *J. Vis.* **3**, 3–3 (2003).

39. Kim, J., Marlow, P. & Anderson, B. L. The perception of gloss depends on highlight congruence with surface shading. *J. Vis.* **11**, 4–4 (2011).

40. Marlow, P., Kim, J. & Anderson, B. L. The role of brightness and orientation congruence in the perception of surface gloss. *J. Vis.* **11**, 16–16 (2011).

41. Marlow, P. J., Kim, J. & Anderson, B. L. The perception and misperception of specular surface reflectance. *Curr. Biol.* **22**, 1909–1913 (2012).

42. Kim, J., Marlow, P. J. & Anderson, B. L. The dark side of gloss. *Nat. Neurosci.* **15**, 1590–1595 (2012).

43. Nishio, A., Goda, N. & Komatsu, H. Neural selectivity and representation of gloss in the monkey inferior temporal cortex. *J. Neurosci.* **32**, 10780–10793 (2012).

44. Sun, H.-C. *et al.* Brain processing of gloss information with 2D and 3D depth cues. *J. Vis.* **15**, 818–818 (2015).

45. Toscani, M., Valsecchi, M. & Gegenfurtner, K. R. Lightness perception for matte and glossy complex shapes. *Vis. Res.* **131**, 82–95 (2017).

46. Miyakawa, N. *et al.* Representation of glossy material surface in ventral superior temporal sulcal area of common marmosets. *Front. Neural Circuits* **11**, 17 (2017).

47. Tsuda, H. & Saiki, J. Constancy of visual working memory of glossiness under real-world illuminations. *J. Vis.* **18**, 14–14 (2018).

48. Sawayama, M. & Nishida, S. Material and shape perception based on two types of intensity gradient information. *PLoS Comput. Biol.* **14**, e1006061 (2018).

49. Harvey, J. S. & Smithson, H. E. Low level visual features support robust material perception in the judgement of metallicity. *Sci. Reports* **11**, 1–15 (2021).

50. Cheeseman, J. R., Ferwerda, J. A., Maile, F. J. & Fleming, R. W. Scaling and discriminability of perceived gloss. *JOSA A* **38**, 203–210 (2021).

51. Schmid, A. C., Barla, P. & Doerschner, K. Material category of visual objects computed from specular image structure. *bioRxiv* 2019–12 (2021).

52. Ho, Y.-X., Landy, M. S. & Maloney, L. T. How direction of illumination affects visually perceived surface roughness. *J. Vis.* **6**, 8–8 (2006).

53. Pont, S. C. & Koenderink, J. J. Shape, surface roughness and human perception. In *Handbook of texture analysis*, 197–222 (World Scientific, 2008).

54. Kawabe, T., Maruya, K., Fleming, R. W. & Nishida, S. Seeing liquids from visual motion. *Vis. Res.* **109**, 125–138 (2015).

55. Paulun, V. C., Kawabe, T., Nishida, S. & Fleming, R. W. Seeing liquids from static snapshots. *Vis. Res.* **115**, 163–174 (2015).

56. van Assen, J. J. R., Barla, P. & Fleming, R. W. Visual features in the perception of liquids. *Curr. Biol.* **28**, 452–458 (2018).

57. van Assen, J. J. R., Nishida, S. & Fleming, R. W. Visual perception of liquids: Insights from deep neural networks. *PLoS Comput. Biol.* **16**, e1008018 (2020).

58. Paulun, V. C., Schmidt, F., van Assen, J. J. R. & Fleming, R. W. Shape, motion, and optical cues to stiffness of elastic objects. *J. Vis.* **17**, 20–20 (2017).

59. Schmidt, F., Paulun, V. C., van Assen, J. J. R. & Fleming, R. W. Inferring the stiffness of unfamiliar objects from optical, shape, and motion cues. *J. Vis.* **17**, 18–18 (2017).

60. Schmid, A. C. & Doerschner, K. Shatter and splatter: The contribution of mechanical and optical properties to the perception of soft and hard breaking materials. *J. Vis.* **18**, 14–14 (2018).

61. Alley, L. M., Schmid, A. C. & Doerschner, K. Expectations affect the perception of material properties. *J. Vis.* **20**, 1–1 (2020).

62. Bi, W., Jin, P., Nienborg, H. & Xiao, B. Manipulating patterns of dynamic deformation elicits the impression of cloth with varying stiffness. *J. Vis.* **19**, 18–18 (2019).

63. Sawayama, M., Adelson, E. H. & Nishida, S. Visual wetness perception based on image color statistics. *J. Vis.* **17**, 7–7 (2017).

64. Fleming, R. W., Jäkel, F. & Maloney, L. T. Visual perception of thick transparent materials. *Psychol. Sci.* **22**, 812–820 (2011).

65. Kawabe, T., Maruya, K. & Nishida, S. Perceptual transparency from image deformation. *Proc. Natl. Acad. Sci.* **112**, E4620–E4627 (2015).

66. Motoyoshi, I. Highlight–shading relationship as a cue for the perception of translucent and transparent materials. *J. Vis.* **10(9):**, 1–11 (2010).

67. Nagai, T. *et al.* Image regions contributing to perceptual translucency: A psychophysical reverse-correlation study. *i-Perception* **4**, 407–428 (2013).

68. Xiao, B., Zhao, S., Gkioulekas, I., Bi, W. & Bala, K. Effect of geometric sharpness on translucent material perception. *J. Vis.* **20(7):**, 1–17 (2020).

69. Todo, H., Yatagawa, T., Sawayama, M., Dobashi, Y. & Kakimoto, M. Image-based translucency transfer through correlation analysis over multi-scale spatial color distribution. *The Vis. Comput.* **35**, 811–822 (2019).

70. Chadwick, A. C., Cox, G., Smithson, H. E. & Kentridge, R. W. Beyond scattering and absorption: Perceptual unmixing of translucent liquids. *J. Vis.* **18**, 18–18 (2018).

