The psychophysics of style

Tal Boger[1*], Chaz Firestone[1*]

Affiliations:

[1]Department of Psychological and Brain Sciences, Johns Hopkins University; Baltimore, 21218, USA.

*Corresponding authors. Email: tboger1@jhu.edu, chaz@jhu.edu

Abstract:

Among the most significant modes of human creative expression is style: the capacity to represent objects, events, and scenes (e.g., lilies dotting a pond) in some distinctive manner (e.g., Monet's broken brushstrokes and blended colors). Diverse research traditions analyze the social, political, and aesthetic significance of stylistic representation. But what are the cognitive and computational foundations of this capacity? Here, we characterize style perception as a process that "parses" form from content, and adapt classic psychophysical paradigms to discover multiple new phenomena of style perception. Using both naturalistic images and synthetic stimuli, ten experiments reveal perceptual 'tuning' to stylistic information, representational constancy over stylistic variation, and mental rendering of novel styled objects. Moreover, an object recognition model further grounds style perception by capturing human judgments of image similarity over different styles. Together, this work illuminates the psychological foundations of stylistic perception, and opens the door to further investigation of styled media.

In Press, *Nature Human Behaviour*

Main Text

Introduction

When looking at a painting, such as Van Gogh's Starry Night (Figure 1A), what do we see? Certainly we see the painting's subject — a French village beneath a night sky, viewed from the window of an elevated monastery. Equally salient, though, is the painting's style — its dark palette of blues and yellows, whirl of spiraling textures, and dreamlike aura. In other words, the scene is portrayed in a certain manner, which is as much a part of the painting as the scene itself.

A distinctive aspect of style is that it can vary independently of content. For example, the same village scene might look entirely different if painted by a realist aiming to preserve naturalistic details, and different still if painted by an abstract expressionist wishing to convey emotion or inner experience. This distinction also arises outside of art galleries and museums, as when we appreciate a piece of clothing or furniture, an unusual set of cutlery, or a row of homes in a neighborhood. For example, a fork in a cutlery set is likely to have tines and a handle — but its shape, finish, and ornamentation may be subject to stylistic variation. Similarly, a house generally requires a door, roof, windows, and space for inhabitants — but the size, layout, and appearance of these elements may differ in a Victorian home as compared to a cottage or ranch.

The ubiquity and salience of style have generated longstanding scholarly interest in a variety of research traditions, including sociology (1, 2), history (3), and of course art theory (4, 5). But how does the mind separate style from content in the first place? Surprisingly, little is known about the psychological basis of visual style perception. Of course, there is a rich psychological literature on visual aesthetics (6-8), which has explored the patterns humans prefer (9-13), to what extent aesthetic preferences reflect stable traits of individuals (14), and which patterns of neural activity are associated with aesthetic experiences (15-17). However, a psychologically grounded account of visual style itself has been elusive (cf. 18-21). Thus, a fundamental question remains unanswered (and even unasked): What is the nature of stylistic perception, what psychological mechanisms does it draw upon, and what are its psychophysical signatures?

Here, we address these questions by drawing on methods and insights from both classic psychophysical studies and recent advances in generative artificial intelligence. A long tradition in experimental psychology explores how human perception 'parses' the content of a stimulus from its context or conditions of presentation — as when we achieve color constancy over different illumination conditions (22, 23), adapt to accented speech (24), or extract letter identities from different typefaces (25). More recently, modern machine-learning technologies have enabled the synthesis of stylized images, whereby a model can extract aspects of artistic style from one image (e.g., Starry Night) and then flexibly apply them to any other image (e.g., ordinary natural scenes) — a technique known as 'style transfer' (26, 27; Figure 1B). For example, this process can create novel images of mountains, beaches, bedrooms, and libraries in the style of Starry Night, or indeed any other painting (Figure 1C).

The present work combines these approaches to investigate the cognitive mechanisms underlying style perception. We conceive of style perception as akin to the well-characterized parsing processes mentioned above (for a similar approach outside the context of artistic style, see 28), and then explore those processes using newly available style-transfer techniques (as well as naturally occurring styled images, such as styled sets of cutlery). Framing style perception as an instance of the mind parsing content from form opens the door to adapting established psychophysical paradigms to the study of artistic style. And using style-transfer techniques allows for the generation of a large imageset that varies mostly or only in style (while preserving underlying content, composition, etc.) in ways that would be difficult or impossible to achieve with purely naturalistic images.

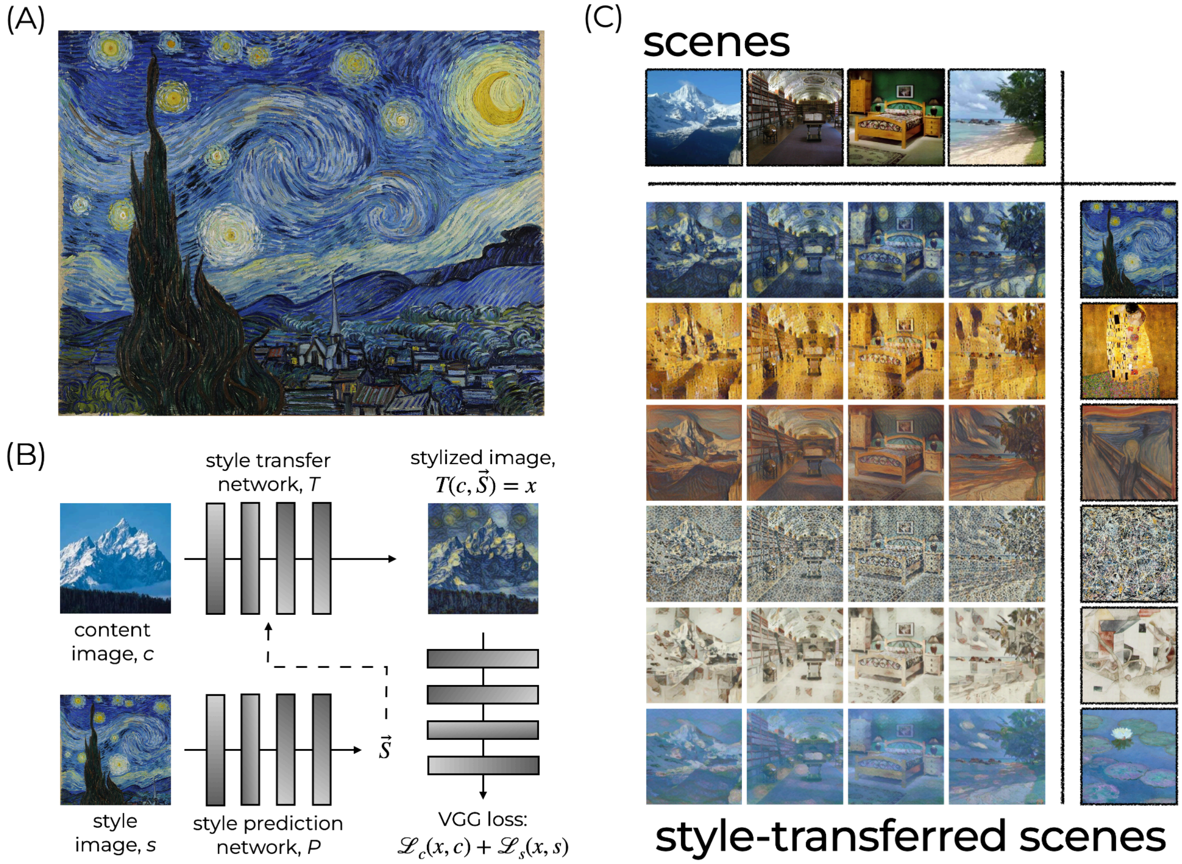
(A)
(B)
style transfer network, T
stylized image, $T(c, \vec{S}) = x$
content image, c
style image, s
style prediction network, P
$\vec{S}$
VGG loss: $\mathscr{L}_c(x, c) + \mathscr{L}_s(x, s)$
(C)
scenes
styles
style-transferred scenes

Figure 1. Extracting and applying artistic style. (A) Starry Night by Vincent Van Gogh. Readers are invited to notice not only the content of the image (a French village beneath a night sky), but also its style, including its palette, swirling textures, and dreamlike aura. (B) Schematic depiction of 'style transfer' (26, 27), a process that analyzes a 'style image' s to infer an embedding vector $\vec{S}$, which is then transferred to a 'content image' c. This results in an image, x, which depicts the content image in the given style. In generating this image, the network is trained to minimize loss in both style (i.e., the style difference between x and s) and content (i.e., the content difference between x and c), here defined by VGG embedding vector distances. (C) Many of our experiments exploit this process to study the perception of style, by generating natural scenes (mountains, libraries, bedrooms, beaches; 29) in the styles of famous paintings (e.g., Van Gogh's Starry Night, Monet's Water Lilies, Klimt's The Kiss, etc.) and placing them in adaptations of classic psychophysical paradigms.

Here, in 10 pre-registered experiments, this approach reveals multiple new phenomena of style perception (Figure 2). These results both (a) constitute new discoveries in their own right, and (b) testify to the promise of 'parsing' as a working model for the study of style perception. Finally, we show that a computer-vision model trained on object recognition (ResNet-18, 30) predicts subjective impressions of similarity across styles. Together, this theoretical framework and empirical results help to illuminate the cognitive and computational basis of style perception.

