A deep-learning framework for human perception of abstract art composition

Pierre Lelièvre

Laboratoire des systèmes perceptifs, Département d'études cognitives Science Arts Création Recherche (EA 7410), Paris, France
École normale supérieure, PSL University, CNRS, Paris, France

Peter Neri

Laboratoire des systémes perceptifs, Département d'études cognitives, Paris, France
École normale supérieure, PSL University, CNRS, Paris, France

Artistic composition (the structural organization of pictorial elements) is often characterized by some basic rules and heuristics, but art history does not offer quantitative tools for segmenting individual elements, measuring their interactions and related operations. To discover whether a metric description of this kind is even possible, we exploit a deep-learning algorithm that attempts to capture the perceptual mechanism underlying composition in humans. We rely on a robust behavioral marker with known relevance to higher-level vision: orientation judgements, that is, telling whether a painting is hung "right-side up." Humans can perform this task, even for abstract paintings. To account for this finding, existing models rely on "meaningful" content or specific image statistics, often in accordance with explicit rules from art theory. Our approach does not commit to any such assumptions/schemes, yet it outperforms previous models and for a larger database, encompassing a wide range of painting styles. Moreover, our model correctly reproduces human performance across several measurements from a new web-based experiment designed to test whole paintings, as well as painting fragments matched to the receptive-field size of different depths in the model. By exploiting this approach, we show that our deep learning model captures relevant characteristics of human orientation perception across styles and granularities. Interestingly, the more abstract the painting, the more our model relies on extended spatial integration of cues, a property supported by deeper layers.

Introduction

Artistic graphical composition can be roughly defined as the structural organization of pictorial elements on a canvas. Art history offers some basic rules and heuristics for understanding the qualitative characteristics of this phenomenon; however, it does not codify processes such as segmentation/interaction of pictorial elements to the degree of specification required by quantitative analysis. Modern artists such as Kandinsky or Klee initiated some systematic and almost scientific studies on this topic (Kandinsky, 1989, 1991; Klee, 1961, 1973, 1998), but they struggled with the combinatorial complexity afforded by compositional questions. Despite more recent progress in this area (Arnheim, 2004), composition remains a complex amalgam of different phenomenona, highly dependent on context and other aspects that are not easily quantified. Composition also represents a versatile experimental tool for empirical aesthetics (Locher et al., 1999; McManus et al., 1993; Schwabe et al., 2018); however, this approach focuses primarily on aesthetic judgements, rather than the compositional processes associated with those judgments.

Recent advances in machine learning, and particularly deep architectures, have demonstrated the ability of artificial neural networks to extract hidden structure from high-dimensional data and solve complex problems with human-level performance (Dodge & Karam, 2017; Serre, 2019). Our goal is to discover whether deep learning tools can advance our understanding of composition and whether, by relying on those tools, we may define a partial, yet relevant, metric description of this phenomenon that is available for quantitative scrutiny (see Iigaya et al., 2020 for related methodology). To achieve this goal, we rely on a well-defined and robust perceptual judgment of visual orientation that is related to composition: telling whether a painting is hung "right-side up."

Citation: Lelièvre, P., & Neri, P. (2021). A deep-learning framework for human perception of abstract art composition. *Journal of Vision*, *21*(5):9, 1–18, https://doi.org/10.1167/jov.21.5.9.

https://doi.org/10.1167/jov.21.5.9 Received April 8, 2020; published May 11, 2021 ISSN 1534-7362

Under the assumption that the orientation of reference for a painting is that selected by the artist, previous work has demonstrated that humans can perform this task well above chance, even for abstract paintings, and regardless of their level of familiarity with painting material (Lindauer, 1969; Mather, 2012). Therefore, it seems that orientation judgments represent a robust behavioral metric, even for material with no recognizable content. Orientations other than the reference orientation may elicit equally valuable subjective interpretations and/or aesthetic experiences in the viewer; however, existing empirical evidence indicates that part of the orientation judgment is consistent across observers: not necessarily directed *toward* the orientation of reference, but at least directed *away* from some of the alternative options. Furthermore, orientation judgments are of immediate relevance to the study of visual perception, an area where image orientation is often manipulated to selectively target higher level processing (see for example the well-known inversion effect (Neri, 2014; Valentine, 1988) and its numerous applications (Cusack et al., 2015; Gaspar et al., 2008; Kelley et al., 2003; Neri et al., 2006, 2007 Yovel & Kanwisher, 2005)).

