The role of texture summary statistics in material recognition from drawings and photographs

Benjamin Balas

Psychology Department,
North Dakota State University,
Fargo, ND, USA

Michelle R. Greene

Psychology Department, Barnard College,
Columbia University, New York, NY, USA

Material depictions in artwork are useful tools for revealing image features that support material categorization. For example, artistic recipes for drawing specific materials make explicit the critical information leading to recognizable material properties (Di Cicco, Wjintjes, & Pont, 2020) and investigating the recognizability of material renderings as a function of their visual features supports conclusions about the vocabulary of material perception. Here, we examined how the recognition of materials from photographs and drawings was affected by the application of the Portilla–Simoncelli texture synthesis model. This manipulation allowed us to examine how categorization may be affected differently across materials and image formats when only summary statistic information about appearance was retained. Further, we compared human performance to the categorization accuracy obtained from a pretrained deep convolutional neural network to determine if observers' performance was reflected in the network. Although we found some similarities between human and network performance for photographic images, the results obtained from drawings differed substantially. Our results demonstrate that texture statistics play a variable role in material categorization across rendering formats and material categories and that the human perception of material drawings is not effectively captured by deep convolutional neural networks trained for object recognition.

Introduction

The depiction of materials in artwork is a useful means to examine critical features for material perception. Painting and drawing materials necessarily entail feature selection on the part of the artist: What contours and gradients must be included to successfully communicate what an object or surface is made of? What spatial layout of textures, contours, and colors will most effectively convey a specific material property? Beside these questions regarding the presence or absence of specific features, there are also decisions to be made with regard to the techniques used to create specific features in paintings and drawings. In general, the choices that lead to successful material depiction thus highlight what the visual system needs to be presented with in which positions for materials to be perceived correctly. For example, the recipes used by artists to depict objects with complex material properties like grapes or lemons (Di Cicco, Wjintjes, & Pont, 2019; Di Cicco, Wjintjes, & Pont, 2020) offer insights into the micropatterns that support inferences of qualities like glossiness and wetness. In the absence of explicit instructions describing how to paint or draw specific materials, examining how materials are depicted across many works of art is a useful way to extract these same insights, especially when coupled with observer evaluations of material properties (Van Zuijlen, Pont, & Wjintjes, 2020).

Drawings are a particularly interesting vehicle for studying the depiction of materials in artwork. Compared with paintings, drawings are often limited to either grayscale values or a two-tone black and white palette, limiting the tools available to the artist to render materials. Also, although the perception of line drawings has been previously explored in the context of shape recovery (Sayim & Cavanagh, 2011; Hertzmann, 2021) to our knowledge, there is as yet far less work examining how drawings may successfully signal material properties. Hertzmann (2020), for example, makes a compelling argument that line drawings (which lack the use of techniques like shading, hatching, and in-painting of tone) work largely by virtue of the artist choosing to include features that capture shape effectively and the visual system processing drawings in the same manner as realistic images.

Citation: Balas, B., & Greene, M. R. (2023). The role of texture summary statistics in material recognition from drawings and photographs. *Journal of Vision*, *23*(14):3, 1–13, https://doi.org/10.1167/jov.23.14.3.

https://doi.org/10.1167/jov.23.14.3 Received March 28, 2023; published December 8, 2023 ISSN 1534-7362

Material properties are not solely communicated by shape, however (Motoyoshi, Nishida, Sharan, & Adelson, 2007; Baumgartner & Gegenfurtner, 2016; Balas & Schmidt, 2017), and in some cases material properties are conveyed graphically both by realistic rendering of natural images and by iconic features that signal material properties by abstraction, similar to the techniques used in comics to communicate visual information without adhering to realism (Cohn & Ehly, 2016). Are drawings of materials then perceived by the human visual system in the same way as realistic images of materials? Alternatively, could it be the case that material drawings rely on specific aspects of image structure more than realistic images do, and vice versa? The answer to this question may also not be uniform across material categories. The perception of different classes of textures and different material categories relies to varying extents on distinct feature classes (Kung & Richards, 1988; Balas & Schmidt, 2017), so the critical features that support drawings and realistic images of various materials may also be a function of the specific categories or qualities under investigation.

We chose to examine one particular aspect of how realistic images depicting materials may be perceived differently than drawings of the same: How does the perception of these images rely on summary statistics as opposed to the joint measurement of position and appearance? Put more simply, does either type of image rely more heavily on seeing specific features in specific positions or configuration as opposed to a texture-like description of appearance that does not encode feature location precisely? We investigated this question by using texture synthesis models as a tool for generating modified versions of original images that were matched for a broad set of summary statistics, but differed from the original stimuli in terms of the spatial layout of features. Specifically, we used the Portilla–Simoncelli algorithm (Portilla & Simoncelli, 2000) for this purpose, a texture synthesis model that has been successfully used to examine summary statistic perception across a wide range of tasks including texture perception (Balas, 2006), visual crowding (Balas, Nakano, & Rosenholtz, 2009), visual search (Rosenholtz, Huang, Raj, Balas, & Ilie, 2012), and the neural processing of natural textures (Balas & Conlin, 2015). By rendering new images using parent images of natural materials and drawings of materials, we are able to create stimuli that lack the joint relationships between position and local feature appearance, but preserve many of the statistical properties of the original image. Comparing the perception of the original images to synthetic images thus permits inferences about the importance of position information to successful material categorization (Balas & Schmidt, 2017; Balas, Auen, Thrash, & Lammers, 2020).

We applied this model to realistic images of different materials and drawings of the same materials obtained from Japanese manga to determine whether either class of stimuli relies on positional information to a greater extent. We hypothesized that realistic images would be sufficiently rich in both texture-like features and position-dependent features signaling material properties to be robust to texture synthesis, but that line drawings would rely more heavily on position-dependent features and thus be harder to recognize from our synthetic images, which do not preserve location information from the parent image. The latter one-half of this hypothesis was based on the observation that in material drawings, material properties are frequently conveyed either by exaggerated or abstract depictions of image features like specularities that are positioned at specific locations on objects or surfaces (the stellate glint of a sharp sword, for example) or by small patches of hatching or shading at crucial positions, implying that the remaining empty surface is filled with the same texture (e.g., an artist drawing only scattered batches of bricks in a brick wall). However, we also hypothesized that this effect may vary across different material categories: Some materials may be especially reliant on position-dependent features, while others may largely depend on the depiction of distributed texture-like features. Specifically, we hypothesized that glossy or shiny materials like water or metal may depend heavily on the position of highlights relative to contours that signal local surface geometry (see Anderson & Kim [2009] for a demonstration that human vision is sensitive to these relationships), whereas matte materials like stone and wood may be less dependent on these conjunctions. Whatever pattern of results we may observe across our different stimulus classes and material categories, we also wished to know the extent to which these effects may reflect image-level properties of our stimuli versus higher level aspects of visual recognition. To examine this issue, we complemented our psychophysical results with a DNN analysis of the discriminability of our stimuli across all stimulus categories.