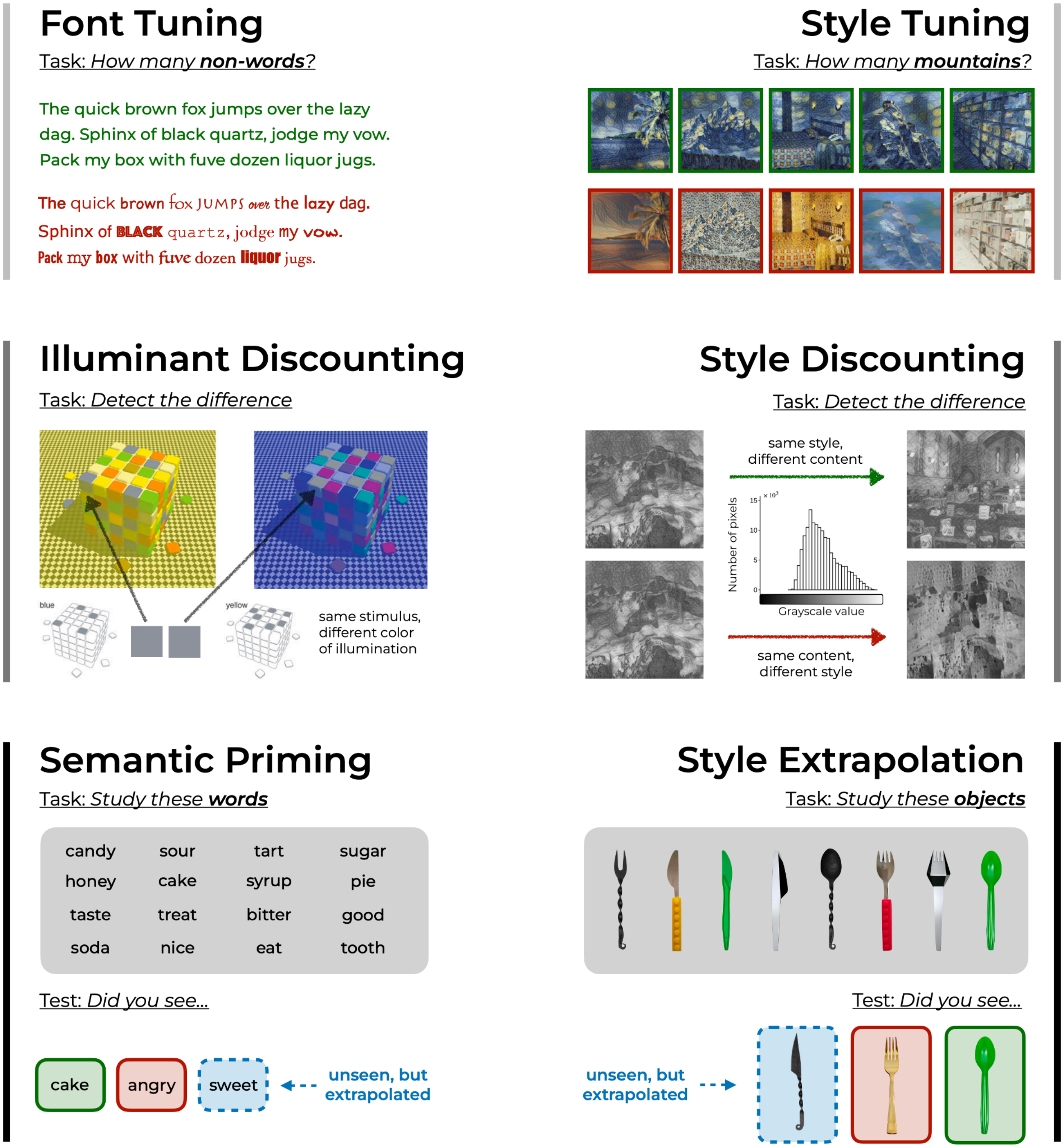
Font Tuning
Task: How many non-words?
The quick brown fox jumps over the lazy dag. Sphinx of black quartz, jodge my vow. Pack my box with fuve dozen liquor jugs.
The quick brown fox JUMPS over the lazy dag.
Sphinx of BLACK quartz, jodge my vow.
Pack my box with fuve dozen liquor jugs.
Style Tuning
Task: How many mountains?
Illuminant Discounting
Task: Detect the difference
blue
yellow
same stimulus, different color of illumination
Style Discounting
Task: Detect the difference
same style, different content
× 10³
15
10
5
0
Number of pixels
Grayscale value
same content, different style
Semantic Priming
Task: Study these words
candy
sour
tart
sugar
honey
cake
syrup
pie
taste
treat
bitter
good
soda
nice
eat
tooth
Test: Did you see...
cake
angry
sweet
unseen, but extrapolated
Style Extrapolation
Task: Study these objects
Test: Did you see...
unseen, but extrapolated

Figure 2. Adapting psychophysical paradigms to study artistic style. (Top) In font tuning, readers become more fluent and accurate after adjusting to the typeface of text; the present experiments adapt this paradigm to introduce style tuning, whereby target detection improves over time when artistic style is held constant. (Middle) A core visual process is discounting the illuminant, enabling perceivers to see the same surface colors across different illumination conditions. Here, we explore an analogous process—style discounting—enabling perceivers to see the same scene content across different artistic styles. A change-detection task tests whether scene changes are more easily detected than style changes, even when the two change types are equated for image similarity. (Bottom) In semantic priming, salient properties of items one has encountered can create false memories in which one internally generates representations of items one has not encountered (e.g., misremembering 'sweet' after encountering 'candy', 'taste', and 'treat'). The present work reveals a similar pattern in style perception: After seeing a spoon and fork from a cutlery set with a given style, the mind may generate a representation of the knife from that set, despite never having encountered it before. Such style extrapolation implies that the mind integrates the style of items one has seen with other background knowledge to infer the likely appearance of unseen objects.

Results

Experiments 1–4: Style tuning

Our first set of experiments was inspired by font tuning, whereby observers adapt to typefaces in ways that aid reading fluency and letter recognition (25). In a typical font tuning paradigm, subjects see a passage of text in either a single typeface or multiple typefaces, and are tasked with making judgments about the presented text (e.g., counting how many tokens are non-words as opposed to words). The key finding is that text appearing in a single typeface is more easily read than text appearing in multiple typefaces, even if each typeface is familiar and otherwise readable on its own — suggesting a tuning process whereby perception extracts (and then adapts to) the font in which the text is rendered. However, letters and typefaces constitute a fairly circumscribed case, due to relatively limited dimensions of typeface variation and a constrained set of underlying contents (i.e., the letters of the alphabet). Might a similar phenomenon arise in the more complex and open-ended context of artistic style?

Experiment 1 adapted the font-tuning paradigm to visual style (Figure 3). Using the style transfer model described in Ghiasi et al. (27, which adapts a model proposed in 26), we generated a stimulus set consisting of natural images of scenes (e.g., mountains and libraries; 29) rendered in the style of famous paintings (e.g., Van Gogh's Starry Night, Monet's Water Lilies, etc.). We then designed a 'style tuning' task using these images. On each trial, subjects viewed a row of images (analogous to a sentence in font tuning) and simply had to count how many images depicted mountains (or one of the other scene types, randomly assigned to each subject). Crucially, in half of trials, the images appeared in a single style; in the other half of trials, the images appeared in multiple styles. (Interested readers may view all tasks, along with a repository containing all pre-registrations, data, and analyses, at https://perceptionresearch.org/style). We predicted that, just as same-typeface sentences are easier to read than mixed-typeface sentences (resulting in faster reading times), same-*style* image arrays would be more easily processed than mixed-*style* image arrays (resulting in faster scene-identification times) — reflecting the visual system's ability to learn a mapping between styles and scene identities.

Indeed, subjects were significantly faster to enumerate scenes in same-style trials (M = 5418ms) than in mixed-style trials (M = 5707ms, difference = 289ms; $t(43) = 4.93$, $p < 0.001$, $d = 0.74$, 95% CI = [171ms, 407ms]; this and all other *t*-tests reported here are two-tailed dependent-samples tests over subject-level means). This speed advantage did not come at the expense of accuracy, which was also higher on same-style trials (M = 72.03%) than on mixed-style trials (M = 65.29%, difference = 6.74%; $t(43) = 6.10$, $p < 0.001$, $d = 0.92$, 95% CI = [4.51%, 8.97%]; pre-registered as a secondary analysis for this experiment). Thus, just as perception adapts to the typeface of text or even the accent of a speaker, it also adapts to an image's style — a novel phenomenon we refer to as 'style tuning'.