71. Chadwick, A., Heywood, C., Smithson, H. & Kentridge, R. Translucence perception is not dependent on cortical areas critical for processing colour or texture. *Neuropsychologia* **128**, 209–214 (2019).

72. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis Mach. Intell.* **35**, 1798–1828 (2013).

73. Kriegeskorte, N. Deep neural networks: a new framework for modelling biological vision and brain information processing. *biorxiv* 029876 (2015).

74. O'Toole, A. J. & Castillo, C. D. Face recognition by humans and machines: Three fundamental advances from deep learning. *Annu. Rev. Vis. Sci.* **7**, 543–570 (2021).

75. Van Zuijlen, M. J., Lin, H., Bala, K., Pont, S. C. & Wijntjes, M. W. Materials In Paintings (MIP): An interdisciplinary dataset for perception, art history, and computer vision. *Plos One* **16**, e0255109 (2021).

76. Prokott, K. E., Tamura, H. & Fleming, R. W. Gloss perception: Searching for a deep neural network that behaves like humans. *J. Vis.* **21**, 14–14 (2021).

77. Tamura, H., Prokott, K. E. & Fleming, R. W. Distinguishing mirror from glass: A "big data" approach to material perception. *J. Vis.* **22**, 4–4 (2022).

78. Gulrajani, I. *et al.* Pixelvae: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013* (2016).

79. Storrs, K. R., Anderson, B. L. & Fleming, R. W. Unsupervised learning predicts human perception and misperception of gloss. *Nat. Hum. Behav.* **5**, 1402–1417 (2021).

80. Testolin, A., Stoianov, I. & Zorzi, M. Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nat. Hum. Behav.* **1**, 657–664 (2017).

81. Suchow, J. W., Peterson, J. C. & Griffiths, T. L. Learning a face space for experiments on human identity. *arXiv preprint arXiv:1805.07653* (2018).

82. Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J. & Kriegeskorte, N. Diverse deep neural networks all predict human it well, after training and fitting. *BioRxiv* (2020).

83. Kasahara, S., Ienaga, N., Shimizu, K., Takada, K. & Sugimoto, M. Human latent metrics: Perceptual and cognitive response corresponds to distance in gan latent space. (2022).

84. Zhuang, C. *et al.* Unsupervised neural network models of the ventral visual stream. *Proc. Natl. Acad. Sci.* **118**, e2014196118 (2021).

85. Higgins, I. *et al.* Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat. Commun.* **12**, 1–14 (2021).

86. Karras, T., Laine, S. & Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410 (2019).

87. Karras, T. *et al.* Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119 (2020).

88. Karras, T. *et al.* Training generative adversarial networks with limited data. *Adv. Neural Inf. Process. Syst.* **33**, 12104–12114 (2020).

89. Richardson, E. *et al.* Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2287–2296 (2021).

90. Goodfellow, I. *et al.* Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **27** (2014).

91. Radford, A., Metz, L. & Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

92. Yang, C., Shen, Y. & Zhou, B. Semantic hierarchy emerges in deep generative representations for scene synthesis. *Int. J. Comput. Vis.* **129**, 1451–1466 (2021).

93. Shen, Y., Yang, C., Tang, X. & Zhou, B. InterFaceGAN: Interpreting the disentangled face representation learned by GANs. *IEEE Transactions on Pattern Analysis Mach. Intell.* (2020).

94. Abdal, R., Qin, Y. & Wonka, P. Image2StyleGAN: How to embed images into the StyleGAN latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4432–4441 (2019).

95. Wu, Z., Lischinski, D. & Shechtman, E. StyleSpace analysis: Disentangled controls for StyleGAN image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12863–12872 (2021).

96. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O. & Cohen-Or, D. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graph. (TOG)* **40**, 1–14 (2021).

97. Zhou, S. *et al.* Hype: A benchmark for human eye perceptual evaluation of generative models. *Adv. Neural Inf. Process. Syst.* **32** (2019).

98. Kruschke, J. K. Rejecting or accepting parameter values in bayesian estimation. *Adv. Methods Pract. Psychol. Sci.* **1**, 270–280 (2018).

99. Abdal, R., Zhu, P., Mitra, N. J. & Wonka, P. Labels4Free: Unsupervised segmentation using StyleGAN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13970–13979 (2021).

100. Barlow, H. B. *et al.* Possible principles underlying the transformation of sensory messages. *Sens. Commun.* **1** (1961).

101. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).

102. Simoncelli, E. P., Olshausen, B. *et al.* Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).

103. Smith, E. C. & Lewicki, M. S. Efficient auditory coding. *Nature* **439**, 978–982 (2006).

104. Hyvärinen, A. & Oja, E. Independent component analysis: algorithms and applications. *Neural Networks* **13**, 411–430 (2000).

105. Wijntjes, M., Spoiala, C. & De Ridder, H. Thurstonian scaling and the perception of painterly translucency. *Art & Percept.* **8**, 363–386 (2020).

106. Giesel, M. & Zaidi, Q. Frequency-based heuristics for material perception. *J. Vis.* **13**, 7–7 (2013).

107. Sawayama, M. & Kimura, E. Stain on texture: Perception of a dark spot having a blurred edge on textured backgrounds. *Vis. Res.* **109**, 209–220 (2015).

108. Cheeseman, J. R., Fleming, R. W. & Schmidt, F. Scale ambiguities in material recognition. *iScience* **25**, 103970 (2022).

109. Gkioulekas, I., Walter, B., Adelson, E. H., Bala, K. & Zickler, T. On the appearance of translucent edges. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5528–5536 (2015).

110. Gigilashvili, D., Dubouchet, L., Hardeberg, J. Y. & Pedersen, M. Caustics and translucency perception. *Electron. Imaging* **2020**, 33–1 (2020).

111. Olkkonen, M., Hansen, T. & Gegenfurtner, K. R. Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. *J. Vis.* **8**, 13–13 (2008).