Briefly, we found that the impact of texture synthesis on material categorization varied across materials and also depended on the stimulus type (drawing vs. realistic image). More important, this pattern of results was not predicted by our deep convolutional neural network (dCNN) analysis, suggesting that these effects are not a simple reflection of the low-level structure in our images, but depend on higher-order aspects of material perception in the human visual system that are not captured by this dCNN. We discuss these results in terms of emergent theories regarding the relationship between drawing and visual perception and the potential for a unified versus material-specific model of human material perception and categorization.

Methods

Stimuli

We used two different sets of stimulus images in our task. Drawings depicting metal, stone, water and wood were selected from the Manga Materials Database (Saito, Kirai, & Horiuchi, 2015). We selected these categories for two reasons. First, in prior work both with adults (Balas, Conlin, & Shipman, 2017) and children (Balas & Schmidt, 2017), these categories were used to include variation in material properties across categories, including a matte versus a glossy appearance. Our use of these categories in the current study helps to align our present results with those previous reports. Second, these four categories were both among those with the greatest number of unique patches in the Manga Material Database and could also be found in the database of photographic material images we chose to use (see below). Individual images were 512 × 512 pixels and depicted two-tone (black and white only) drawings of objects and surfaces matching each of our four target categories. An important note is that these are not line drawings owing to the presence of cross-hatching, in-painted areas (regions filled with uniform black ink), and textured linework. Photographs depicting the same material categories were selected from the Flickr Materials Database (Sharan, Rosenholtz, & Adelson, 2014) and cropped to a square aspect ratio. To convert the original full-color photographs into two-tone images, we applied the function imbinarize.m in MATLAB, which uses Otsu's method (Otsu, 1979) to select a threshold to convert greyscale intensities into a two-tone image (Figure 1). Briefly, this method identifies the threshold between light and dark pixels that minimizes the intra-class variance, where pixel class is defined by pixel intensity being either above or below the candidate threshold value.

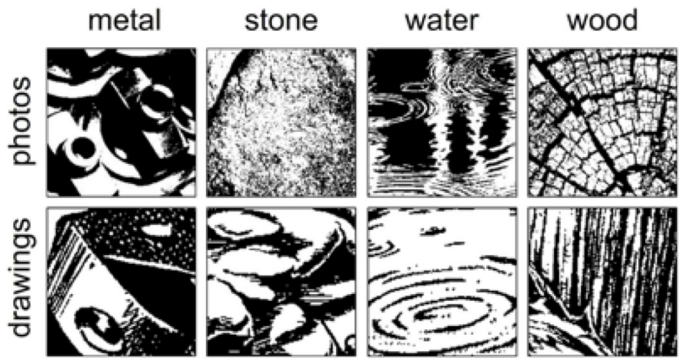

Figure 1. Examples of images from each material category used in our experiments. The top row contains photographic images of each material, and the bottom row contains drawings.

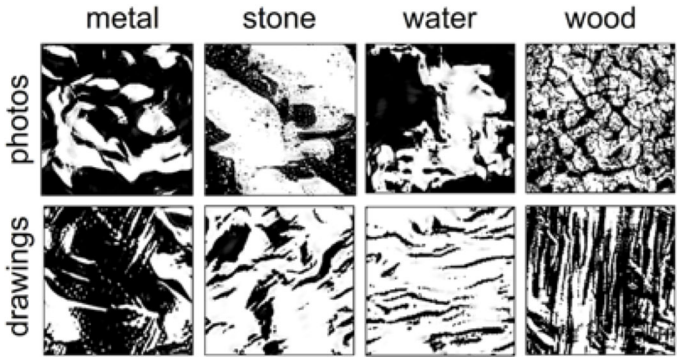

Figure 2. Examples of synthetic texture images from each material category used in our experiments. The top row contains syntheses made from photographic images of each material, and the bottom row contains synthetic textures made from drawings.

Participants

We recruited a total of 53 participants from the North Dakota State University undergraduate psychology study pool to complete this experiment. A total of 35 participants completed the task using the original stimulus images described above, and a total of 17 participants completed the task using images created via the application of the Portilla–Simoncelli texture synthesis algorithm (Portilla & Simoncelli, 2000) to these original images. Specifically, we applied the texture analysis and synthesis routines to each of the two-tone drawings and photographs using the default parameters in the MATLAB implementation of the algorithm. This yielded a new set of two-tone images (Figure 2), one per original image, with joint wavelet statistics that closely match the parent image. The appearance of these images is such that, although local structures tend to be preserved, the global layout of the original image, including large-scale pictorial elements, is disrupted.

All participants were between the ages of 18 and 27 years of age and self-reported normal or corrected-to-normal visual acuity. Participants received course credit for completing the experiment and only began the testing session after providing written informed consent. All recruitment and testing procedures were approved by the North Dakota State University IRB (Protocol #SM11167). The sample sizes of our two participant groups are different owing to restrictions on human subjects testing arising from the coronavirus disease 2019 pandemic.

Procedure

We asked participants in each group to complete a four-alternative forced choice material categorization task using both the drawings and photographs

described above. Participants in our first group were only shown the original two-tone photographs and drawings from each of our target material categories, whereas participants in the second group were only shown the images resulting from the application of the Portilla–Simoncelli algorithm. For each trial, participants were presented with a single image and asked to categorize it as either metal, stone, water, or wood. Image presentation time was unlimited and participants were free to take as much time as they liked to respond. We presented drawings and photographs in separate blocks (2 blocks per stimulus type) and randomized the order of these blocks for each participant. Each block contained all of the stimulus images of that type (192 images per block) for a grand total of 384 trials in the entire testing session. Within each block, stimulus order was randomized independently for each participant.