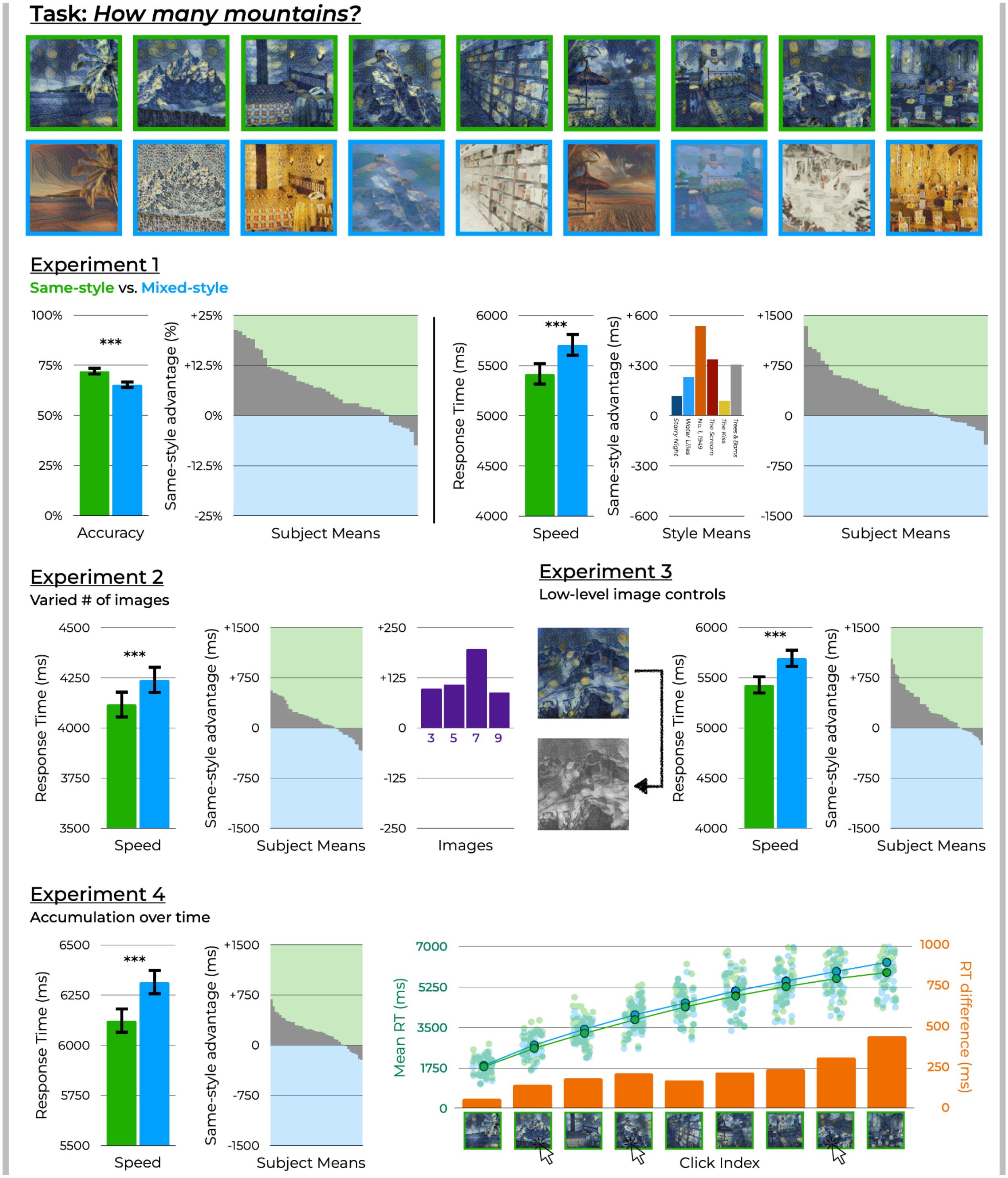

Figure 3. Style tuning. Experiments 1–4 adapted font tuning paradigms by asking subjects to enumerate a target scene type (e.g., mountains) within an array of styled images. Half of the time, all of the images appeared in the same style (e.g., as seen in the green outlines above); the other half of the time, the images appeared in a mix of styles (e.g., as seen in the blue outlines above). Subjects were significantly faster and more accurate on same-style arrays than mixed-style arrays, not only on the level of same-style vs. mixed-style but also on the style- and subject-level (Experiment 1, N = 44). This effect arose for arrays of as few as three images (Experiment 2, N = 46), survived controls for low-level image properties (Experiment 3, N = 44), and accumulated over time (Experiment 4, N = 42). Data are presented as means +/- 95% confidence intervals of the difference between conditions. All statistical tests are paired, two-sided t-tests; *** = *p* < .001.

Experiment 2 varied the number of images in the arrays (3, 5, 7, or 9) to ask how quickly the mind adapts to style. Might these effects arise with very few examples? Indeed, we found that style tuning occurred both across the full sample (i.e., at the same-style vs. mixed style level; mean difference = 123ms; $t(45) = 3.53$, $p < 0.001$, $d = 0.52$, 95% CI = [53ms, 193ms]) and at each subsample (three-image arrays: 98ms; five-image arrays: 108ms; seven-image arrays: 197ms; nine-image arrays: 89ms). Thus, style tuning occurs with as few as 3 examples. The rapid nature of style tuning continues to mirror font tuning, which arises not just for words in a longer paragraph but even for individual letters in a short string (25).

Experiment 3 asked whether low-level image differences may be driving style tuning effects. Notice, for example, that Van-Gogh-styled scenes tend to be blue and yellow, while Munch-style scenes tend to be reddish, and that the images vary along other dimensions as well (e.g., luminance); perhaps, then, our results are driven by the difficulty of jumping from a blue scene to a red one, or from a dark scene to a bright one. To control for such factors, we repeated Experiment 1 with grayscale, luminance-matched versions of our style-transferred stimuli (created using the SHINE toolbox in MATLAB; *31*). Previous work shows that neural networks trained to classify style do so in ways that go beyond the color distributions of the images (32), implying that style perception might persist even without these features. Indeed, we found that style tuning survives these low-level controls: Subjects in Experiment 3 were still faster (mean difference = 267ms, $t(43) = 5.61$, $p < 0.001$, $d = 0.85$, 95% CI = [171ms, 364ms]) and more accurate (mean difference = 8.26%, $t(43) = 7.37$, $p < 0.001$, $d = 1.11$, 95% CI = [6.00%, 10.53%]) on same-style trials than mixed-style trials.

Finally, Experiment 4 explored the timecourse of style tuning. Instead of enumerating the mountain scenes by entering a single response at the end of a trial, subjects in this experiment clicked on each target image with their cursor, thereby providing multiple responses to analyze in each trial. We discovered that tuning accumulates over the course of a trial: The further into an image array, the larger the same-style advantage ($r(376) = 0.27$, $p < 0.001$, 95% CI = [0.17, 0.36]). To further examine this pattern, we fit a linear mixed-effects model that predicts response times, with a random effect of subject and fixed effects of click index (i.e., how many images were previously clicked), trial type (same-style or mixed-style), and their interaction. We hypothesized that the interaction of click index and trial type would significantly predict response times, as expected if the same-style advantage accumulates over time; indeed, this interaction was significant ($t(711) = 3.25$, $p < 0.01$). As before, subjects were also faster at same-style trials than mixed-style trials (mean difference = 192ms, $t(41) = 5.89$, $p < 0.001$, $d = 0.91$, 95% CI = [126ms, 258ms]). Together, these results demonstrate that style tuning exists, onsets rapidly, survives low-level controls, and accumulates over time.

Experiments 5–6: Style discounting

As noted earlier, our approach is to conceive of style perception as a process that parses an image into two components: The content being portrayed, and the manner in which it is portrayed. Perhaps the most foundational example of such a process in visual perception is color constancy — the ability to perceive the 'same' reflectance properties under different conditions of illumination (22, 23). For example, we can see the blue, yellow, red and green squares of a Rubik's cube, and they will typically look to have those colors even across different lighting conditions (e.g., yellowish daylight or neutral fluorescent light). In such cases, vision 'discounts the illuminant' — essentially seeing through the lighting conditions to extract the underlying color of the surface being depicted. Recent work in experimental psychology has shown that such discounting occurs even for higher-level visual processes, such as when we ‘see through’ a cloth to discern the shape of the object beneath it (33; see also 34). This process was investigated using a change-detection task, wherein subjects saw a sequence of two images depicting cloth-covered objects. Changes were more perceptible when the second image showed a different object draped similarly (an ‘underlying object change’) than when the second image showed a similar object draped differently (a ‘cloth change’), even when these changes were equated on certain image metrics. Does a similar process arise for style, and can it be studied the same way?

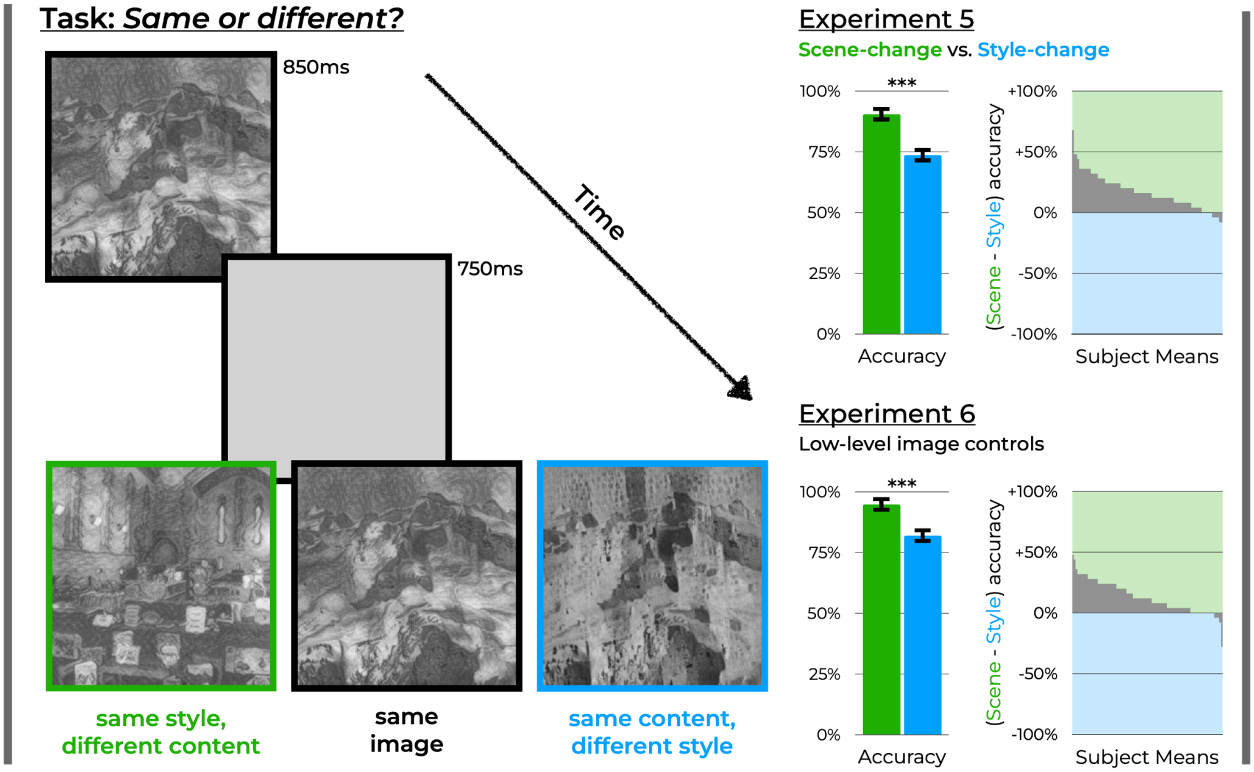

Figure 4. Style discounting. In Experiments 5–6, subjects judged whether two sequentially appearing images were the same or different. We found that changes to the underlying scene (with style held constant, shown in green) were more detectable than changes to the style (with the underlying scene held constant, shown in blue; Experiment 5, N = 87), even when low-level image statistics would predict the opposite pattern (Experiment 6; N = 89). This effect mirrors other discounting effects in vision, demonstrating that visual processing can subtract out the effects of style so as to represent the underlying scene content. Data are presented as means +/- 95% confidence intervals of the difference between conditions. All statistical tests are paired, two-sided t-tests; *** = $p < .001$ in paired, two-sided t-test.