The exact mechanisms underlying orientation judgements are not fully understood. Some authors have suggested that the perception of orientation depends more on low-level stimulus properties than higher level object recognition and/or image interpretation (Lindauer, 1987), prompting others to investigate the potential role of relatively simple cues, such as Fourier amplitude spectrum slope (Mather, 2012), or image statistics based on explicit rules gathered from several art theories incorporated into a machine learning algorithm (Liu et al., 2017) (see Elgammal et al., 2018; Rodriguez et al., 2018 for related applications).

In approaching these issues, we do not commit to restrictive assumptions or purpose-built schemes. Our model is structured around a general architecture not originally devised for application to art material. We exploit a large database of paintings to train the model, and in so doing we automatically approximate the perceptual mechanisms underlying composition. Despite not being hand-engineered to tailor our specific problem of interest, the trained model outperforms previous applications and extends to a greater variety of painting styles.

It is generally believed that orientation judgments are supported by global analysis of the scene (Oliva & Torralba, 2006). The role of local cues has been relatively unexplored, and more generally the granularity of this phenomenon is not well-understood (Gong et al., 2018). Within the context of our approach, we can naturally probe the issue of granularity and identify the appropriate scale for understanding pictorial elements. More specifically, by exploiting the hierarchical architecture of our model, we can explore how information is represented at different depths within the network. We find that the use of small-scale patterns and deeper level features shows qualitative differences between abstract paintings and more realistic pictorial styles.

To validate the applicability of our model to human visual perception, we carried out a web-based experiment with human observers. They were asked to perform the orientation judgment task on whole paintings as well as fragments of different sizes, corresponding with the different extent covered by the receptive field of distinct depth levels in the model. These experiments were designed with the following goals in mind: establish whether human performance on the orientation task can survive a wider range of stimulus manipulations (painting style, abstraction level, fragment size) than previously tested in the literature; and determine whether our model provides a satisfactory account of the human process. We find positive answers to both questions, although we did identify some discrepancies between human and simulated results, which serve as useful starting points for us to elaborate on how the proposed model may be augmented in future work.

Methods

Database

Our image database is derived from the WikiArt web encyclopedia (WikiArt). The associated API returns metadata such as artist identification and painting styles of each image. At the time of this experiment (May 2019), the WikiArt database contained 157,291 entries. We excluded non-painting styles (e.g., performing arts) and pictures of painting details, reducing this figure to 141,892 items. To make our results directly comparable with those reported by Mather (2012), we manually added 18 entries and moved all paintings from this paper in the validation set. Because our interest is mainly in how model performance varies with style (e.g. abstract vs. figurative), we ensured that different styles and artists were comparably distributed between the training and validation sets. With a target validation ratio of 0.1, the final split is 126,451/15,459. We grouped entries into the genres and styles detailed in Supplementary Tables S1 and S2. Representative examples from this selection are shown in Figure 1. Chosen classification is largely unambiguous, but there are instances for which the specific choice of genre/style may be disputable from historic and/or artistic perspectives. For instance, abstract style is often associated with modern/contemporary Western movements; from such a viewpoint, our decision to