112. Yoonessi, A. & Zaidi, Q. The role of color in recognizing material changes. *Ophthalmic Physiol. Opt.* **30**, 626–631 (2010).

113. Flachot, A. & Gegenfurtner, K. R. Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks. *Vis. Res.* **182**, 89–100 (2021).

114. Conway, B. R. The organization and operation of inferior temporal cortex. *Annu. Rev. Vis. Sci.* **4**, 381 (2018).

115. Glorot, X., Bordes, A. & Bengio, Y. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323 (JMLR Workshop and Conference Proceedings, 2011).

116. LeCun, Y., Bengio, Y. *et al.* Convolutional networks for images, speech, and time series. *The Handb. Brain Theory Neural Networks* **3361**, 1995 (1995).

117. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning* (MIT press, 2016).

118. Geirhos, R. *et al.* Generalisation in humans and deep neural networks. *Adv. Neural Inf. Process. Syst.* **31** (2018).

119. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125 (2017).

120. Metzger, A. & Toscani, M. Unsupervised learning of haptic material properties. *Elife* **11**, e64876 (2022).

121. Peirce, J. W. PsychoPy—psychophysics software in Python. *J. Neurosci. Methods* **162**, 8–13 (2007).

122. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

123. Bürkner, P.-C. brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* **80**, 1–28 (2017).

Acknowledgements

We thank Eric Schuler for the valuable discussion on the statistical analysis of this work and Alex Godwin for discussion of data visualization.

Supplementary Information

Experiment 1: Real-versus-generated discrimination

To further illustrate the results of Figure.2(d) in main paper, Figure S. 1 shows examples of real-versus-generated judgment agreed by the majority of observers (at least 50% of observers).

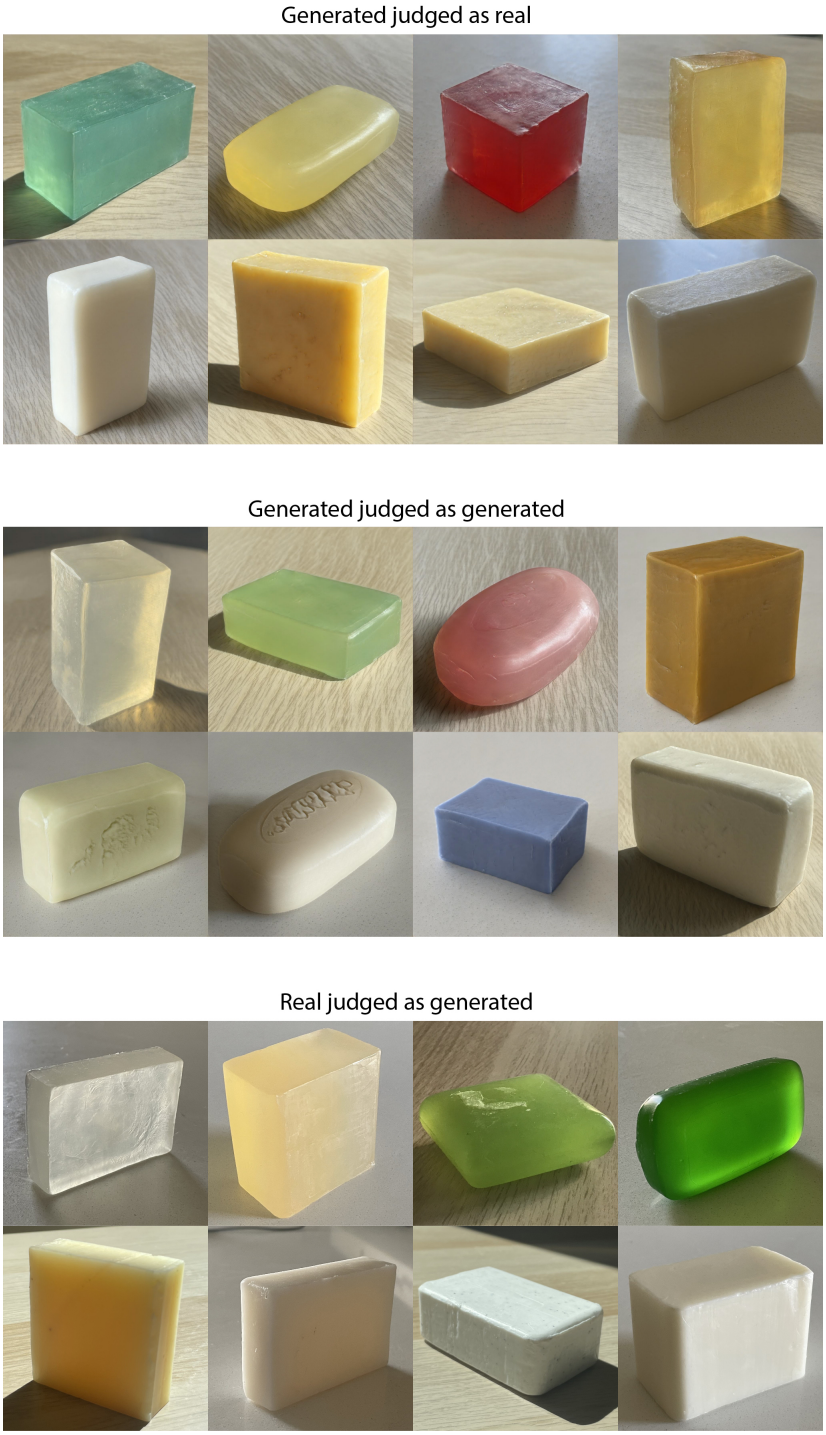

Figure S. 1. Example stimuli from the real-versus-fake experiment agreed by the majority of observers. Each image is resized for display.

Experiment 2: Material attribute rating

Figure S. 2 shows the illustrations of light transport process for opaque and translucent objects that we presented to the observers in the experiment instruction.