Experiments 5–6 adapted this design for style perception, essentially asking whether vision engages in an analogous ‘style discounting’ process (Figure 4). In Experiment 5, subjects briefly viewed an image (e.g., a Van-Gogh-styled beach), which then disappeared and was replaced by a new image. Subjects then had to say whether the two images were the same or different from each other. Half of trials depicted the same image, and the other half depicted a different image. The different-image trials were themselves equally split between images depicting the same scene in a different style (e.g., the same beach but now in the style of Munch; style-change trials), or a different scene in the same style (e.g., a library in the style of Van Gogh’s Starry Night; scene-change trials). Following previous work (33), both change types were equated in terms of embedding distances in a Convolutional Neural Network (here, ResNet-18; 30), such that the images were equally different from one another (from the point-of-view of the CNN) across the two experimental conditions. Nevertheless, subjects performed significantly better at scene-change trials than style-change trials (mean difference = 16.83%, *t*(86) = 11.30, *p* < 0.001, *d* = 1.21, 95% CI = [13.87%, 19.79%]), as would be expected if vision engages in style discounting.

Experiment 6 replicated this result with more extensive controls for image similarity. Although Experiment 5 equated the distance of ResNet embeddings, it is possible that the change types still differed on other lower-level image metrics. Thus, Experiment 6 subsampled the most similar scene-change pairs and least similar style-change pairs, such that multiple image statistics — including mean-squared error of pixel changes, structural similarity (35), and ResNet embeddings — were not only similar but if anything would predict the opposite trend (because scene-change pairs were more similar to one another on average than style-change pairs). Remarkably, subjects still performed significantly better on scene-change trials (mean difference = 12.90%, *t*(88) = 9.44, *p* < 0.001, *d* = 1.00, 95% CI = [10.18%, 15.61%]). This provides especially compelling evidence for style discounting, a new phenomenon in which perception sees through the style of an image to extract its underlying content, in ways analogous to other discounting processes in vision.

Experiments 7–9: Style extrapolation

Whereas ‘style tuning’ and ‘style discounting’ suggest that the mind extracts an image’s style to better perceive its content, we may also extract an image’s style for use in mental functions further downstream. In a third set of experiments, we explore how style affects memory, by asking whether the mind extrapolates styles we have seen to anticipate the appearance of completely unseen objects.

These experiments were inspired by semantic priming, the phenomenon whereby semantic processing of one word spreads to other words whose meanings are related (36). While semantic priming effects are often studied as small reaction-time benefits (e.g., in lexical decision tasks), they may also manifest in false memories of words one has not actually seen. For example, after reading the words ‘candy’, ‘taste’, and ‘treat’, subjects may misremember having seen the word ‘sweet’. Might a similar phenomenon arise in the perception of style?

Our next experiments adapted this task to style perception by exploring false memories for members of sets of styled objects (Figure 5). Instead of the synthetic images used in previous experiments, here we used naturalistic images of utensils (i.e., forks, knives, and spoons) to ask whether the mind generates representations of new objects in a given style after previously seeing examples of other objects in that same style (see also 28, who explore transfer of letter identities to new typefaces). By extending our approach to naturalistic stimuli (i.e., beyond the artificial stimuli in Experiments 1-6), these experiments also served to test the generality of our approach. The styles of cutlery sets — and naturalistic objects more broadly — vary in ways that differ from neural style transfer: While artificial style transfer primarily applies a textural transformation (and preserves, for example, global shape and scene composition), naturally styled objects often vary considerably in shape. In part for this reason, stimuli of this sort have been the subject of considerable scholarly attention in the fields of computer vision and machine learning (37–40).

In Experiment 7, subjects first performed a simple identification task in which they saw a series of utensils, one at a time, and judged whether they were forks, knives, or spoons. Then, subjects performed a recall task in which they saw an array of images (some novel, some shown previously) and had to click the utensils they had remembered seeing. Each subject was randomly assigned a 'recall utensil' (either a fork, knife, or spoon) which determined the utensil in the recall task. (At the start of this experiment and Experiment 8, subjects were explicitly informed that their memory for the utensils would be tested; this was not the case for Experiment 9.) Crucially, there were three types of images in the recall task: (1) Images that were seen previously in the identification task ('seen'); (2) images that were not previously seen ('unseen'); (3) images that were not themselves previously seen but that appeared in a style that was previously seen (e.g., a medieval knife, having previously seen a medieval fork and spoon; 'extrapolated'). We suspected that, just as subjects who see 'candy' and 'taste' misremember having seen 'sweet', subjects who see a fork and a spoon in a given style would misremember having seen the knife in that style — a behavior that draws on the capacity to anticipate what a knife from that style would look like, despite never having seen it before.

Indeed, subjects falsely remembered seeing the recall utensil for 'extrapolated' utensils significantly more often than for 'unseen' utensils (mean difference = 29.87%, $t(74) = 11.43$, $p < 0.001$, $d = 1.32$, 95% CI = [24.66%, 35.07%]). This was not just due to poor memory for all utensils, as subjects also successfully remembered 'seen' utensils at a higher rate than 'extrapolated' utensils (mean difference = 36.00%, $t(74) = 15.48$, $p < 0.001$, $d = 1.79$, 95% CI = [31.37%, 40.63%]). This result provides initial evidence for style extrapolation: To reliably select extrapolated utensils (e.g., the medieval knife) more often than unseen utensils suggests that false memories for styled images arise in ways analogous to other established memory phenomena. Moreover, it also indicates that the mind generalizes the styles it learns to novel instances; in order to recognize an object as a medieval knife, one must have abstracted that style from one or more seen examples (here, the medieval fork and spoon) to this new instance (see also *28*).

An important confound in Experiment 7 is image similarity: Because medieval knives look more similar to medieval forks and spoons than they do to other images, false memories for those objects could arise from that similarity alone (rather than from an internal model of style applied to novel objects). Experiment 8 addressed this confound by presenting subjects with either one unique or two unique examples of each style in the identification task, and comparing performance between these two cases. Subjects performed the same task as in Experiment 7, but here the trials were split as follows: For half of the styles shown, two unique examples were seen in the identification task (e.g., a medieval fork and a medieval spoon); for the other half of styles shown, one unique example was seen twice (e.g., two medieval forks). Each set was also split evenly across the 'held out' utensil (i.e., of the two-example styles, one third depicted a fork and a knife, one third depicted a fork and a spoon, and one third depicted a knife and a spoon). If style extrapolation is merely explained by image similarity, then styles containing two unique examples should behave similarly to styles containing one unique example shown twice, because the presented images would be equally similar to the held-out utensil in both cases. However, this is not what we observed; instead, there were higher rates of false memories (i.e., higher rates of generating the held out utensil) for styles containing two unique examples than styles containing one unique example shown twice (mean difference = 5.04%, $t(89) = 2.71$, $p < 0.01$, $d = 0.29$, 95% CI = [1.34%, 8.73%]). Evidently, exposure to different instances of the same style aids in extracting common stylistic features and applying them to novel cases — as would be expected if style extrapolation goes beyond image similarity.

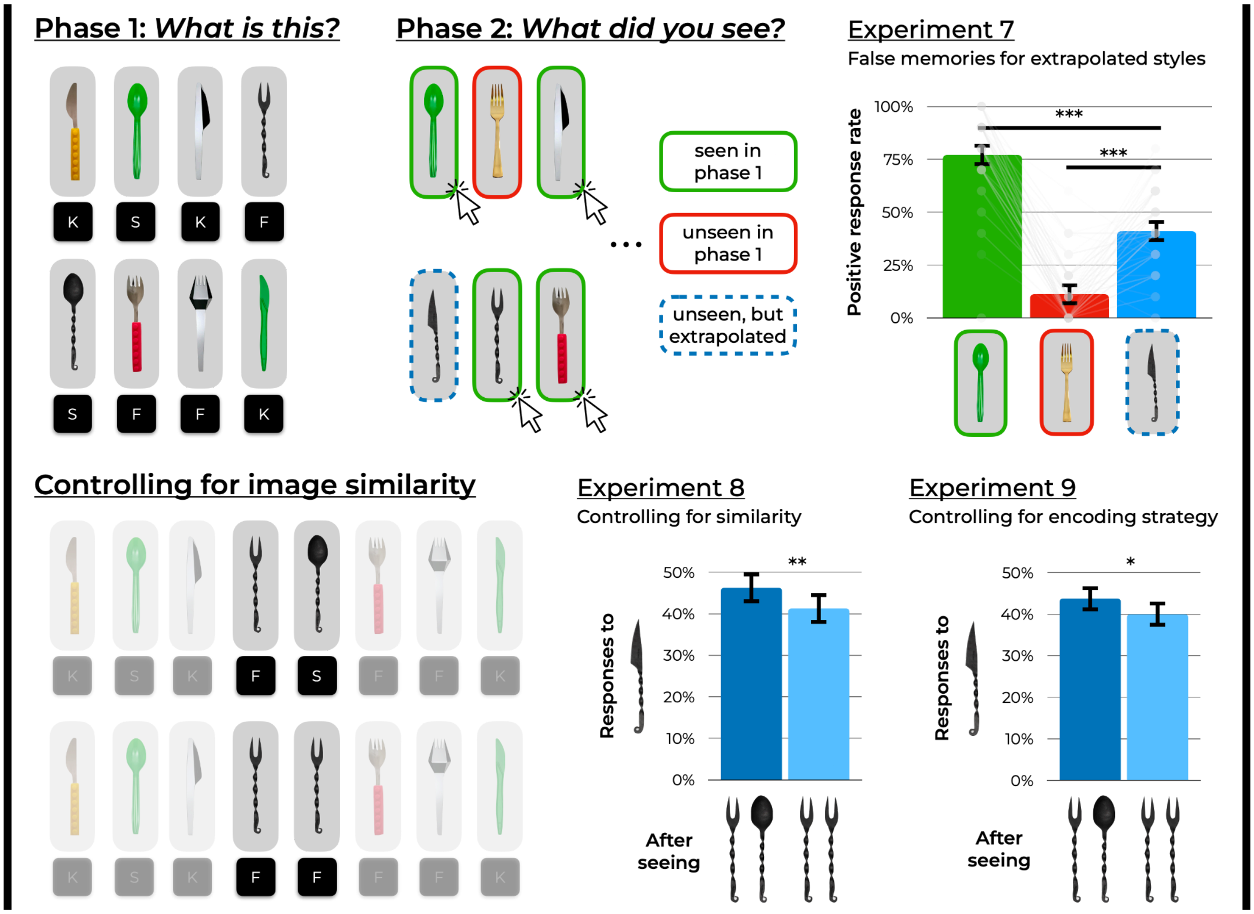

Figure 5. Style extrapolation. In Experiments 7–9, subjects first identified utensils, classifying them as forks, knives, or spoons; then, they completed a recall task in which they selected the utensils they remembered seeing. Experiment 7 (N = 75) revealed false memories for objects that hadn't themselves appeared earlier if other members of that cutlery set had appeared — suggesting that the mind was able to extrapolate the style of seen utensils to anticipate the appearance of other objects in that style. Experiment 8 (N = 90) controlled for image similarity by presenting either two unique utensils of a given style (e.g., fork and spoon from one cutlery set), or two of the same utensils from that style (e.g., two forks from that cutlery set). Even though the presented objects and the lures (e.g., the knife from that set) were now equated across conditions, subjects nevertheless had more false memories after seeing two unique utensils than two of the same utensils, further implicating a process whereby the mind generates representations of new objects in the style of previously seen ones. Experiment 9 (N = 92) showed that such generation may occur automatically, since the effects arise even when the recall task comes as a surprise to participants. Data are presented as means +/- 95% confidence intervals of the difference between conditions. All statistical tests are paired, two-sided t-tests; *** = $p < .001$; ** = $p < .01$; * = $p < .05$.