All stimulus presentation and response collection routines were written in PsychoPy v3.0 and administered as an online experiment via the Pavlovia platform. Because participants completed the task online, we cannot comment on precise characteristics of individual observers' displays, but we did limit the availability of the experiment so that smartphones and tablets could not be used to complete the task.

dCNN modeling

We extracted layerwise activations for each image from a dCNN (Krizhevsky, Sutskever, & Hinton, 2017) that was pretrained on the ImageNet database (Deng et al., 2009). This network contains eight layers, with the first five being convolutional and the last three fully connected. For the convolutional layers, layer activations were obtained after max pooling. For all layers, activation patterns were vectorized to create a 192-image by M-feature matrix. The number of features in each layer varied from 4,096 in layers 6 and 7 to 64,896 in layer 3. Our goal was to use the activation patterns in each layer of the network (excluding the final classification layer) to investigate the discriminability of material categories using both the original photographs and drawings and the synthetic versions of these stimulus sets. To achieve this, we carried out two stages of dimensionality reduction. First, we applied principal components analysis to the activation patterns in each layer, decreasing the full activation pattern to a 192-dimensional vector, which corresponds with the number of unique stimuli in each image condition. This ensures that at this first stage of dimensionality reduction we are retaining the full variance across inputs. Subsequently, we applied t-distributed stochastic neighbor embedding (t-SNE) (Hinton & Roweis, 2002) to reduce this 192-dimensional description of the images to a 2-dimensional description. The latter supports straightforward visualization of the data and also simplifies the classification step, which we will describe in the Results section (see also Figure 7).

Because t-SNE is a stochastic procedure that results in different low-dimensional embeddings, each time it is executed, we performed a linear classification analysis on 100 independently generated t-SNE solutions at each of the eight dCNN layers. We evaluated the amount of category information present in each layer via a linear support vector machine, using leave-one-observation-out) cross-validation. This analysis was conducted independently for original photographs, original drawings, and texturized photographs and drawings. We averaged across the 100 t-SNE solutions to obtain an average accuracy for each of the eight dCNN layers.

To compare the performance of the dCNN-based classifier to human observers, we computed the accuracy ratio of the classifier to humans as:

$$\hat{r} = \frac{k_c/n_c}{k_h/n_h},$$

where n_c and n_h denote the number of total trials for the classifier and human observers, and k_c and k_h denote the number of those trials that are correctly-classified, respectively.

Additionally, we examined the agreement between the dCNN model and human observers by calculating the expected number of human-classifier agreements as suggested by Tadros, Cullen, Greene, and Cooper (2019). Briefly, this method was inspired by measures of inter-rater reliability, such as Cohen's Kappa, and accounts for the fact that two independent classifiers that are at either floor or ceiling performance will have very high agreement by definition. This is especially helpful when considering all layers of a dCNN, as classification performance increases with increasing layers. Following Tadros et al. (2019), we computed the expected number of agreements as:

$$E[a] = n_h\left(\left(\frac{k_c}{n_c}\right)\left(\frac{k_h}{n_h}\right) + \left(\frac{w_c}{n_c}\right)\left(\frac{w_h}{n_h}\right)\right),$$

where w_c and w_h represent the number of incorrect responses from the classifier and human observers, respectively. From this, we can compute the difference between observed agreement (a/n_h) and expectation ($E[a]/n_h$), and assess whether this deviation was significantly different from chance.

Finally, we used the level of agreement between human observers to compute an upper bound for the agreement that can be expected between human observers and the classifier. If humans misclassify an image in a consistent manner, then it is reasonable to expect the classifier to also demonstrate this consistency.

However, if misclassification results from random guessing, then the classifier should not be expected to replicate them. We created both an upper and a lower bound for human agreement. For the upper bound, we computed the average confusion matrix across observers and then computed the expected agreement between each individual observer and the group average. For the lower bound, we created the group average by holding out the observer being compared with the group.

Results

Participant accuracy

Within each material category, we calculated the proportion of correct responses for both the photographs and drawings of our four material categories (Figure 3). We analyzed these values using a 4 × 2 × 2 mixed-design analysis of variance with material category (metal, stone, water, or wood) and image type (photograph or drawing) as within-subject factors and with participant group (original images or synthetic images) as a between-subjects factor. All statistical analyses were carried out using JASP (2022).

This analysis revealed main effects of both material category, $F(3,150) = 7.50$, $p < 0.001$, and the type of image, original stimuli or texture-synthesized versions, $F(1,50) = 46.739$, $p < 0.001$, that participants categorized. The main effect of material category was driven by a significant difference between performance with images of wood and water, $t = -4.7$, $p < 0.001$, such that wood was categorized more accurately than water across all image manipulations, as well as marginal differences between metal and water, $t = 2.5$, $p = 0.075$, and between stone and water, $t = 2.4$, $p = 0.098$. The main effect of participant group was driven by significantly better performance with the original images as compared with the texture-synthesized versions, $t = 6.8$, $p < 0.001$.

These main effects were qualified by two interactions: A two-way interaction between material category and image format, photographs vs. drawings, $-F(3,150) = 17.96$, $p < 0.001$, and a three-way interaction between all of our factors, $F(3,150) = 4.80$, $p = 0.003$. Because the two-way interaction is qualified by the three-way interaction, we will confine our discussion here to the three-way interaction. To examine this interaction, it is useful to consider how the application of texture synthesis changes the profile of performance across material categories for both our photographs and our drawings. Comparing performance with drawings in Figure 3 and Figure 4, it is evident that the application of texture synthesis did not change the ordinal relationships across material categories: Water remains the most difficult

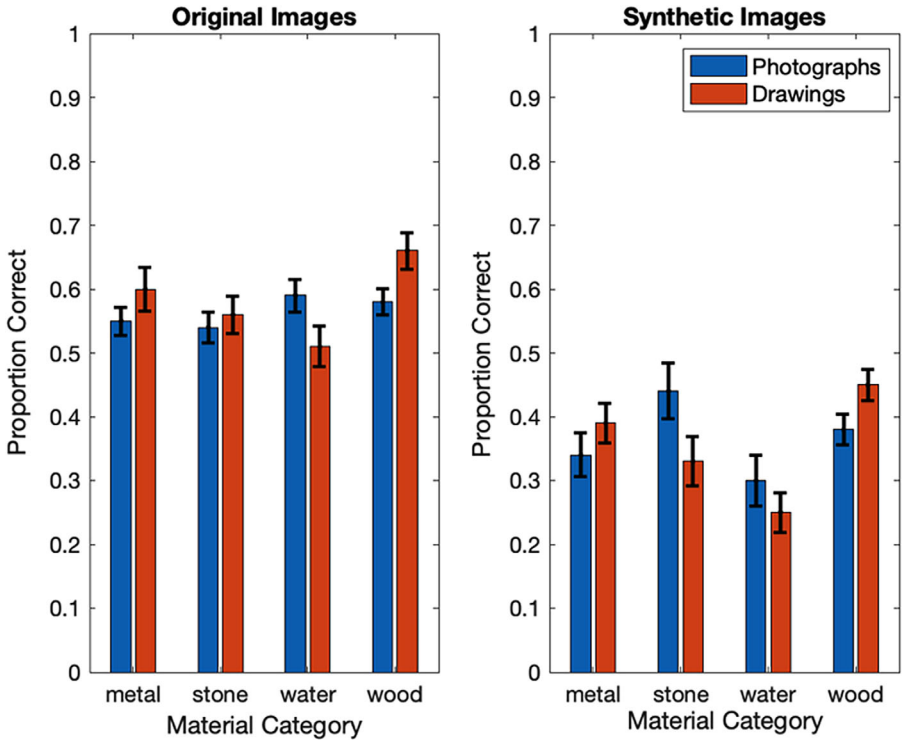

Figure 3. Average proportion correct across participants for all images of original photographs (left) and synthetic images (right). Error bars represent 95% confidence intervals.