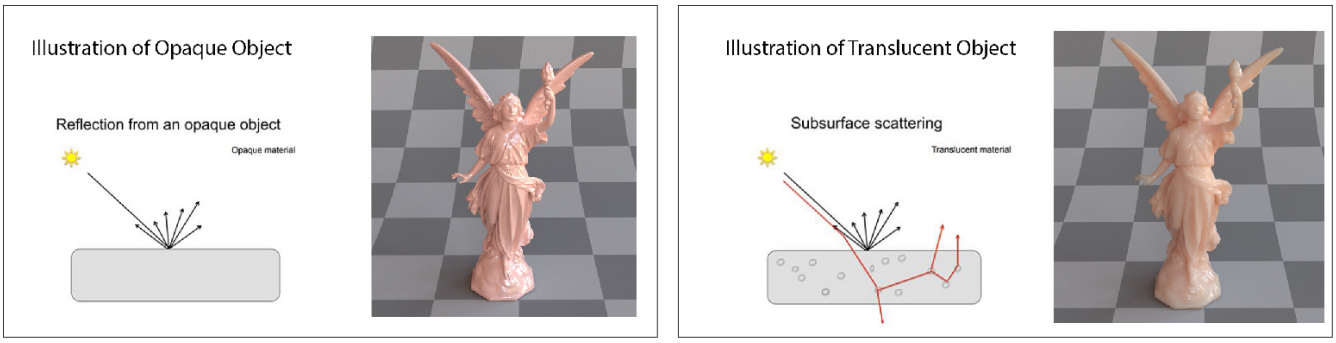

Figure S. 2. Illustrations of the simplified light transport process. Left: light transport process for an opaque object. Right: subsurface scattering for a translucent object.

Experiment 3: Scene attribute evaluation

Table. 1 summarizes the results from the Bayesian multilevel multinomial logistic regression model. We set the baselines of the variables in the following way: "lighting" for the scene attribute, "later-layer manipulation" for the layer-manipulation method, and "opaque-opaque (OO)" for the type of source-target pair.

The most salient output from the model is the mean of posterior distribution for early-layer manipulation. The relative risk ratio of selecting "shape" in comparison to selecting "lighting" when the layer-manipulation method switches from later-layer manipulation to early-layer manipulation has an estimated mean posterior of 2644.90 (95% Highest Density Interval (HDI), [1106.94, 6783.08]). We could make the decision of whether to accept or reject the null value of a parameter based on the relation between HDI and region of practical equivalence (ROPE). ROPE is set for each parameter with a 0.1 range around 0 using the unexponentiated model. If none of the 95% HDI of the parameter distribution falls into the ROPE, we can reject the null value. If the 95% HDI of the parameter distribution completely falls into the ROPE, we accept the null as a credible value. Otherwise, the decision remains undecided. Since zero percent of the 95% HDI of the parameter distribution for early-layer manipulation falls inside the ROPE, we can reject the null value. Therefore, early-layer manipulation is a credible parameter and it increases the probability that the observers choose "shape/orientation" as the most prominent change in the image sequence, regardless of the source-target pair of the materials. Secondly, the relative risk ratio of selecting "material" in comparison to selecting "lighting" when the layer-manipulation method switches from later-layer manipulation to middle-layer manipulation has an estimated mean posterior of 2.11 (95% HDI, [1.58, 2.80]). Middle-layer manipulation is also a credible parameter (Inside.ROPE = 0), and it increases the probability for the observers to select "material" as the most prominent visual attribute changed as compared to selecting "lighting". It is also important to note that the middle-layer manipulation applied on opaque-translucent pairs increases the probability that observers choose "material" as the most prominent visual attribute being changed (Mean Est = 11.61, 95% HDI, [7.34, 18.39], Inside.ROPE = 0). Lastly, later-layer manipulation is most likely to lead to change in "color" across all source-target pair conditions.

Figure S. 3 illustrates the conditional effect of layer-manipulation on the prediction of the most prominent scene attribute chosen by the observers. For the early-layer manipulation, the estimated probability of selecting "shape" is close to 1 across all three types of source-target pair. For the later-layer manipulation, the estimated probability of selecting "color" is approximately 77.4% (95% HDI, [71.4%, 82.5%]) for OT pairs, 68.7% (95% HDI, [61.4%, 75.1%]) for OO pairs, and 80.0% (95% HDI, [73.9%, 85.0%]) for TT pairs. For the middle-layer manipulation, the estimated probability of selecting "material" is 77.9% (95% HDI, [69.5%, 84.5%]) for the OT pair, which is higher than that of the OO or TT condition.