Finally, we explored whether style extrapolation might occur implicitly or unintentionally (perhaps as a strategy during encoding) by making the recall task a surprise. Whereas Experiments 7 and 8 alerted subjects to the upcoming recall task at the start of the experiment, Experiment 9 presented the identification task without any further context, and then surprised subjects with the recall task. Even in this case, participants extrapolated styles containing two unique examples more than styles containing two of the same examples (mean difference = 3.70%, $t(91) = 2.43$, $p = 0.02$, $d = 0.25$, 95% CI = [0.68%, 6.71%]). Thus, style extrapolation occurs spontaneously; even when there is no independent pressure to do so, the mind implicitly generates representations of unseen objects in the style of previously encountered ones.

Experiment 10: Modeling style perception with CNN embeddings

Our results thus far suggest that the mind adapts to, sees through, and extrapolates style. However, we also judge style more explicitly than in the above contexts. For example, we may appreciate that a certain Van Gogh painting is more stylistically similar to a Monet painting than a Pollock painting (and incorporate these judgments into decisions about which paintings to hang where in a gallery or collection). How systematic and predictable are such judgments, and what is their relationship to more basic mechanisms of visual perception?

In a final experiment, subjects saw two style-transferred images on each trial and simply rated their similarity on a 9-point scale (Figure 6). Each pair of images consisted of one scene type (e.g., a mountain) depicted in two different styles (e.g., Van Goh and Monet). We modeled these judgments by extracting ResNet embeddings (i.e., the network's final layer before classification) for each image in our style-transfer stimulus set and then reducing them from 512 dimensions to 2 dimensions using t-distributed Stochastic Neighbor Embedding (t-SNE; 41). This allowed for visualization of the embeddings, following similar approaches elsewhere in this literature (e.g., separating human sketches by style or drawing pattern; 42). Even before incorporating human judgments, note that t-SNE creates large, well-defined clusters that are naturally separated by style. This may be expected given earlier work, including results from Karayev et al. (32), who demonstrate that ImageNet models implicitly learn features that can be used to classify style, as well as demonstrations of the salience of image style in neural network models (43, 44).

More relevant to our research question is how these similarities track human judgments. While it is expected that the embeddings of style-transferred images should naturally cluster by style rather than scene type (given the training objective of style transfer), it is not obvious that the distance *between* these embeddings would match human judgments in any particularly robust way. However, we found that mean t-SNE distance for all image pairs grouped by similarity rating (1 to 9) decreased monotonically (indeed, perfectly so; $\rho = -1.0$, $p < 0.001$); in other words, images that were rated as less similar by subjects consistently had more distant t-SNE embeddings. (And this relationship was stronger for t-SNE embedding distance between images than for mean squared error in pixel values between images; $\rho = -0.85$, $p < .01$). While there was already good reason to expect most of these results, together they (a) dovetail with demonstrations showing that artistic style may be salient to computer vision models (32, 43, 44), (b) show that these models' representations of style track with human judgments of similarity across styles, and (c) demonstrate that style not only drives performance on the implicit tasks explored earlier but also grounds explicit similarity judgments. (Note that style transfer models may not encompass *all* aspects of stylistic variation. For example, such models capture aspects of texture and color, but not composition and framing. See our General Discussion for more detail on future directions exploring other models and approaches to style transfer.)

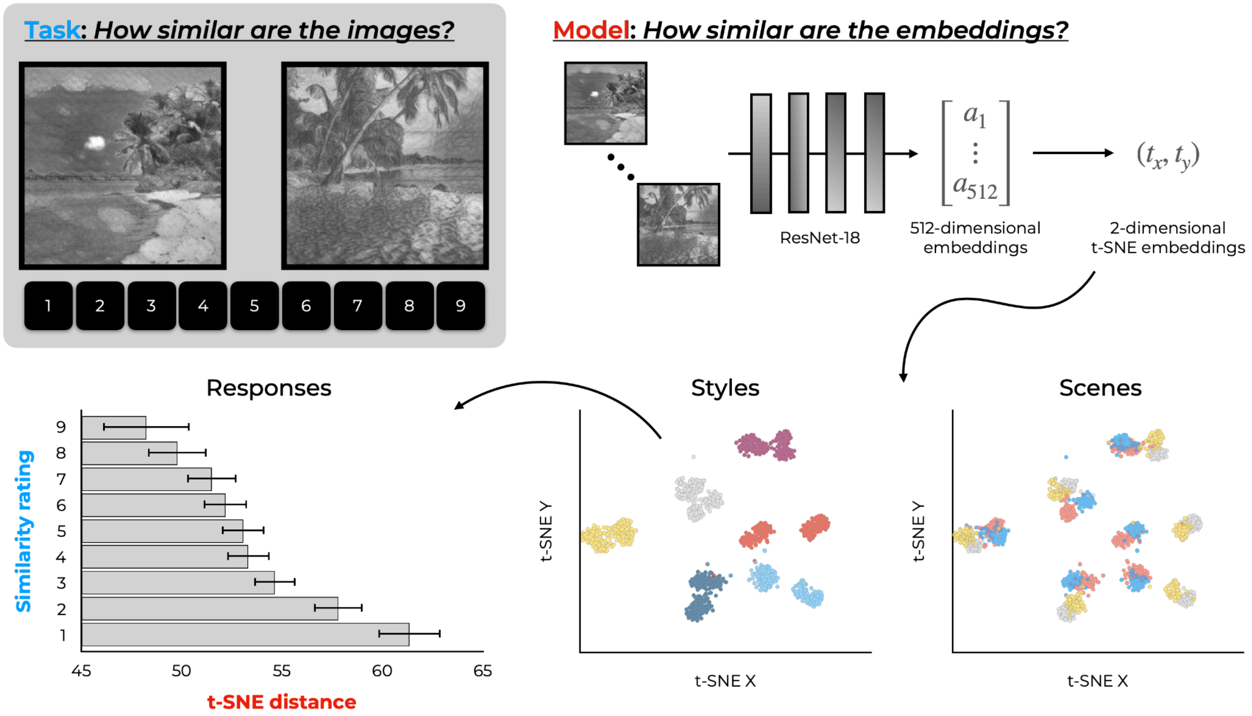

Figure 6. Modeling style perception with CNN embeddings. In Experiment 10 (N = 95), subjects rated the similarity of pairs of styled images. We extracted ResNet-18 embeddings for each image, which we then reduced to 2 dimensions using t-SNE. Notice that the clusters produced by the embeddings are naturally grouped by style rather than scene type. These embeddings also map on to human judgments; similarity judgments decrease monotonically as t-SNE distance between the judged images increases, both for each response and for each style pair. Data are presented as means +/- 95% confidence intervals for each group.

More generally, these results may guide future behavioral experiments on style perception akin to Experiments 1–9. For example, one could further explore the abstractness of style tuning; would style tuning be observed for impressionist-style images broadly, and not just Monet or Van Gogh paintings? Our results could offer a helpful guide for this experiment by informing the selection of styles that are appropriately close together (or far away) in the embedding space. In other words, cognitive questions concerning style perception (e.g., how abstract is style tuning?) can be usefully grounded in the sort of computational approach taken here (e.g., how close do the embeddings of two styles have to be in order to observe style tuning for both?).

Discussion

What is the psychological basis of our capacity to perceive style? The results reported here explore how well-characterized cognitive mechanisms in which the mind parses ‘content’ from ‘form’ underlie this ability and leave psychophysical traces of their operation. This approach revealed several new phenomena of style perception that share key signatures with these other parsing processes. Experiments 1–4 revealed style tuning, whereby observers adapt to the style of scenes, leading to increased processing fluency akin to font tuning and speech adaptation. Experiments 5–6 demonstrated style discounting, a process in which vision ‘sees through’ the style of a scene to discern its underlying content. Experiments 7–9 explored downstream effects of style perception, through style extrapolation — a phenomenon in which perceived style is used to mentally render new objects (creating false memories of having seen them). Finally, Experiment 10 demonstrated that subjective impressions of style are captured by computer vision models in ways that could ground future behavioral experiments.

Our findings join a growing empirical literature at the intersection of visual art, perceptual psychology, and computational aesthetics (6-17, 32, 37-40, 43, 44). While this literature has shed light on questions related to those we explore here (see also 18-21, 28, 45-47), uncovering the nature of style perception has remained an elusive goal — in part owing to the lack of suitable tools to study it (e.g., methods for generating well-controlled styled images). By combining classic psychophysical approaches with recent advances in generative artificial intelligence, our work helps to elucidate this process, showing how style perception can arise from core psychological mechanisms for parsing the content of an image from its form.