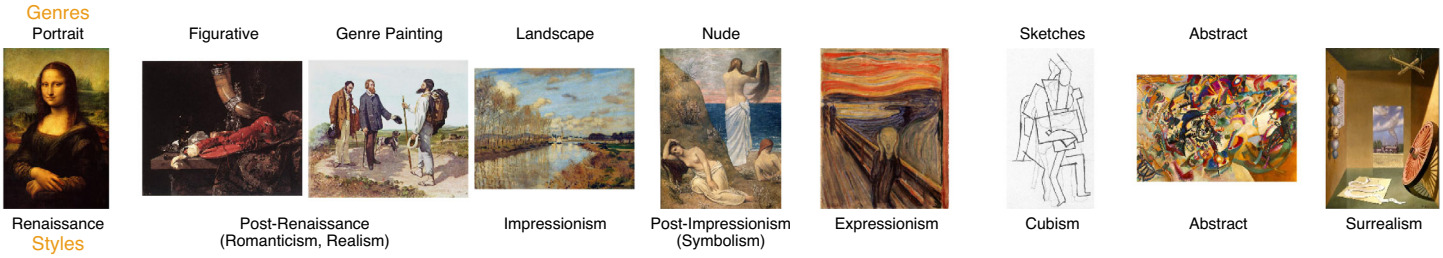

Figure 1. Gallery of genres and styles mentioned throughout the paper. Ordering is chronological. (*Mona Lisa* by *Leonardo da Vinci* (1503-1519), *Still-Life with Drinking-Horn* by *Willem Kalf* (1653), *The Meeting (Bonjour Monsieur Courbet)* by *Gustave Courbet* (1854), *Argenteuil seen from the small arm of the Seine* by *Claude Monet* (1872), *Young Girls on the Edge of the Sea* by *Pierre Puvis de Chavannes* (1879), *The Scream* by *Edvard Munch* (1893), *Seated man with his arms crossed* by *Pablo Picasso* (1915), *Komposition VII* by *Wassily Kandinsky* (1913), *A Naturalist's Study* by *Pierre Roy* (1928)).

include Native art in the abstract category may seem questionable. This decision, however, is motivated by our focus on visual abstraction, rather than abstraction as defined by historical criteria. Furthermore, the questionable instances represent <1% of the total, rendering this issue of little concern. A more probable source of bias is represented by the portrait/landscape aspect ratio. We address this issue in the Supplementary Material, where we demonstrate that this bias is negligible and that the aspect-ratio distribution is well-balanced for abstract paintings, the class we are most interested in.

Model architecture

The task of orienting an image can be thought of as a simple classification problem with four classes, each class corresponding with one possible orientation for the painting. Within this family of machine learning problems, the classification of items from ImageNet (Russakovsky et al., 2015) has led to the development of several deep learning models dedicated to image processing, in particular convolutional neural networks. There is now extensive evidence highlighting similarities between convolutional neural networks and the mammalian visual pathway (Kriegeskorte, 2015; Yamins & DiCarlo, 2016). Among such artificial neural architectures, the most popular are AlexNet (Krizhevsky et al., 2012) and VGG (Simonyan & Zisserman, 2014). Based on its complexity and reported accuracy on ImageNet, we selected VGG-16 (PyTorch implementation; Paszke et al., 2019) as an appropriate starting point for this study.

Figure 2 shows the schematic architecture of our network. All convolutional blocks in gray (1–5) are directly ported from VGG. They consist of multiple convolutional layers with rectified linear units (ReLU) activation functions followed by max-pooling. Our implementation does not use batch-normalization and we removed the original linear layers of the classifier to be replaced by a custom-designed classifier-5, composed of a convolutional layer (kernel size = 7, stride = 3) and linear layers (sizes = [512, 128, 32]). ReLU activation functions and dropout units are applied to all layers except for the last one, to which we applied a softmax function for classification purposes. The dropout rate is of 0.30, except for units before the last layer with a rate of 0.15.

The main feature of our network is that its linear layers are convolutional with kernel size 1. We adopted this formulation to enable inspection of the spatial distribution associated with classified outputs. The consequence on classifier-5 is null because, at this depth in the network, its output (height = 1, width = 1, classes = 4) is generated by a receptive field covering the entire input image. The implication for the other classifiers (1–4), inserted after each convolutional block corresponding to earlier visual areas, is that they have access to small receptive fields. As a consequence, classifier-1 (earliest level) produces for example a classification output of shape (36, 36, 4), as if the network simultaneously judged the orientation of multiple fragments across the picture. This architecture makes it possible for us to inspect network behavior at different depth and for cues of differing granularity.