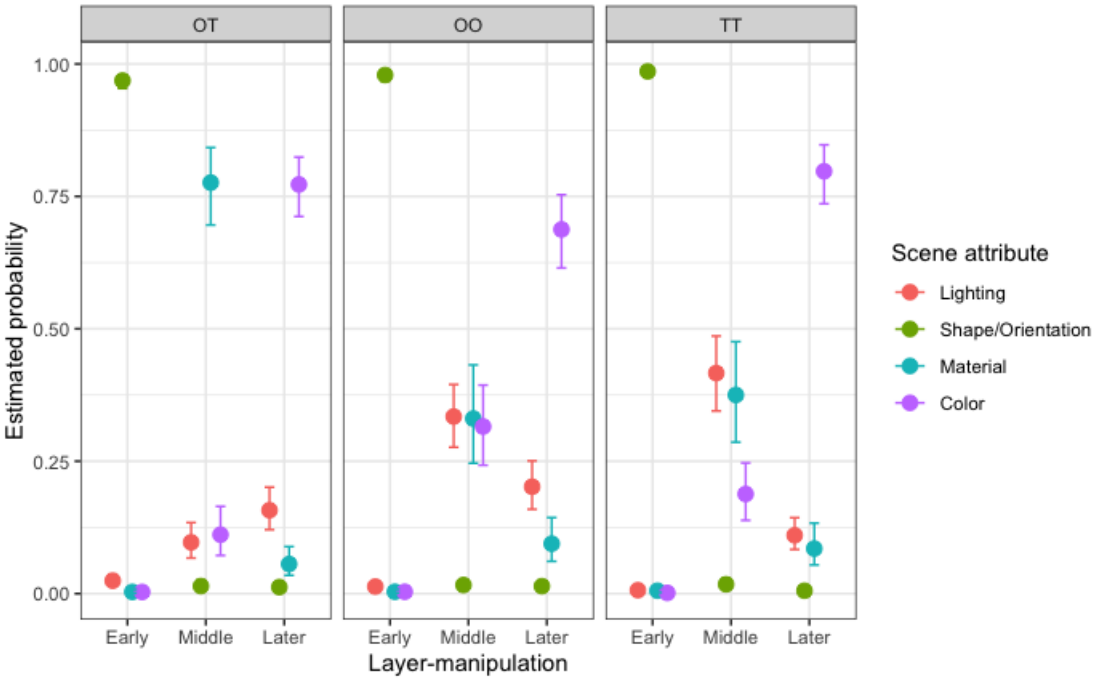

Figure S. 3. Conditional effect of layer-manipulation method on the prediction of scene attribute using Bayesian multilevel multinomial logistic regression model. The x-axis is the layer-manipulation method, and the y-axis is the estimated probability that certain scene attribute is selected as the most prominent attribute that has been changed in an image sequence. The error bar indicates the upper and lower bounds of the estimation at confidence level of 95%. The panels show the predicted results for three source-target pair conditions: opaque-translucent (OT), opaque-opaque (OO), and translucent-translucent (TT).

Response_Predictor	Mean Est	Est.Error	HDI Lower	HDI Upper	Inside.ROPE (%)
Shape_Early-layers	**2644.90**	**1.59**	**1106.94**	**6783.08**	**0.00**
Shape_Middle-layers	0.94	1.50	0.43	2.11	20.25
Shape_Source-target pair(TT)	0.79	1.77	0.24	2.32	13.57
Shape_Source-target pair(OT)	1.63	1.52	0.72	3.69	10.36
Shape_Early-layers and Source-target pair(TT) interaction	1.86	2.14	0.44	8.82	8.10
Shape_Middle-layers and Source-target pair(TT) interaction	1.24	1.96	0.34	4.85	12.80
Shape_Early-layers and Source-target pair(OT) interaction	0.19	1.76	0.06	0.58	0.00
Shape_Middle-layers and Source-target pair(OT) interaction	2.26	1.75	0.77	6.86	5.14
Material_Early-layers	0.72	1.90	0.19	2.35	11.98
Material_Middle-layers	**2.11**	**1.16**	**1.58**	**2.80**	**0.00**
Material_Source-target pair(TT)	1.70	1.20	1.18	2.43	0.00
Material_Source-target pair(OT)	0.73	1.22	0.50	1.07	12.38
Material_Early-layers and Source-target pair(TT) interaction	1.83	2.32	0.38	10.12	8.81
Material_Middle-layers and :Source-target pair(TT) interaction	0.54	1.24	0.35	0.82	0.00
Material_Early-layers and Source-target pair(OT) interaction	0.44	2.37	0.08	2.41	5.99
Material_Middle-layers and Source-target pair(OT) interaction	**11.61**	**1.27**	**7.34**	**18.39**	**0.00**
Color_Early-layers	**0.11**	**1.89**	**0.03**	**0.34**	**0.00**
Color_Middle-layers	**0.28**	**1.12**	**0.22**	**0.35**	**0.00**
Color_Source-target pair(TT)	2.22	1.14	1.73	2.85	0.00
Color_Source-target pair(OT)	1.47	1.12	1.16	1.85	0.00
Color_Early-layers and Source-target pair(TT) interaction	0.10	4.36	0.00	1.29	1.58
Color_Middle-layers and Source-target pair(TT) interaction	0.21	1.19	0.15	0.30	0.00
Color_Early-layers and Source-target pair(OT) interaction	0.15	2.49	0.02	0.89	0.07
Color_Middle-layers and Source-target pair(OT) interaction	0.86	1.22	0.59	1.27	32.32
Shape_Intercept	0.05	1.41	0.02	0.09	0.00
Material_Intercept	0.46	1.26	0.30	0.72	0.00
Color_Intercept	3.33	1.17	2.45	4.58	0.00

Table 1. Summary of Bayesian multilevel multinomial logistic regression model outputs. The left most column shows the name of the parameter. The names of the response variable and the predictor are separated by "_". The second to fifth columns are the exponentiated mean (Mean Est), the standard error (Est.Error), and the lower (HDI Lower) and upper bounds (HDI Upper) of the 95% credible interval of the posterior distribution for each parameter. The last column is the percentage of the 95% HDI of parameter distribution falls inside the ROPE.

Visualizing the effect of layer manipulation based on learned decision boundary of milky-versus-glycerin soap classification

As an extension of Figure.5(d) in the main paper, we show the image manipulation results based on the learned SVM decision boundary for each of 18 layers of $W+$ latent space. Figure S. 4 and Figure S. 5 demonstrate the manipulation along the positive and negative direction of the normal of the learned decision boundary.

Independent Component Analysis (ICA) for the intermediate generative representation

We used FastICA from *scikit-learn* for the independent component analysis of image patches extracted from the intermediate results of the generative process our trained generator. For the intermediate generative result from 64×64 tRGB layer, we conducted the analysis using 64 and 100 components, and found that similar sets of sparse features (i.e. middle-layer ICA kernels) were extracted. More details can be found in the Result section of the main paper.