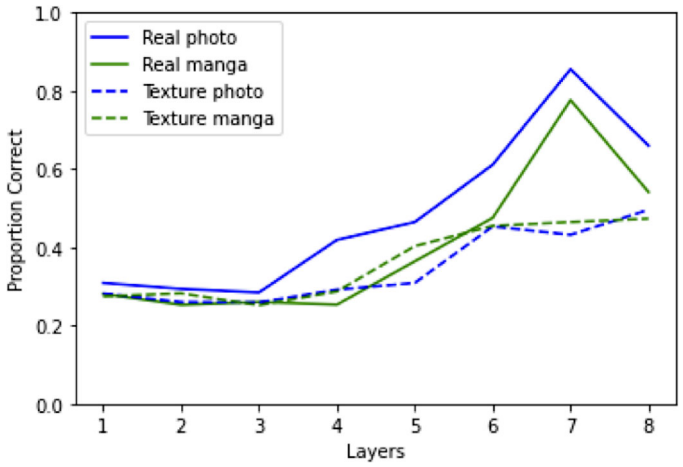

Figure 4. Classification accuracy for each of the four image conditions across each of the eight dCNN layers.

category (statistically lower performance than all three alternative categories), wood the easiest (statistically higher than all three categories), with stone and metal in between. By comparison, performance with photographs does change significantly across these two figures: Performance across material categories is approximately uniform when participants categorized the original stimuli, but the application of texture synthesis introduces significant differences between material categories. Stone emerges as the easiest to categorize correctly, whereas performance with water drops to near-chance levels. Metal and wood are categorized more poorly, but not as poorly as water.

In summary, our three-way interaction can be described in terms of these differential effects of texture synthesis on performance across material categories and image formats: Removing nontexture information from drawings does not differentially impact recognition accuracy across material categories, but removing nontexture information from photographs does do this. We take this as evidence that the selection process artists engage in to present material properties through drawings relies on texture-like features versus location-dependent features to a uniform degree across material categories. By contrast, photographs of these same materials rely on these two classes of features to different degrees as a function of material category.

Participant response time

We also analyzed participants' average response time for correct responses across material categories and stimulus types using the same $4 \times 2 \times 2$ mixed design analysis of variance. This analysis revealed only a significant main effect of material category on performance. This main effect was driven by significantly slower correct classification of metal relative to water and wood, and also significantly slower classification of stone relative to water and wood.

Error analysis

To complement our analysis of correct responses and the response latencies associated with correct material categorization, we also examined the nature of the errors participants made across material categories and stimulus appearance manipulations. Specifically, we wished to examine how confusability between materials may have differed as a function of the presentation of materials as drawings or photographs and the imposition of texture synthesis. We investigated this by tabulating the miscategorizations made of each type per material, segregated both by image type (photographs vs. drawings) and also by participant group (original images vs. synthetic images) and subjecting these contingency tables to a multinomial frequency analysis using JASP. This analysis was necessary to carry out separately for each material because the errors participants can make differ by material category: metal is an error response if the correct category is wood, for example, but cannot be part of the error counts if the correct category is metal. We, therefore, present the outcome of these tests separately for each material, in each case indicating how both image type and participant group may have affected the distribution of errors across material categories. For each material category, we analyzed our count data of errors with error type entered as rows in the contingency table, texture appearance entered as columns in the table, and image type (photographs or drawings) entered as layers in the table. The resulting analysis allows us to comment via χ^2 tests whether the distribution of errors across material categories was significantly affected by texture synthesis for photographs and drawings.

For metal images, we found that the χ^2 test for photographs, $\chi^2 = 64.2$, $p < 0.001$, and the χ^2 test for drawings, $\chi^2 = 25.0$, $p < 0.001$, reached significance. For images of stone, the χ^2 test for photographs did not reach significance, $\chi^2 = 1.6$, $p = 0.46$, while the test for drawings did, $\chi^2 = 38.9$, $p < 0.001$. For images of water, the χ^2 test for photographs, $\chi^2 = 7.80$, $p = 0.020$, and the χ^2 test for drawings, $\chi^2 = 7.84$, $p = 0.020$, reached significance. Finally, for images of wood, the χ^2 test for photographs, $\chi^2 = 141.6$, $p < 0.001$, and the χ^2 test for drawings, $\chi^2 = 42.2$, $p < 0.001$, reached significance. Overall, this analysis demonstrates that texture synthesis tends to have a differential impact on the distribution of errors for photographs versus drawings of materials.

dCNN categorization accuracy

Given these behavioral results, our key questions with regard to dCNN categorization accuracy are two-fold: 1) Does this network replicate the pattern of performance we observed across conditions? and 2) Does the network's agreement with human performance vary according to which layer of the network we query? The former question is a way of asking whether our results can be accounted for by the computations enacted to achieve accurate recognition in another problem domain, whereas the latter helps us to determine the relative contribution of low-level versus high-level processing to the agreement we may observe between humans and the model.

We first examined the classification accuracy at each of the eight dCNN layers. This initial analysis is an important to see if the network is capable of performing above-chance at all and if this depends on which layer of the network we query. As shown in Figure 4, classification accuracy increased across the layers, reaching maximum accuracy in the seventh layer. We performed one-sample t tests at each dCNN layer across the 100 classification results. We found that original photograph classification accuracy was above chance level at each layer, all $p < 0.001$. However, original manga classification accuracy did not exceed chance until the fifth dCNN layer. For texturized photographs, classification accuracy started to exceed chance in the fourth dCNN layer. Finally, the texturized manga images were classified at above-chance levels at all dCNN layers except for the third. In terms of our goal to understand how human–model agreement may depend on the level of processing within the network, this limits our analysis to layers 5 through 7; this constrains our analysis, but also demonstrates that the lowest levels of the network are not effectively capturing the variation we have observed across conditions. Conservatively, this at least suggests that some form of high-level integration of visual features is likely the basis of our behavioral results.