Training procedure

Input images conform to the VGG format with resolution 224 × 224 pixels and color normalization computed from the ImageNet database. In principle, all parts of a painting may be relevant to judging its orientation, making it inappropriate to crop images into a square shape. We therefore scaled images so that their largest dimension was 224, and fill the remaining empty space with the ImageNet mean value (Figure 3a). These manipulations raise two possible concerns. First, downsampling to a lower resolution may leave out useful orientation cues from the original

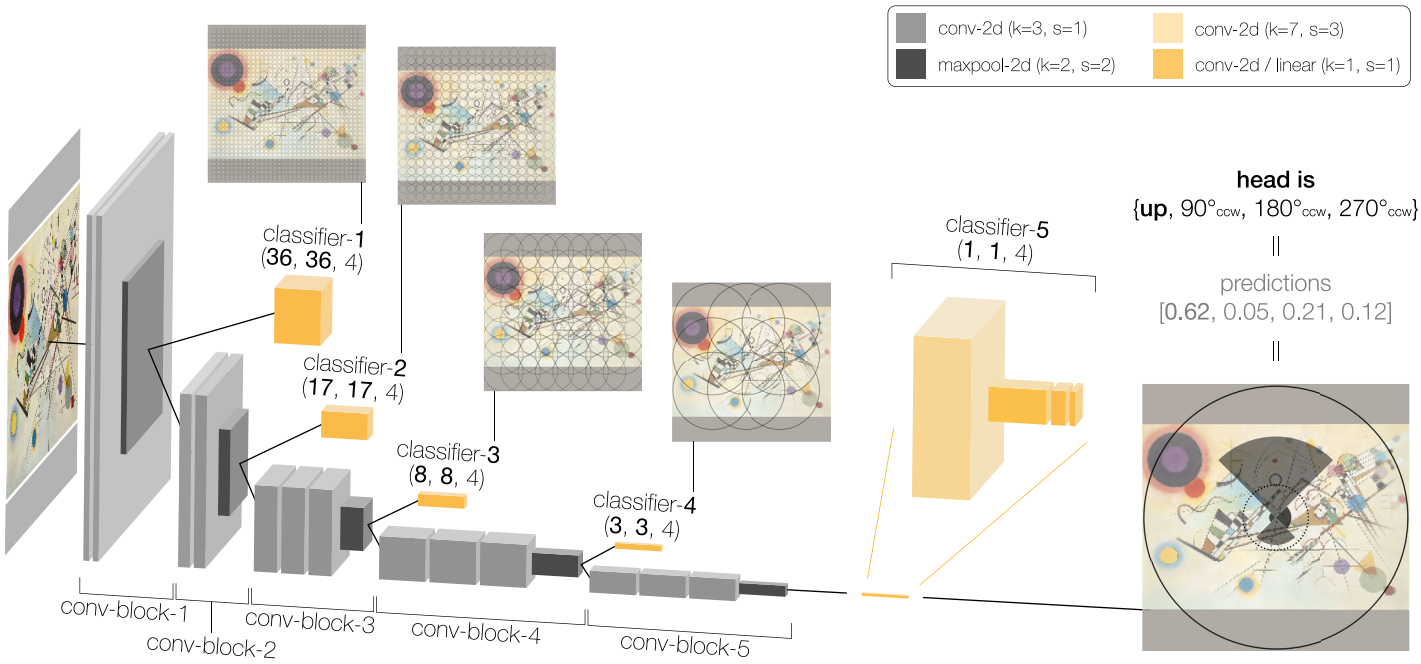

Figure 2. Schematic architecture of the multilevel orientation classification model employed in this study. Each of five convolutional blocks is associated with a classifier (indicated by classifier-*n* with $n = 1$ to 5). The output dimensionality of each classifier is indicated by (*x*, *x*, 4), where *x* is the number of samples across each spatial dimension (see density of circle array within insets overlaying local filters onto painting), and 4 is the number of orientation labels {up,90,180,270}. The four values within [ ] show one example of the categorical distribution generated by the network for *Komposition VIII* by *Wassily Kandinsky* (1923). In the legend, k/s stand for kernel/stride size.