We also conducted a similar control analysis based on the intermediate generative result from the early-layers. For the 3160 "high" translucent images used to obtain the middle-layer ICA kernels, we extracted their corresponding intermediate generative results from 16×16 tRGB layer (Figure S. 7(a)). Since a great number of the 16×16 generative results were generated from the same early-layer latent vectors, we randomly sampled 1000 of them to reduce the redundancy. For each image, we first resized it to 512 pixels $\times$ 512 pixels resolution, and then sampled 10 image patches of 96 pixels $\times$ 96 pixels from random locations. We then applied FastICA on the 1000×10 image patches to learn 64 basis functions (i.e. early-layer ICA kernels) (Figure S. 7(b)). The effect of convolving these 96×96 kernels with a real photograph of translucent soap is shown in Figure S. 7(c). Compared with the information extracted from middle-layer kernels, the early-layer chromatic kernels detect coarser information of the edges, and cannot capture the fine spatial color variations that are indicative of the "glowing" effect of translucent materials.

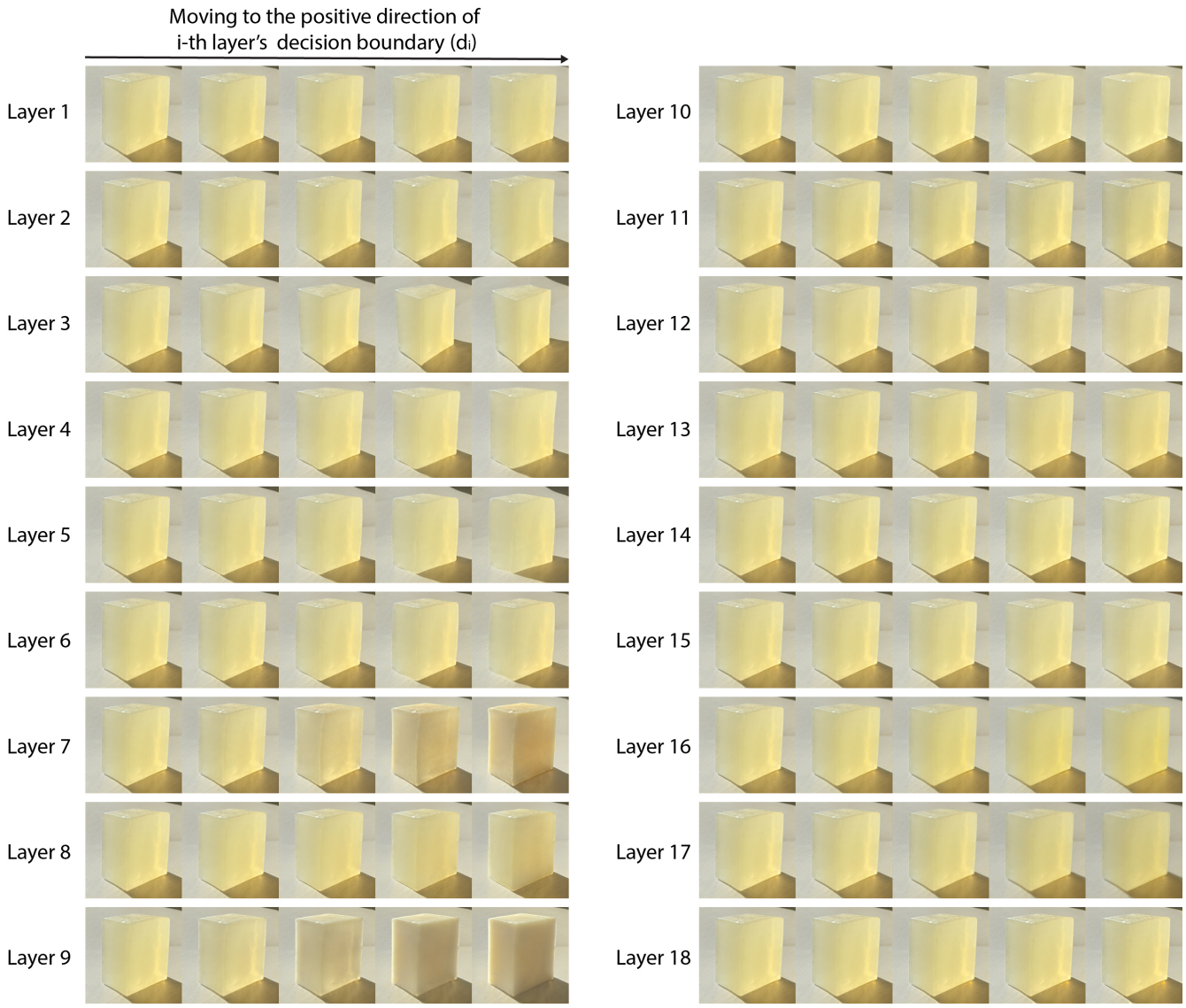

Figure S. 4. Manipulating the i-th layer's latent vector of the original image (left most) along the positive direction of the normal of the learned translucent decision boundary (d_i). The displacement on the middle-layers (layers 7 to 9) can mainly affect translucent appearance.