We then examined the classification accuracy across material in detail for layer 7. This layer was chosen as the highest performing across stimulus type and condition. Our main goal is to determine whether or not the dCNN results account for our behavioral data. More specifically, do we find that variation across material categories, synthetic versus original appearance, and photographic versus line drawing presentation is similar in both cases? We present the results of this classification analysis in Figure 5.

Comparing Figure 3 with Figure 5, there are several key similarities that are apparent. The application of texture synthesis incurs a substantial performance cost in both cases, for example. Despite the tendency of dCNNs to rely heavily on texture-like features for

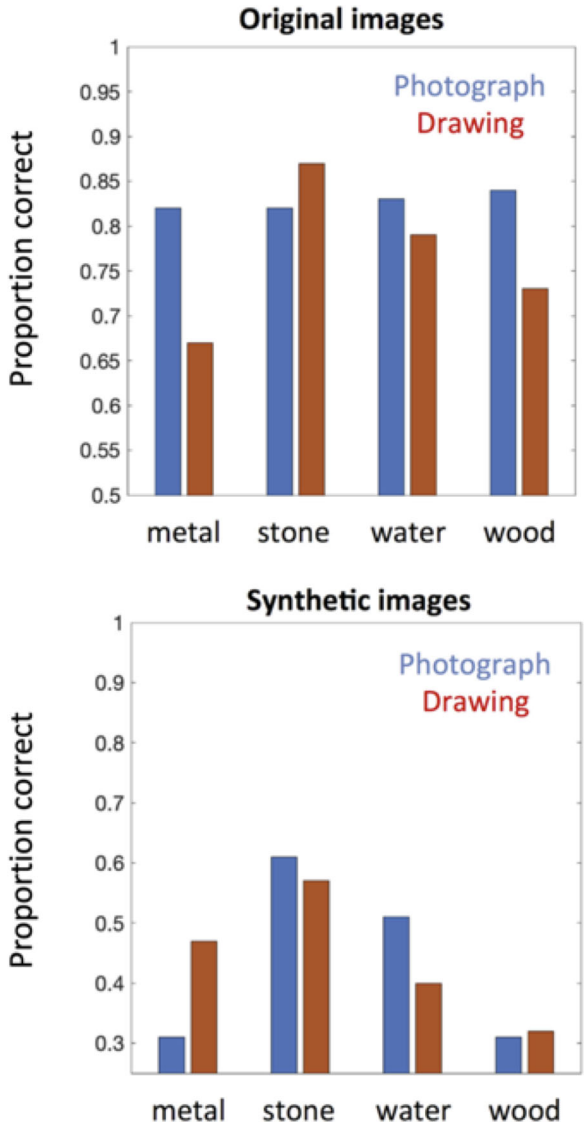

Figure 5. dCNN material categorization accuracy as a function of material, image type, and original (top) versus synthetic appearance (bottom). As described in the main text, multiple iterations of the classification procedure were carried out, but variation across these was sufficiently small that we have not included error bars here.

image categorization (Geirhos et al., 2019; Laskar, Sanchez, & Schwartz, 2020), this result demonstrates that either higher-order texture statistics than those included in the Portilla–Simoncelli feature set or more object-like joint statistics of features and position have relevance to material categorization for both humans and the network. In terms of the ordinal ranking of materials categories across our various stimulus appearance conditions, however, we find there is not substantial agreement between the network and human performance. For example, if we consider the categorization results obtained from the original photographs and drawings, we see that humans'

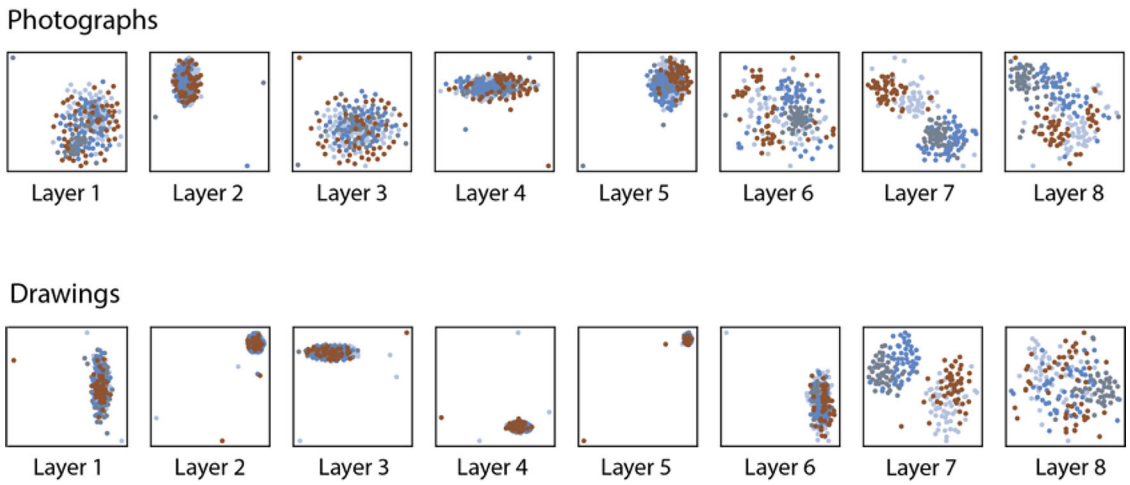

Figure 6. Sample tSNE embeddings of dCNN activations across layers to photographs (top row) and drawings (bottom row) of original material images. Colors indicate different material categories (blue = water; brown = wood; light gray = stone; dark grey = metal).