The present work focuses on cases where the distinction between style and content seems natural and intuitive. Of course, it is not always trivial (or even sensible) to separate style from content in this way: In many artistic contexts, style is content in an important sense. And even in the design of a home or a tool, stylistic choices may carry functional consequences. Nevertheless, the existence of such examples need not detract from the cases where style and content separate more cleanly, as in the phenomena we explore and investigate here.

These results open the door to further experimental investigations of style perception. A natural extension would be to examine style perception in other modalities. For example, just as the same village scene may be depicted in different visual styles, so too can the same melody or chord progression be realized in different auditory styles. Musical style arguably respects the same distinction between ‘content’ — the underlying melody, perhaps as expressed in sheet music — and ‘form’ — the instruments used to play the melody, the character and emotion with which the instruments are played, the musical genre they belong to, and so on (see, for example, 48). We speculate that musical style might engage the same types of parsing processes in the mind, leading to the prediction that auditory style perception would share many of the signatures we explore here (e.g., listeners ‘tuning’ to the style of an orchestral melody in ways that improve detection of tempo changes or errant notes).

Within the domain of visual style, further advances may be made by loosening various constraints on style-transfer approaches to produce images with more variability. Our experiments employed leading neural style transfer models (26, 27) that render a scene in a given style while preserving its underlying composition. Thus, “style” in such models generally consists of changes in texture, color, and other lower- and mid-level image features. However, artistic styles may also diverge in even higher-level ways. For example, a different artist painting the same scene may choose to vary which objects are present in the first place, where they are located, what viewpoint they are seen from, and so on. Leading style-transfer models do not permit such variation; this is why they were an appropriate choice for the controlled psychophysical setting of our behavioral experiments, which were designed to vary style while holding scene content constant. However, more recent approaches to image synthesis (such as diffusion models) could capture these additional aspects of artistic style (e.g., 49), opening the door for new questions — but also new methodological challenges — concerning the perception and representation of artistic style.

More generally, our work here shows how seemingly abstract or rarefied questions about human creativity and expression may be bound up with more basic psychological capacities — and how quantitative and experimental approaches can complement more qualitative or humanistic traditions to shed light on questions of interest to both.

Methods

General Methods (All Experiments)

Readers can experience all of our experiments, in the same way as our participants did, at https://perceptionresearch.org/style. All sample sizes, designs, and analysis plans were pre-registered; these pre-registrations, along with the stimuli, experimental code, data and analysis code are available at https://osf.io/mb3nh/.

Subjects. All subjects were adults recruited from the online platform Prolific (for a discussion of the reliability of this subject pool, see 50). We coded each experiment using Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript (JS), and then posted our web experiments on Prolific for subjects to complete. Experiments 1–4 and each recruited 50 subjects each; Experiments 5-10 recruited 100 subjects each. Sample sizes were chosen to be sufficiently large based on pilot studies we conducted. All subjects in each experiment were unique; no subject completed multiple experiments. Subjects received financial compensation upon completing the experiment. The experiments were approved by the Homewood Institutional Review Board of Johns Hopkins University (HIRB00005762). All subjects in these experiments provided consent for their participation. In all experiments, subjects who did not submit a complete dataset were excluded.

Style transfer stimuli (Experiments 1–6; 10). We created a stimulus set of artificially-styled scenes using the style-transfer model from (27). We used the model to apply 6 styles from famous paintings (Demuth's Trees and Barns Bermuda, Van Gogh's Starry Night, Klimt's The Kiss, Monet's Water Lilies, Pollock's Number 1, 1949, and Munch's The Scream) to images of 4 different scene types (beaches, bedrooms, libraries, and mountains). We chose these styles because they were prominent and even famous, while also still being sufficiently different from one another. Each scene type contained 64 images from the stimulus set in (29). This resulted in a style-transferred stimulus set of 1,536 images; all these images, along with the code needed to run the style-transfer model, are available in our data archive.

Utensil stimuli (Experiments 7–9). Along with our style-transfer stimuli, we created a stimulus set of naturally styled objects, specifically sets of cutlery. We gathered images of 30 sets of cutlery from various online sources. The sets of cutlery come from a range of styles (e.g., some depict kids cutlery, some depict plastic cutlery, some depict medieval cutlery, etc.).

Experiment 1: Style tuning

Stimuli and Procedure. Participants completed 200 trials of our style tuning task. At the start of the experiment, each participant was randomly assigned a target scene type (either beaches, bedrooms, libraries, or mountains); this determined the scene type they had to count in each trial. Each trial contained an array of 9 images laid out horizontally, with the specific scenes randomly chosen (such that the target scene type was between 1 and 9; distractor scene types were also chosen randomly). Subjects had to count how many times the target scene appeared in the array, and then respond using their keyboard.

Half of trials contained scenes of all the same randomly-chosen style (i.e., for each of 100 same-style trials, a random style was chosen from our set of 6 styles, and then all 9 images appeared in that style). The other half of trials contained scenes of multiple different styles, chosen randomly for each image. The order of same-style and mixed-style trials was shuffled randomly for each participant. Subjects received feedback at the end of each trial.

This study was pre-registered on 31 August 2023 (https://aspredicted.org/kv3v-z8nv.pdf).

Exclusions. We excluded participants who responded correctly on less than 30% of trials. This excluded 6 subjects total. Then, we excluded trials with response times below 200ms or above 10000ms; this excluded 741 of the remaining 8800 trials.

Results. As reported in the main text, subjects were significantly faster at enumerating scenes in same-style trials than in mixed-style trials (mean difference = 289ms, $t(43) = 4.93$, $p < 0.001$, $d = 0.74$, 95% CI = [171ms, 407ms]; this and all other t-tests reported here are two-tailed dependent-samples tests over subject-level means). This analysis was performed only over trials in which participants responded correctly. Subjects were also significantly more accurate on same-style trials than on mixed-style trials (mean difference = 6.74%; $t(43) = 6.10$, $p < 0.001$, $d = 0.92$, 95% CI = [4.51%, 8.97%]; pre-registered as a secondary analysis).

Experiment 2: Style tuning for varying array length

Stimuli and Procedure. The task in Experiment 2 was the same as in Experiment 1 except as noted below. Instead of all trials presenting 9 images, here trials presented different numbers of images. The 200 trials were equally split among arrays of 3 images, 5 images, 7 images, and 9 images. This balance was preserved across same- or mixed-style trials; in other words, of the 100 same-style trials, 25 contained 3 images, 25 contained 5 images, 25 contained 7 images, and 25 contained 9 images.

This study was pre-registered on 10 September 2023 (https://aspredicted.org/fpkz-qpmk.pdf).

Exclusions. We excluded participants who responded correctly on less than 40% of trials. Note that this accuracy criterion is higher than in Experiment 1; this is because the task is easier, as chance performance is higher on trials with fewer images. This excluded 2 subjects. 2 additional subjects were excluded because they did not complete the experiment. As before, we excluded trials with response times below 200ms or above 10000ms, which resulted in excluding 241 of 9200 trials.

Results. Collapsing across all image-array sizes, we observed significantly faster response-times on same-style trials than mixed-style trials (mean difference = 123ms, $t(45) = 3.53$, $p < 0.001$, $d = 0.52$, 95% CI = [53ms, 193ms]). Accuracy was also higher on same-style trials than on mixed-style trials (mean difference = 4.33%, $t(45) = 4.50$, $p < 0.001$, $d = 0.66$, 95% CI = [2.39%, 6.27%]). Finally, we observed either significantly or marginally significantly faster response times for 3-image trials (mean difference = 98ms, $t(45) = 2.53$, $p = 0.02$, $d = 0.37$, 95% CI = [20ms, 176ms]), 5-image trials (mean difference = 108ms, $t(45) = 1.93$, $p = 0.06$, $d = 0.28$, 95% CI = [-5ms, 221ms]), 7-image trials (mean difference = 197ms, $t(45) = 3.15$, $p < 0.01$, $d = 0.46$, 95% CI = [71ms, 323ms]), and 9-image trials (mean difference = 89ms, $t(45) = 1.36$, $p = 0.18$, d

= 0.20, 95% CI = [-42ms, 220ms]). Note that the experiment was not powered to test these specific sub-sample differences, and thus these latter analyses are purely exploratory.

Experiment 3: Style tuning with color and luminance controls

Stimuli and Procedure. This experiment proceeded the same way as Experiment 1; the only difference was in the stimuli. Whereas Experiment 1 presented stimuli which varied in color and luminance (simply using the output of the style-transfer method), Experiment 3 presented grayscale, luminance-matched versions of those stimuli (created using the SHINE toolbox in MATLAB, 31).

This study was pre-registered on 13 September 2023 (https://aspredicted.org/5v4x-rhw2.pdf).

Exclusions. Exclusion criteria here were the same as in Experiment 1: Subjects were excluded if their accuracy was below 30% (which excluded 6 subjects), and trials were excluded for response times below 200ms or above 10000ms (which excluded 721 or 8800 trials).

Results. As in Experiment 1, subjects were significantly faster (mean difference = 267ms, $t(43)$ = 5.61, $p < 0.001$, $d = 0.85$, 95% CI = [171ms, 364ms]) and more accurate (mean difference = 8.26%, $t(43)$ = 7.37, $p < 0.001$, $d = 1.11$, 95% CI = [6.00%, 10.53%]) on same-style trials than mixed-style trials.

Experiment 4: The timecourse of style tuning

Stimuli and Procedure. The stimuli here were the same as in Experiment 1, and, as before, participants were assigned a target scene type at the start of their experiment. However, instead of counting the number of times the target scene appears in an array of 9 images, subjects used their mouse to click on each of the target scene images. When a subject was satisfied and thought they had clicked all the images of the target scene type, they could advance to the next trial by pressing "enter" on their keyboard. At the end of each trial, subjects received feedback, alerting them of both incorrect clicks (i.e., false alarms) and incorrect non-clicks (i.e., misses).

This study was pre-registered on 14 September 2023 (https://aspredicted.org/95g8-cmgv.pdf).

Exclusions. Subjects with accuracy below 40% were excluded (resulting in 4 exclusions). 4 additional subjects were excluded for not completing the experiment. Note that we defined a "correct" trial as one where the subject clicks on all the correct scenes, and only those scenes (and thus has no false alarms or misses). This accuracy criterion was higher than in Experiments 1 and 3 because the responses were slower and more intentional (i.e., participants could unclick and double-check their responses). Trials with a response time below 200ms or above 10000ms were excluded (resulting in 1201 of 8400 trials being excluded).