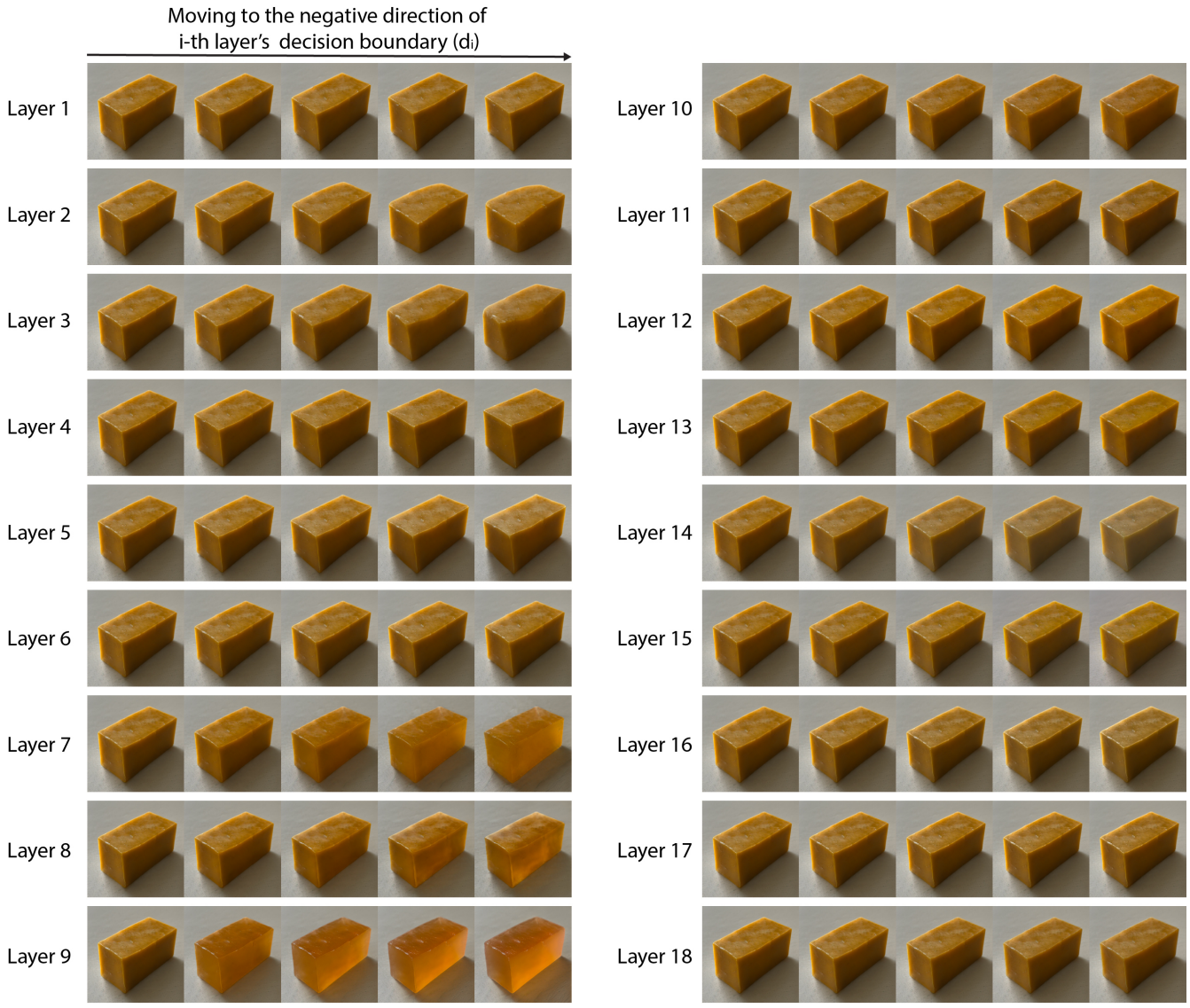

Figure S. 5. Manipulating the i-th layer's latent vector of the original image (left most) along the negative direction of the normal of the learned translucent decision boundary (d_i). The displacement on the middle-layers (layers 7 to 9) can mainly affect translucent appearance.