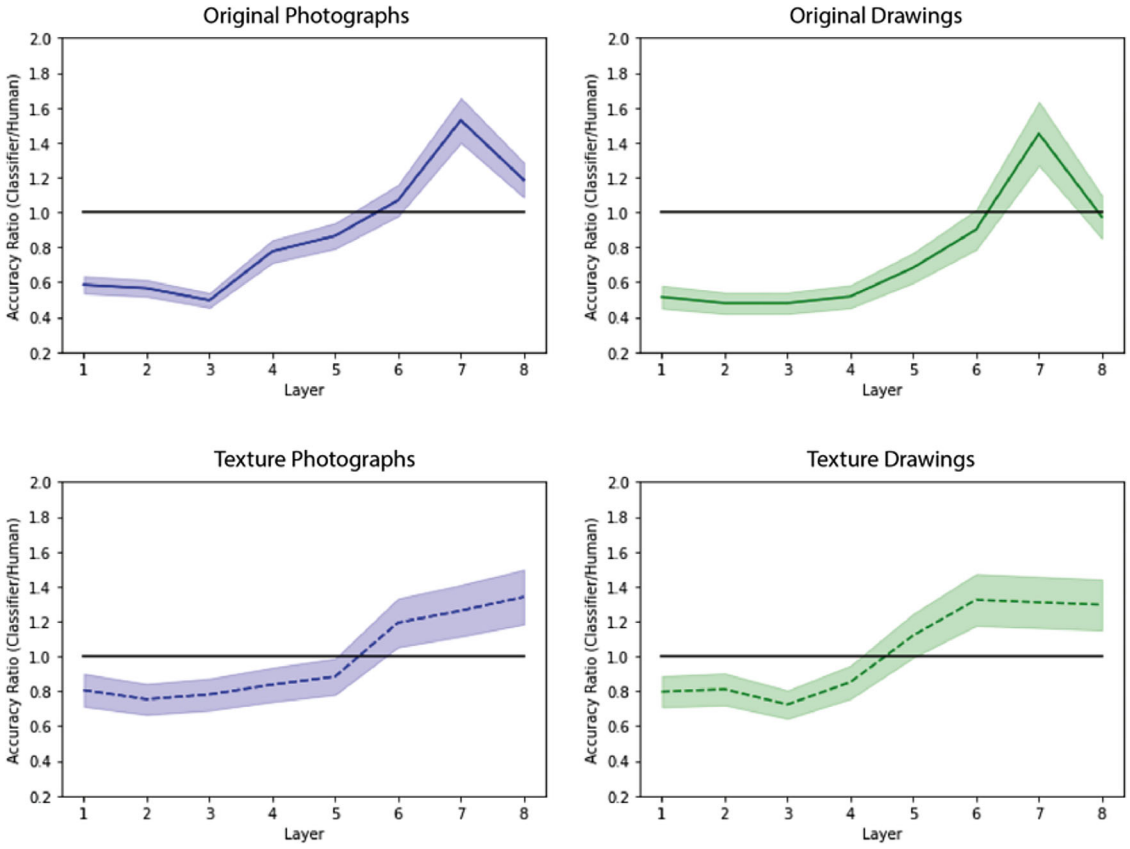

Figure 7. Ratio of classifier to human accuracy as a function of dCNN layer. Shaded error region reflects 95% confidence intervals.

relatively flat performance across material categories is reflected in network accuracy, but that the ordinal ranking of drawings is quite different. Humans find drawings of water the hardest to categorize accurately, whereas the network finds drawings of metal to be the most difficult. Likewise, stone and wood drawings are also reversed relative to human observers in terms of categorization accuracy.

Examining the distribution of these categories in a t-SNE embedding (Figure 6), it is potentially important that metal/water and wood/stone are separable clumps in this space: The reversal of which category is

categorized most easily across humans and the network could be the result of a criterion difference humans maintain favoring one member of each category pair over the other. Considering the results obtained from synthetic images, we find once again that, although the overall pattern of results across photographic stimuli is matched, drawing performance differs between humans and the network. In this case, we observe the same reversal of stone/wood accuracy in the network, but in this case the difference between metal and water accuracy is in the same direction as human observers.

We next examined the ratio of the classifier's accuracy at each of the eight dCNN layers to human observers. As shown in Figure 7, for each stimulus type and condition, human observers outperformed the classifier from the early dCNN layers, but were outperformed by the later dCNN layers. A three-way mixed model analysis of variance with condition (original or texturized) as a between-subjects factor, and stimulus type (photograph or drawing) and dCNN layer as within-subjects factors revealed a significant main effect of condition, $F(1,50) = 7.9$, $p = 7.0e\text{-}3$, ges $= 0.103$, whereby the human observers outperformed the classifier for the original images, ratio $= 0.82$ on average, but not the texturized images, ratio $= 1.01$ on average. We observed no significant main effect of stimulus type, $F(1,50) = 2.02$, $p = 0.16$. We observed a significant main effect of dCNN layer, $F(7,350) = 472.2$, $p = 4.9e\text{-}174$, ges $= 0.49$, whereby the accuracy ratio between the classifier and humans increased with increasing dCNN layer. Moreover, we observed a significant interaction between condition and stimulus type, $F(1,50) = 8.72$, $p = 5e\text{-}3$, ges $= 0.03$. For both drawings and photographs, the accuracy ratio of the classifier to humans was larger for the texturized images than for the originals, suggesting that the classifier outperforms humans in these circumstances. This effect was larger for the drawings than the photographs. We also observed a significant interaction between condition and dCNN layer, $F(7,350) = 36.25$, $p = 5.4e\text{-}26$, ges $= 0.07$. For texture images, the increase in accuracy ratio over dCNN layer was shallower than for original images. Furthermore, we observed a significant interaction between stimulus type and dCNN layer, $F(7,350) = 23.8$, $p = 2.1e\text{-}26$, ges $= 0.008$. Although photographs and drawings had similar accuracy ratios at lower dCNN layers, the accuracy ratio for photographs exceeded drawings in later layers. Finally, we observed a significant three-way interaction between condition, stimulus type, and dCNN layer, $F(7,350) = 42.3$, $p = 4.3e\text{-}43$, ges $= 0.02$.

To more fully examine the similarities and differences between the dCNN and the human observers, we examined the level of agreement in responses relative to what would be expected from the overall performance (see Methods for details). In Figure 8, we scaled the agreements between the expected level (shown as zero on the *y* axis) and the noise ceiling that reflects how well one observer can predict another (shown as one on the *y* axis). For original content (both photos and drawings), human classifier agreement was relatively constant across dCNN layers. However, we observed that human classifier agreement tended to increase with increasing dCNN layer for texturized photos and drawings.

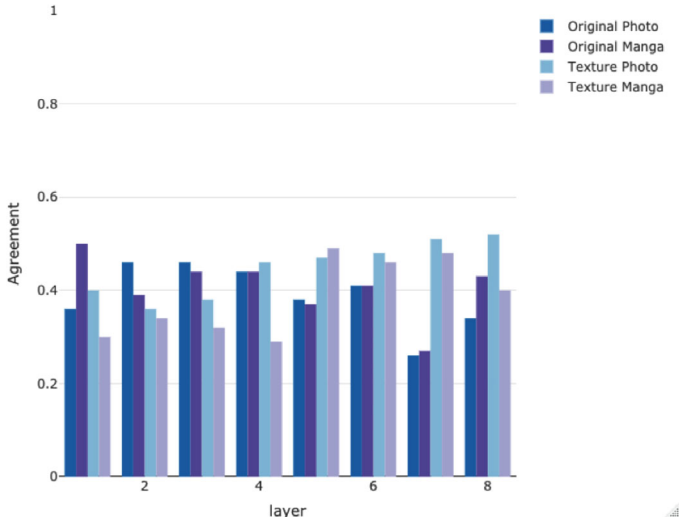

Figure 8. Agreement between classifier and human observers. Zero on the *y*-axis represents the expected level of agreements, based on the overall classification accuracy of that layer, and 1 on the *y*-axis represents the noise ceiling (i.e., how well one observer agrees with another observer).