Results. Total response time — i.e., from the presentation of the images until the subject pressed "enter" — was faster for same-style trials than for mixed-style trials (mean difference = 192ms, $t(41)$ = 5.89, $p < 0.001$, $d = 0.91$, 95% CI = [126ms, 258ms]). Subjects were also more accurate on same-style trials (mean difference = 6.94%, $t(41)$ = 6.93, $p < 0.001$, $d = 1.07$, 95% CI = [4.92%, 8.96%]). Crucially, however, we also analyzed our data by click index to investigate how style tuning evolves over time. We fit a linear mixed-effects model predicting response times, with a random effect of subject and fixed effects of click index (i.e., how many images were previously clicked), trial type (same-style or mixed-style), and their interaction. We found a significant interaction between click index and trial type ($t(711) = 3.25$, $p < 0.01$). As expected, the fixed effect of click index was significant, as it takes more time to click more images ($t(711) = 69.64$, $p < 0.001$), and the fixed effect of trial type was not significant, suggesting that such a tuning advantage indeed evolves over time (i.e., it is not present immediately at the onset of a trial; $t(711) = 0.63$, $p = 0.53$). (Given that these two parameters on their own do not bear on the question of this experiment, we did not pre-register their analysis and merely include them for thoroughness here.) More importantly, we found a significant correlation between click index and same-style advantage on the subject-level ($r(376) = 0.27$, $p < 0.001$, 95% CI = [0.17, 0.36]), such that the same-style advantage increased as click index increased.

Experiment 5: Style discounting

Stimuli and Procedure. We used the luminance-matched stimuli from Experiment 3 in a same-different task to examine style discounting. On each trial, a base image appeared for 850ms, followed by a blank screen for 750ms, followed by a new image, which stayed visible until response. The two images appeared in random locations, and also with random rotations for each trial; the rotations were introduced to make the task more difficult so that subjects would make errors (which are the targets of our analyses). Subjects had to say whether the two images were the same or different (irrespective of rotation; i.e., an image and its 90-degree-rotated version are the same image for this purpose). They made this response using their keyboard ("S" for same, "D" for different).

Participants completed 100 trials of this task. In 50 of the trials, the two images were the same, and in 50 of the trials the two images were different. Of the 50 trials depicting two different images, 25 depicted the same scene as the base image, but in a different style (style-change trials); and 25 depicted a different scene from the base image, but in the same style (scene-change trials). The order of trials was randomly shuffled for each participant.

This study was pre-registered on 15 November 2023 (https://aspredicted.org/c6b9-4jh8.pdf).

Exclusions. We excluded subjects who did not perform accurately on at least 75% of trials, resulting in 11 exclusions. 2 additional subjects were excluded for failing to complete the experiment.

Results. Subjects were more accurate on scene-change trials than style-change trials (mean difference = 16.83%, $t(86) = 11.30$, $p < 0.001$, $d = 1.21$, 95% CI = [13.87%, 19.79%]). They were also numerically faster on scene-change trials, though this trend was not significant (mean difference = 17ms, $t(86) = 1.45$, $p = 0.15$, $d = 0.16$, 95% CI = [-6ms, 41ms]). Additionally, we report an unbiased measure of sensitivity, d'. This revealed high sensitivity for scene-change trials over style-change trials (mean subject $d' = 1.88$, 95% CI = [1.75, 2.01]), confirming our accuracy-based metrics. Note that this d' analysis was not pre-registered, and thus is purely exploratory; however, it suggests that our results still hold with unbiased measures.

Experiment 6: Style discounting with balanced metrics

Stimuli and Procedure. Participants performed the same task from Experiment 5. However, we sub-sampled our stimulus set to contain only the 25 most-similar pairs of scene-change images, and the 25-least similar pairs of style-change images. This resulted in a stimulus set where MSE of the pixel values, structural similarity, and ResNet embedding distance all rated scene-change pairs as more-similar than style-change pairs. The image metrics calculated for all relevant pairs of images are available in our OSF repository.

This study was pre-registered on 20 November 2023 (https://aspredicted.org/43pc-xt3d.pdf).

Exclusions. As in Experiment 5, we excluded subjects who responded correctly on less than 75% of trials, resulting in 10 exclusions. 1 additional subject was excluded for failing to complete the experiment.

Results. As before, subjects were more accurate on scene-change trials (mean difference = 12.90%, $t(88) = 9.44$, $p < 0.001$, $d = 1.00$, 95% CI = [10.18%, 15.61%]). Subjects were also faster on scene-change trials (mean difference = 55ms, $t(88) = 3.53$, $p < 0.001$, $d = 0.37$, 95% CI = [24ms, 87ms]). As above, these results were confirmed by an exploratory analysis using d', an unbiased measure (mean subject d' for scene-change vs. style-change trials = 2.42, 95% CI = [2.28, 2.56]).

Experiment 7: Style extrapolation

Stimuli and Procedure. The experiments for style extrapolation use the utensil stimuli described above.

This experiment consisted of two parts: an identification task and a recall task. In the identification task, participants saw images of utensils, one at a time, and simply classified them as forks, knives, or spoons by pressing “F”, “K”, or “S” on their keyboards. After completing the identification task, participants were then given the recall task, which consisted of multiple trials each displaying a 3x2 grid of utensil images. (Participants were explicitly told that a recall task would follow the identification task.) Each trial of the recall task had the same utensil-types — i.e., all forks, all knives, or all spoons, randomly decided for each subject. (We refer to this utensil-type below as the ‘recall’ utensil.) Subjects were simply instructed to click on the images they had remembered seeing in the identification task and leave unclicked any images they hadn’t seen.

The images appearing in each of these tasks were chosen very intentionally, and as follows (with these choices and parameters not revealed to participants). Of the 30 styles in our stimulus set, 10 were chosen to be left out of the identification task (we refer to these as the ‘unseen’ styles). Of the remaining 20 styles, 10 contained two instances of the recall utensil in the identification task (‘seen’ styles), and 10 contained two examples of the two non-recall utensils in the identification task (‘extrapolated’ styles). For example, if the recall utensil for a given participant was a knife, then that participant would, in their identification task, see: 0 images of any kind from styles 1-10; 20 knives from styles 11-20 (10 knives, each appearing twice); and one fork and one spoon each from styles 21-30.

Then, in that participant’s recall task, they would see 30 images of knives, one from each style. Of those 30 images of knives, 10 would have actually appeared in the identification task (‘seen’);

10 would have came from styles that appeared in the identification task as forks and knives, even though the knives from that style did not appear themselves but the recall utensil itself did not appear in that style ('extrapolated'); and 10 would have been genuinely novel, neither having appeared in the identification task themselves, nor any members of their style having appeared in the identification task ('unseen').

The order of the identification trials, the position of the utensils in the recall grid, and the order of the recall trials were all randomized.

This study was pre-registered on 13 November 2023 (https://aspredicted.org/fpsn-yrdj.pdf).

Exclusions. We excluded participants who responded correctly on less than 90% of trials in the identification task, or got less than 66% of trials in the recall task correct. 5 subjects were excluded because of their performance in the identification task, and 19 more were excluded for their performance in the recall task. 1 additional subject was excluded for failure to complete the experiment.

Results. Subjects had a significantly higher false-positive rate for extrapolated images than for unseen images (mean difference = 29.87%, $t(74) = 11.43$, $p < 0.001$, $d = 1.32$, 95% CI = [24.66%, 35.07%]). This positive response rate was in turn significantly lower than the positive responses for the seen images (which in this case, are correct responses; mean difference = 36.00%, $t(74) = 15.48$, $p < 0.001$, $d = 1.79$, 95% CI = [31.37%, 40.63%]).

Experiment 8: Style extrapolation, equating image similarity

Stimuli and Procedure. As in Experiment 7, this task contained two parts. However, the styles were split differently than in that experiment. Here, of the 30 styles, 15 contributed two of the same examples in the identification task (e.g., two forks), and 15 contributed two unique examples (e.g., a fork and a knife). Furthermore, rather than being tested on just one recall utensil, participants in this task were tested on recalling all three utensils. The 15 styles of each kind were then equally split by recall utensil; 5 held out the fork, 5 held out the knife, and 5 held out the spoon. As before, the styles in each condition were randomly chosen for each participant.

Because of these changes in style, the identification task now contained 60 trials (as each of 30 styles contributed two images). Then, the recall task consisted of 10 trials, each displaying 6 images in the grid. Half the images appeared in the identification task, and half were new. As before, the number of new images was randomized in each trial between 1 and 5 in a way that summed to 30 across the 10 trials.

As stated in the main text, our prediction was that participants would be more likely to misremember having seen a given utensil (e.g., the medieval knife) if they had previously seen both other utensils from that style (e.g., the medieval fork and the medieval spoon) than if they had previously seen only one other utensil from that set twice (e.g., the medieval fork twice). This procedure thus equates for image similarity and frequency of exposure to images from a given style, ensuring that false memories of the additional utensil truly reflect style extrapolation.

This study was pre-registered on 18 October 2023 (https://aspredicted.org/p89s-gtkf.pdf).

Exclusions. We excluded participants who scored below 90% on the identification task, or below 50% on the recall task. This recall exclusion criterion was more lenient than in Experiment 8 because this task is harder than the previous one (i.e., the lures are more similar to the previously seen utensils). 1 subject was excluded based on their performance in the identification task, and 7 more were excluded based on their performance in the recall task. 2 additional subjects did not complete the experiment, and were thus excluded.

Results. Subjects had higher rates of false memories for styles containing two unique examples than for styles containing two examples of the same utensil (mean difference = 5.04%, $t(89) = 2.71$, $p < 0.01$, $d = 0.29$, 95% CI = [1.34%, 8.73%]).

Experiment 9: Implicit style extrapolation

Stimuli and Procedure. This task was exactly the same as Experiment 8. The only difference was that, whereas in Experiment 8 subjects were told at the start of the experiment (i.e., before the identification task) that they would later be tested on their memory of the utensils, here subjects were not alerted of the upcoming recall task, such that it came as a surprise.