Figure 3. Effect of median filtering on network attention, visualized through guided error back-propagation. Error map is inverted and thresholded for legibility. Light gray indicates pixels where attention reaches at least 1% of its maximum (moderate attention); dark gray indicates pixels where it exceeds 10% (high attention). (**a**) shows original images used for training. (**b**) shows directed attention in the absence of median filtering applied to the borders, (**c**) in the presence of median filtering. Two examples by *Paul Klee* are shown: *The Place of the Twins* (1929) and *After Annealing* (1940).

image. This is possible; however, general considerations about the nature of the images, combined with cursory inspection of representative examples, indicates that composition is a global property that is retained at the adopted resolution. For example, the images shown in Figure 1 are downsampled using the same algorithm we used for the experiments: these paintings are still highly recognizable and understandable. Furthermore, our study is designed as a comparative behavioral experiment between humans and a deep learning model; we expect that the two systems should be similarly impacted by downsampling. The second potential concern relates to color normalization of the paintings. If carried out incorrectly, this procedure may disrupt the perceptual analysis of color and partially alter compositional effects. To avoid this undesirable outcome, we compute mean and standard deviation per channel at the dataset level, not at the level of individual images. Therefore, when normalization is carried out using these mean and standard deviation values, relative color differences and local contrast are conserved at the painting level.

We minimized overfitting using simple data-augmentation techniques: first, we applied random

color gamma correction $output_c = input_c^{\gamma_c}$ where $\gamma_c = 2^{0.5a+0.25b_c}$, a a random scalar and b a random vector sampled from uniform distributions over $[-1, 1]$. Second, images are randomly rotated by up to 5° in either direction and randomly shifted along their shorter dimension within a range such that the whole image remains visible. To accelerate training, we relied on the pretrained model provided by PyTorch. Parameters for the convolutional blocks are not fixed, so they are fine tuned for painting material during training. When filter parameters are fixed, performance is substantially reduced (see Supplementary Material). We used cross-entropy loss for optimisation as is customary in classification problems. Optimization is performed by an Adam algorithm with learning rate $1e^{-4}$ and a scheduler that decreases this learning rate by a factor of 10 when network accuracy remains stable across two epochs.

Testing procedure

At the adopted resolution of the input images, some pictures retained spurious cues to their original orientation, such as artist signatures or handwritten titles near the border. We solved this issue as follows. We initially relied on guided back-propagation to visualize regions emphasized by the model during a preliminary training procedure, and found that the network directed attention to artist signatures and other written characters usually within the bottom region of paintings (Figure 3b). These cues can be trivially exploited to determine picture orientation, but are not connected with composition, so our goal was to remove them as effectively as feasible in automated fashion (manual editing was not an option for such a large database). We applied a median filter with a ramp along all borders of each painting (filter of size 5, full on the outer 5% of the image and with a ramp to zero up to the 20% point). Median filtering is preferable to Gaussian filtering because it removes high-frequency noise while retaining sharp edges. This border-based median-filtering procedure is only applied during validation because it is not useful during training: the network is still able to learn residual artefacts associated with signatures. Figure 3c demonstrates that, even though the network has learned to exploit signatures during training, it successfully reallocates its attention to other parts of the painting when median filtering is applied to borders during validation. Results reported in this article (most importantly validation scores) are averaged separately for each painting over four presentations of that painting in every possible orientation. Model performance refers to average top-1 scores. Top-1 accuracy is 1 if the most probable predicted class is the targeted class, 0 otherwise.