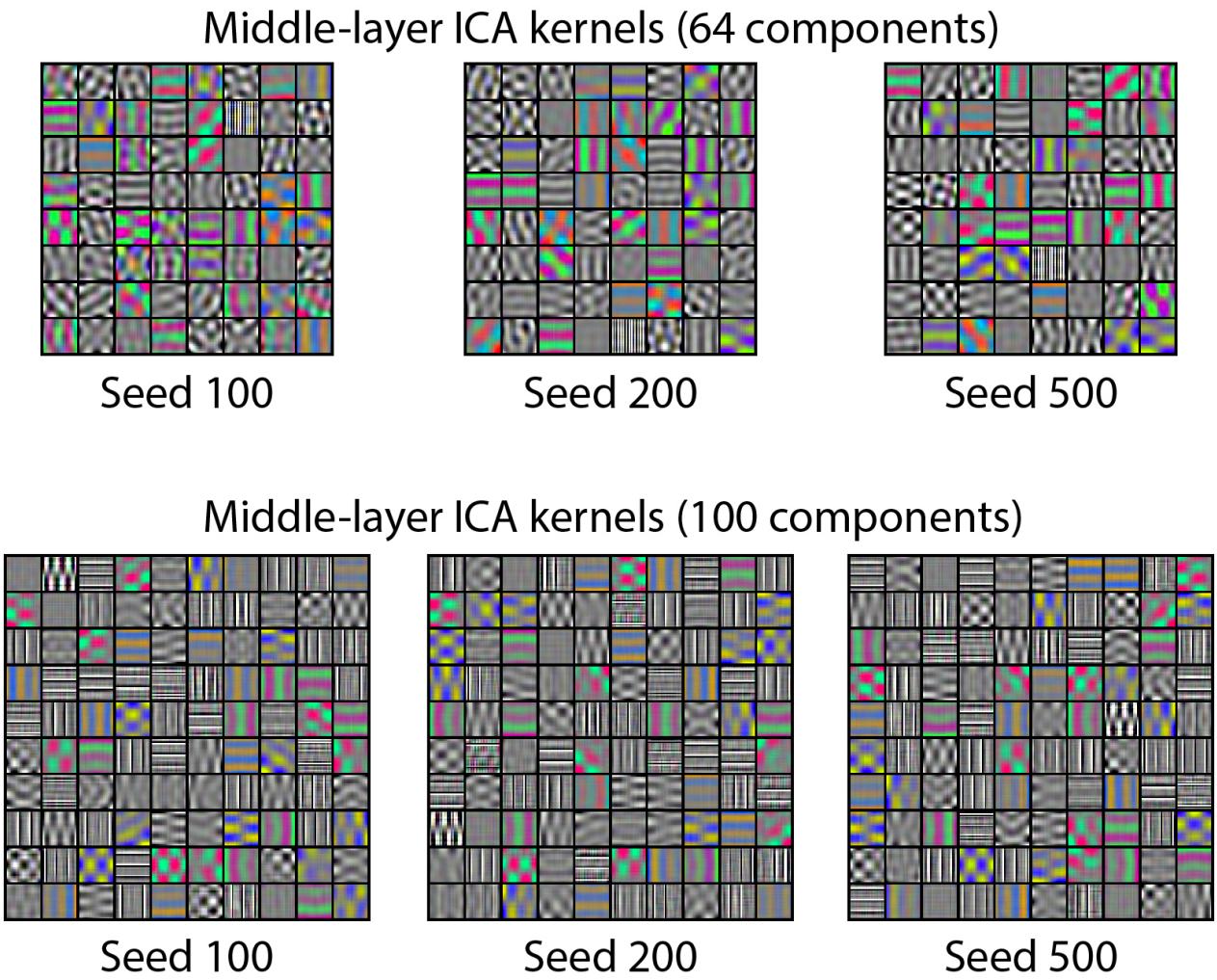

Figure S. 6. Middle-layer ICA kernels extracted from intermediate activation of high-translucency images (see main paper Figure 7). Top and bottom rows show the FastICA results of using 64 and 100 components respectively. Within each row, each panel shows the kernels learned from a different random sampling of the image patches. The kernels are 24×24, and are resized for display.

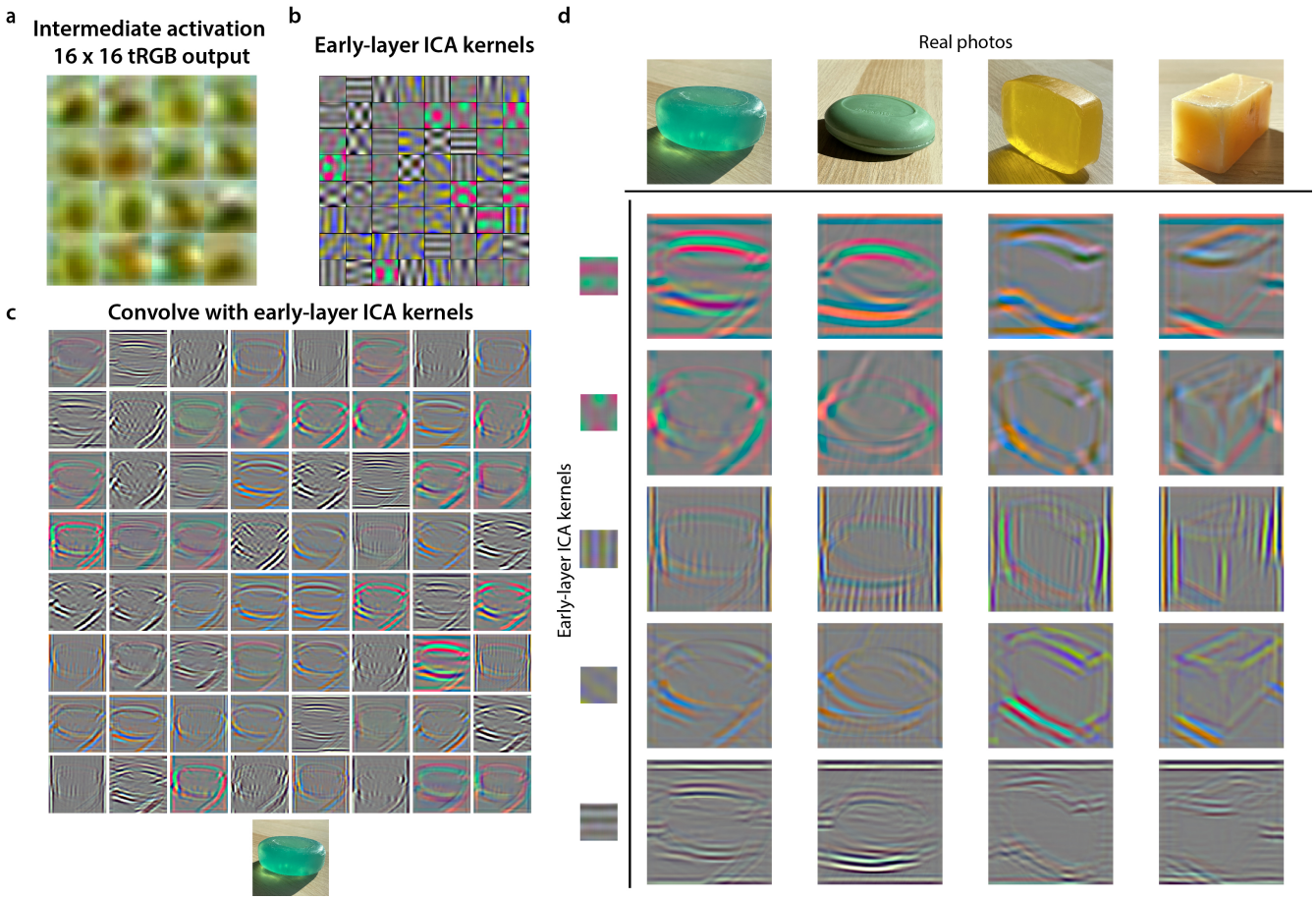

Figure S. 7. Visualization of features captured in the early-layers of the learned latent space. a, The intermediate generated results (tRGB layer output at 16 pixels $\times$ 16 pixels resolution) of the images from the high-translucency dataset. The images are resized for display. **b,** Early-layer ICA kernels obtained by training a system of 64 basis functions on 96 $\times$ 96 image patches extracted from images in (a). The kernels are of size 96 $\times$ 96. **c,** Visualization of applying three-dimensional convolution of the individual early-layer ICA kernels in (b) on a real photograph of translucent soap. **d,** The resulting filtered images of four different soaps with selected chromatic and achromatic kernels.