This study was pre-registered on 17 October 2023 (https://aspredicted.org/6ngy-g3xs.pdf).

Exclusions. Exclusions were the same as in Experiment 8. 3 participants were excluded because of poor performance in the identification task, and 3 more were excluded because of poor performance in the recall task. 2 additional subjects were excluded because they did not complete the experiment.

Results. As in Experiment 8, subjects had more false-positives for styles containing two unique examples than for styles containing two examples of the same utensil (mean difference = 3.70%, $t(91) = 2.43$, $p = 0.02$, $d = 0.25$, 95% CI = [0.68%, 6.71%]).

Experiment 10: Predicting stylistic judgments

Stimuli and Procedure. On each trial of this experiment, participants saw two images (from our grayscale, luminance-matched style-transfer stimuli) and rated how similar they were on a 9-point scale using their keyboard. Participants completed 100 such trials; in each trial, the two images were chosen randomly such that they depict the same scene type (e.g., both depict beaches) but in different styles.

This study was pre-registered on 24 January 2024 (https://aspredicted.org/m439-g7d5.pdf).

Exclusions. We excluded participants who responded with the same number in over 50 trials; this excluded 5 subjects. Then, we excluded trials with a response time below 200ms (which excluded 129 of 9500 remaining trials).

Embeddings. We computed embeddings for each image in the stimulus set as follows: First, we extracted ResNet embeddings for each image (i.e., the final layer of ResNet classification). These embeddings are 512-dimensional; we reduced them to 2 dimensions by first transforming the 512 dimensions into 50 dimensions with PCA, and then transforming those 50 dimensions into 2 with t-SNE (t-distributed Stochastic Neighbor Embedding, 41). (The PCA transformation was done because t-SNE is typically thought to be unstable above 50 dimensions). We then computed Euclidean distance between images in this 2D embedding space. We also calculated the mean squared error (MSE) in pixel values between any two images in the dataset for model comparison.

Note that all correlations regarding the human-response data are Spearman rank-order correlation tests (and not Pearson's correlation tests) because we are concerned with monotonicity and not linearly. Further, we make no assumptions about the distribution of the variables, and thus prefer the non-parametric test. We found that the mean t-SNE distance had a significant monotonic relationship with judged similarity (ρ = -1.0). Meanwhile, the rank-order correlation with MSE was weaker (ρ = -0.85). We also computed the mean response and distances for each pair of styles (of which there are 15); this relationship was ρ = -0.46 for t-SNE distances, and ρ = -0.37 for MSE.

Data Availability: All data and materials are available in our OSF repository here: https://osf.io/mb3nh/.

Code Availability: All code is available in our OSF repository here: https://osf.io/mb3nh/.

Acknowledgments: The authors thank the Johns Hopkins University Perception & Mind Lab for helpful feedback and discussion. This work was supported by NSF BCS #2021053 award to CF, and an NSF Graduate Research Fellowship awarded to TB.

Author Contributions Statement: TB and CF contributed equally to the study's conceptualization, methodology, visualization, writing — original draft, and writing — review & editing. TB served as lead for data curation and formal analysis. CF served as lead for project administration and supervision.

Competing Interests Statement: Authors declare that they have no competing interests.

References

1. Bonnell, V. E. (1998). *Iconography of power: Soviet political posters under Lenin and Stalin* (Vol. 27). Univ of California Press.
2. Zolberg, V. L. (1990). *Constructing a Sociology of the Arts*. Cambridge University Press.
3. Doss, E. (1995). *Benton, Pollock, and the politics of modernism: from regionalism to abstract expressionism*. University of Chicago Press.
4. Bordwell, D. (2002). Intensified continuity visual style in contemporary American film. *Film Quarterly*, *55*(3), 16-28.
5. Callen, A. (2000). *The art of impressionism: Painting technique & the making of modernity*. Yale University Press.
6. Brielmann, A. A., & Pelli, D. G. (2018). Aesthetics. *Current Biology*, *28*(16), R859-R863.
7. Cavanagh, P. (2005). The artist as neuroscientist. *Nature*, *434*(7031), 301-307.
8. Shimamura, A. P., & Palmer, S. E. (Eds.). (2012). *Aesthetic science: Connecting minds, brains, and experience*. OUP USA.
9. Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual Review of Psychology*, *64*, 77-107.
10. Palmer, S. E., & Schloss, K. B. (2010). An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences*, *107*(19), 8877-8882.
11. Brielmann, A. A., & Pelli, D. G. (2017). Beauty requires thought. *Current Biology*, *27*(10), 1506-1513.
12. Brielmann, A. A., & Pelli, D. G. (2019). Intense beauty requires intense pleasure. *Frontiers in Psychology*, *10*, 2420.
13. Chen, Y. C., & Scholl, B. J. (2014). Seeing and liking: Biased perception of ambiguous figures consistent with the "inward bias" in aesthetic preferences. *Psychonomic Bulletin & Review*, *21*, 1444-1451.
14. Chen, Y. C., Chang, A., Rosenberg, M. D., Feng, D., Scholl, B. J., & Trainor, L. J. (2022). "Taste typicality" is a foundational and multi-modal dimension of ordinary aesthetic experience. *Current Biology*, *32*(8), 1837-1842.
15. Chatterjee, A. (2014). *The aesthetic brain: How we evolved to desire beauty and enjoy art*. Oxford University Press, USA.
16. Chatterjee, A., & Vartanian, O. (2014). Neuroaesthetics. *Trends in Cognitive Sciences*, *18*(7), 370-375.
17. Belfi, A. M., Vessel, E. A., Brielmann, A., Isik, A. I., Chatterjee, A., Leder, H., ... & Starr, G. G. (2019). Dynamics of aesthetic experience are reflected in the default-mode network. *NeuroImage*, *188*, 584-597.
18. Augustin, M. D., Leder, H., Hutzler, F., & Carbon, C. C. (2008). Style follows content: On the microgenesis of art perception. *Acta Psychologica*, *128*(1), 127-138.
19. Davis, T. M., & Bainbridge, W. A. (2023). Memory for artwork is predictable. *Proceedings of the National Academy of Sciences*, *120*(28), e2302389120.
20. Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, *95*(4), 489-508.
21. Pelowski, M., Markey, P. S., Lauring, J. O., & Leder, H. (2016). Visualizing the impact of art: An update and comparison of current psychological models of art experience. *Frontiers in Human Neuroscience*, *10*, 160.
22. Ebner, Marc. *Color constancy*. Vol. 7. John Wiley & Sons, 2007.
23. Foster, D. H. (2011). Color constancy. *Vision Research*, *51*(7), 674-700.

24. Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647-3658.
25. Walker, P. (2008). Font tuning: A review and new experimental evidence. *Visual Cognition*, *16*(8), 1022-1058.
26. Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2414-2423).
27. Ghiasi, G., Lee, H., Kudlur, M., Dumoulin, V., & Shlens, J. (2017). Exploring the structure of a real-time, arbitrary neural artistic stylization network. In *Proceedings of the British Machine Vision Conference 2017*. British Machine Vision Association.
28. Tenenbaum, J. B., & Freeman, W. T. (2000). Separating style and content with bilinear models. *Neural computation*, *12*(6), 1247-1283.
29. Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science*, *21*(11), 1551-1556.
30. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).
31. Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: the SHINE toolbox. *Behavior Research Methods*, *42*, 671-684.
32. Karayev, S., Hertzmann, A., Trentacoste, M., Han, H., Winnemoeller, H., Agarwala, A., & Darrell, T. (2014). Recognizing Image Style. In *Proceedings of the British Machine Vision Conference 2014*. British Machine Vision Association.
33. Wong, K. W., Bi, W., Soltani, A. A., Yildirim, I., & Scholl, B. J. (2023). Seeing soft materials draped over objects: A case study of intuitive physics in perception, attention, and memory. *Psychological Science*, *34*(1), 111-119.
34. Phillips, F., & Fleming, R. W. (2020). The Veiled Virgin illustrates visual segmentation of shape by cause. *Proceedings of the National Academy of Sciences*, *117*(21), 11735-11743.
35. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600-612.
36. McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Psychology Press.
37. Li, H., Zhang, H., Wang, Y., Cao, J., Shamir, A., & Cohen‐Or, D. (2013, September). Curve style analysis in a set of shapes. In *Computer Graphics Forum* (Vol. 32, No. 6, pp. 77-88).
38. Xu, K., Li, H., Zhang, H., Cohen-Or, D., Xiong, Y., & Cheng, Z. Q. (2010). Style-content separation by anisotropic part scales. In *ACM SIGGRAPH Asia 2010 papers* (pp. 1-10).
39. Yin, K., Gao, J., Shugrina, M., Khamis, S., & Fidler, S. (2021). 3dstylenet: Creating 3d shapes with geometric and texture style variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 12456-12465).
40. Chen, Z., Kim, V. G., Fisher, M., Aigerman, N., Zhang, H., & Chaudhuri, S. (2021). Decor-gan: 3d shape detailization by conditional refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15740-15749).
41. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

42. Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects?. *ACM Transactions on Graphics (TOG)*, *31*(4), 1-10.
43. Kim, D. S., Liu, B., Elgammal, A., & Mazzone, M. (2018, January). Finding principal semantics of style in art. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)* (pp. 156-163). IEEE.
44. Saleh, B., & Elgammal, A. (2016). Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature. *International Journal for Digital Art History*, (2).
45. Fan, J. E., Bainbridge, W. A., Chamberlain, R., & Wammes, J. D. (2023). Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, 2(9), 556-568.
46. Ajani, K., Lee, E., Xiong, C., Knaflic, C. N., Kemper, W., & Franconeri, S. (2021). Declutter and focus: Empirically evaluating design guidelines for effective data communication. *IEEE Transactions on Visualization and Computer Graphics*, *28*(10), 3351-3364.
47. Wallraven, C., Fleming, R., Cunningham, D., Rigau, J., Feixas, M., & Sbert, M. (2009). Categorizing art: Comparing humans and computers. *Computers & Graphics*, *33*(4), 484-495.
48. Storino, M., Dalmonte, R., & Baroni, M. (2007). An investigation on the perception of musical style. *Music Perception*, *24*(5), 417-432.
49. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*(2), 3.
50. Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153-163.