General discussion

Our behavioral experiment revealed that photographs and drawings of materials differentially rely on nontexture information versus texture statistics across material categories. Joint encoding of local appearance and position varies in its importance in natural images of different materials, but whatever process artists engage in to select and render aspects of material appearance across different categories seems to induce a more consistent balance between these cues. With regard to photographs of images and their texture-synthesized counterparts, our results with these two-tone renderings of natural materials are partly consistent with prior work (Balas et al., 2017). In particular, as we observed in prior work examining how synthetic materials were recognized compared with recognition of original images presented in the fovea and the visual periphery, there is little difference in accuracy across categories when unaltered images are available to observers, but the imposition of texture statistics introduces variation across categories. The same pattern of ordinal results

was observed in our previous work with full-color natural images and the current study, which provides a useful replication of that data and validates our claims with regard to photographic images of natural material categories after the binarization transform applied to make photographs more comparable with our drawing stimuli.

Our initial hypothesis was that materials that are typically glossy in appearance, like metal and water, would be most affected by texture synthesis owing to the diagnostic value of specularities coinciding with specific surface geometry (Fleming & Bulthoff, 2005; Ho, Landy, & Maloney, 2008; Anderson & Kim, 2009). We further expected that artists might preferentially use that information to depict these materials, leading to a disproportionate cost of texture synthesis on drawings of glossy materials. Instead, our behavioral results indicate a different pattern that reveals some surprising properties of drawings and photographs of these natural material categories. Images of water, whether they are photographs or drawings, are profoundly affected by the imposition of texture synthesis. Participants were at or near chance when attempting to categorize synthetic versions of these images in both formats, which demonstrates that texture information alone is definitely not sufficient to signal the presence of water either in drawings or in natural images. This is consistent with some early attempts to capture the appearance of water using computer graphics (Kung & Richards, 1988), which relied heavily on a physical model of light interacting with water. Our data suggest that, although aspects of water may be texture like (e.g., ripples that extend across the surface of a pond), reliable categorization of an image as water is extremely difficult when only texture statistics have been preserved. Another surprising feature of our data is the differential impact of texture synthesis on drawings and photographs of stone. Specifically, drawings of stone are disproportionately impacted by texture synthesis relative to photographs. This pattern of results is what we would expect if artists tend to include specific features in specific positions to communicate stoniness rather than approximate the texture of stone in drawings. Although the presence of this pattern of results with regard to stone in particular is contrary to our initial hypotheses, it nonetheless supports our broader idea that drawings may differ from photographs of materials in terms of the sufficiency of texture statistics for carrying category information. An interesting possibility to consider is whether artists' tend to use specific techniques to render a stony texture that incorporate small patches of texture in critical places within an outline to convey stone-like material without covering the entire surface in stippling or hatching.

None of these intriguing results are reflected in the outcome of our DNN analysis. Although the accuracy we obtain from the model improves as we look at higher levels of the network, the model also predicts a consistent advantage for photographs over drawings, which is not what we observed in human observers. Thus, though our analysis confirms prior work demonstrating that drawings of materials can be classified accurately by DNNs (Horiuchi, Saito, & Hirai, 2017), it also reinforces the fragility of these models to generalize beyond the statistical relationships in their training sets.

Interestingly, although classifiers from early DNN layers were less accurate than human observers, classifiers from later layers exceeded human performance in all conditions. This effect was particularly strong for the original photographs and drawings and may reflect the fact that DNNs have been shown to be more reliant on texture compared with human observers (Geirhos et al., 2019; Laskar et al., 2020). Upon examining the item-by-item agreement between DNNs and human observers, we found that, in all conditions, agreement was higher than what would be statistically expected based on overall performance, but substantially under the average agreement between individual human observers. Additionally, we found subtle differences in the level of agreement across dCNN layers. For original images, we observed the most human classifier agreement in the early-to-mid dCNN layers. By contrast, texturized images had the most agreement in the later DNN layers.

That our DNN analysis does not match the human data raises the interesting question of exactly what the human visual system is doing differently that leads to the pattern of results we observed in our recognition task. As we noted elsewhere in this article, dCNNs that succeed at object recognition seem to do so via the use of more texture-dependent processing than human observers do (Geirhos et al., 2019; Laskar et al., 2020), but what insights do the current results provide about how drawings in particular may be perceived? An interesting possibility suggested by our data is that human observers may be ultimately outperformed by the dCNN at later levels of the network owing to an under-reliance on texture information by humans when shape and other higher-level features are available in original photographs and drawings. The fact that the classifier/human ratio is lower when images have been subject to texture synthesis may be the result of human observers being unable to be misled by these aspects of appearance when they are removed by the Portilla–Simoncelli model. Further exploring this relationship by limiting human observers' ability to use object and shape features in other ways (e.g., aperturing) and looking for increases in material categorization may be an interesting way to characterize feature use in complex texture images.

With regard to photographic renderings versus drawings of materials, another intriguing possibility is

that drawings may include diagnostic features that do not have a direct correspondence to the real appearance of materials. McCloud (1994) suggests that comics (and drawings more generally) are interpretable by observers largely because of an accumulated iconic vocabulary that offers a shorthand for extracting meaning from drawings. That is, observers must learn artistic conventions (that may vary cross-culturally) so that they can successfully recognize the content of line drawings and other artwork based on the presence of specific features that signal objects, surfaces, and materials. The nature of such an iconic vocabulary is perhaps best exemplified by considering the use of zip lines in comics to depict object motion (Ito, Seno, & Yamanaka, 2010). These trailing lines that explicitly depict the path of a point on an object as it passes through space do not reflect a real physical entity observers have experience seeing, but are instead a matter of convention: Artists and viewers have agreed that these elements imply motion. Some of what DNNs may not capture about the human recognition of material categories in drawings may be related to the existence of such an abstract shorthand for signaling material categories that human observers understand, but that is not easily captured in this framework. An interesting question to ask about the nature of such an iconic vocabulary for recognizing materials is how it is acquired. In prior work, school-age children have been found to be poorer than adults at using texture statistics to categorize or name materials, but these performance differences are less evident in a matching task that does not require labeling (Balas & Schmidt, 2017). One possibility is that these results may reflect the slow acquisition and refinement of an iconic vocabulary for material categorization, one that may be especially useful for interpreting drawings. Compared with adults, we would predict that children may not be as affected by the imposition of texture synthesis on material drawings, because they may lack the iconic vocabulary that contributes category information above and beyond what texture statistics provide.

The current study does include some important limitations. First, our choice to examine just four material categories potentially limits the generalizability of our results. Indeed, our data suggest that the impact of texture synthesis on the perception of photographs versus drawings varies across material, so examining a broader range of material categories is likely to yield different outcomes. We suggest that this is a potentially important direction for future work, especially with regard to different strategies for rendering specific materials in drawings that incorporate extended texture patterns versus local features placed in particular locations. Second, our examination of texture statistics is confined to the class of wavelet correlations implemented via the Portilla-Simoncelli algorithm. There are many other candidate models of texture processing to consider, and the different assumptions these models make about the basic vocabulary of texture processing may also influence the outcome of our tasks. Finally, we also note that varying the nature of the dCNN we use to attempt to capture human behavior may also be an important way to understand how drawings and photographs of these materials are processed. What type of training is necessary to replicate our results? Do different regimes of network tuning via structured training lead to closer agreement with human observers?

These limitations notwithstanding, the current results demonstrate that drawings of materials and photographs of materials rely on texture statistics to different extents across material categories. Our data point to intriguing additional questions regarding how materials and material properties are depicted in artwork and recognized by human observers. In future work, we aim to examine these issues using a broader set of material properties and a range of different feature vocabularies for describing image structure in terms of low-level, mid-level, and high-level descriptors.

Keywords: material perception, drawing, texture perception, deep neural networks

Acknowledgments

Commercial relationships: none.
Corresponding author: Benjamin Balas.
Email: benjamin.balas@ndsu.edu.
Address: Psychology Department, 1210 Albrecht Blvd., North Dakota State University, Fargo, ND 58102-6050, USA.

References

Anderson, B. L., & Kim, J. (2009), Image statistics do not explain the perception of gloss and lightness. *Journal of Vision, 9*(11), 1–17, https://doi.org/10.1167/9.11.10.

Balas, B., Conlin, C., & Shipman, D. (2017). Summary statistics and material perception in the visual periphery. *ACM Transactions on Applied Perception, 13*(2):8, 1–13.

Balas, B., & Schmidt, J. (2017). Children's use of visual summary-statistics for material categorization. *Journal of Vision, 17,* 1–11.

Balas, B. J. (2006). Texture synthesis and perception: Using computational models to study texture

representations in the human visual system. *Vision Research, 46*(3), 299–309.

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision, 9*(12), 1–18.

Balas, B., & Conlin, C. (2015). The visual N1 is sensitive to deviations from natural texture appearance. *PloS One, 10*(9), e0136471.

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

Baumgartner, E., & Gegenfurtner, K. R. (2016). Image statistics and the representation of material properties in the visual cortex. *Frontiers in Psychology, 7*, 1185.

Cohn, N., & Ehly, S. (2016). The vocabulary of manga: Visual morphology in dialects of Japanese Visual Language. *Journal of Pragmatics, 92,* 17–29.

Deng, J., Wei, D., Socher, R., Li-Jia, L., Kai, L., & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, 248–255.

Di Cicco, F., Wjintjes, M., & Pont, S. C. (2020). If painters give you lemons, squeeze the knowledge out of them. A study on the visual perception of the translucent and juicy appearance of citrus fruits in paintings. *Journal of Vision, 20*(13), 12.

Di Cicco, F., Wjintjes, M. W. A., & Pont, S. C. (2019). Understanding gloss perception through the lens of art: Combining perception, image analysis and painting recipes of 17th century painted grapes. *Journal of Vision, 19*(3), 1–15, https://doi.org/10.1167/19.3.7.

Fleming, R. W., & Bülthoff, H. H. (2005). Low-level image cues in the perception of translucent materials. *ACM Transactions on Applied Perception, 2*(3), 346–382.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations (ICLR 2019) May 6–9, 2019*, New Orleans, Louisiana.

Hertzmann, A. (2021). The role of edges in line drawing perception. *Perception, 50*(3), 266–275.

Hertzmann, A. (2020). Why do line drawings work? a realism hypothesis. *Perception, 49*(4), 439–451.

Hinton, G., & Roweis, S. (2002). Stochastic neighbor embedding. In *Proceedings of the 15th International Conference on Neural Information Processing Systems (NIPS'02)*. Cambridge, MA: MIT Press; 857–864.

Ho, Y.-H., Landy, M. S., & Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture. *Psychological Science, 19*(2), 196–204.

Horiuchi, T., Saito, Y., & Hirai, K. (2017). Analysis of material representation of manga line drawings using convolutional neural networks. *Journal Imaging Science and Technology, 61*(4), 40404–1–10.

Ito, H., Seno, T., & Yamanaka, M. (2010). Motion impressions enhanced by converging motion lines. *Perception, 39*(11), 1555–1561.

JASP Team. (2022). JASP (Version 0.16.1)[Computer software].

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84–90.

Kung, T. J., & Richards, W. A. (1988). Inferring "water" from images. In W. A. Richards (Ed.) *Natural computation*. Cambridge, MA: MIT Press. pp. 224–233.

Laskar, M., Sanchez Giraldo, L. G., & Schwartz, O. (2020). Deep neural networks capture texture sensitivity in V2. *Journal of Vision, 20*(7), 21.

McCloud, S. (1994). *Understanding comics: Writing and art*. New York: Harper Perennial.

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities, *Nature, 447*(7141), 206–209.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics, 9*, 62–66.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*, 49–70.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision, 12*(4), 12.1.14 14. https://doi.org/10.1167/12.4.14 14.

Saito, Y., Hirai, K., & Horiuchi, T. (2015). Construction of manga materials database for analyzing perception of materials in line drawings. *Color and Imaging Conference, 2015*(1), 201–206.

Sayim, B., & Cavanagh, P. (2011). What line drawings reveal about the visual brain. *Frontiers in Human Neuroscience, 5*, 118.

Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of Vision, 14*(9), 12.

Tadros, T., Cullen, N. C., Greene, M. R., & Cooper, E. A. (2019). Assessing neural network scene classification from degraded images. *ACM Transactions on Applied Perception, 16*(4), 1–20, https://doi.org/10.1145/3342349.

Van Zuijlen, M. J. P., Pont, S. C., & Wjintjes, M. W. A. (2020). Painterly depiction of material properties. *Journal of Vision, 20*(7), 1–17, https://doi.org/10.1167/jov.20.7.7.