Web-based experiments

We developed a dedicated website for human data collection. Before accessing the experimental platform, participants registered and specified their age as well as their general knowledge of art material. In the first experiment, participants were required to select the original orientation of randomly picked abstract paintings successively presented in blocks of 10. Each painting was presented in isolation and could be oriented interactively by the user; once the participant was satisfied with a particular orientation, this was selected by pressing a button and triggered presentation of the next painting in the sequence. If any element in the painting could serve as obvious hint to the correct orientation, like a word or a signature, people were asked to report it via a dedicated button. After each series, a figurative painting of obvious orientation was inserted into the sequence to check whether participants were meaningfully engaging with the task. To motivate their interest and maintain their focus, participants were provided with feedback at the end of each series detailing performance scores and information about the paintings. In the second experiment, participants saw fragments of both abstract and figurative paintings. The fragments were sized to span the approximate size and location of fragments accessible to the network for each classifier. Under these conditions, the task was perceived as challenging and sometimes puzzling owing to the fragments often being small and blurry; however, it produced interesting results for understanding compositional perception at different granularities. Because we sought to randomly sample paintings from the same style distribution as the model dataset, we excluded categories with a small number of entries to avoid unreliable measurements. More specifically, the abstract category included the following styles (in decreasing order of representation): Abstract Expressionism, Abstract Art, Art Informel, Color Field Painting, Minimalism and Lyrical Abstraction; the figurative category only included Romanticism. We collected an average of 50 trials per participant from 71 participants aged between 15 and 67 and coming from 8 different countries. As an indication that our sample is representative of those commonly used in the literature, our measured average accuracy of 47% (Figure 9b) is highly consistent with values reported by existing studies (Lindauer, 1969; Mather, 2012). We excluded eight participants with scores of less than 0.75 for figurative styles and of less than 0.25 for abstract styles who had typically collected fewer than 10 trials. The inclusion of these participants lowers overall accuracy to 46%, but does not alter the general pattern of the results and their interpretation. We also recorded reaction time, age and general knowledge of art material (as self-reported via questionnaire); these factors are tangential to the present study, so

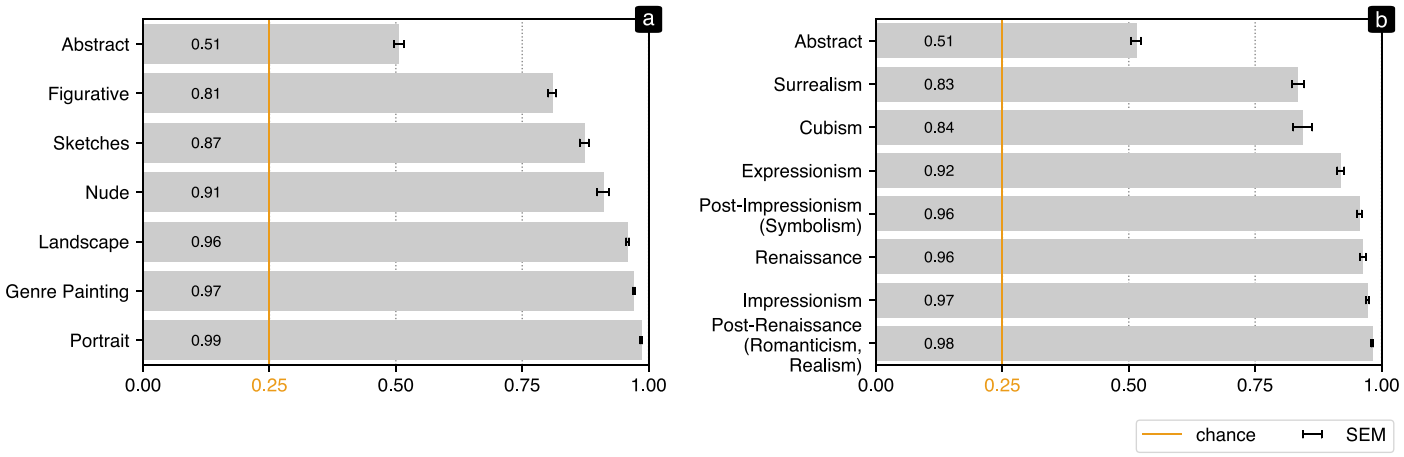

Figure 4. Model performance on whole paintings grouped by genre (**a**) and style (**b**).

they are only briefly discussed in Supplementary Material.

Guided back-propagation

This neural network visualization technique consists of back-propagating the true class/label (binary one-hot distribution) as an error through the network all the way back to the input image. Because the network applies more correction to regions of the input image where information is most useful for achieving categorization of the back-propagated class, those regions map out the equivalent of attentional deployment by the network (Figure 3c, Figure 5b). The *guided* variant of back-propagation was introduced by Springenberg et al. (2015) to improve back-propagation of the gradient through ReLU activation units.

Cross-entropy

Given target probability distribution p and estimated probability distribution q, cross-entropy is defined as $H(p, q) = H(p) + D_{\mathrm{KL}}(p \parallel q)$ where $H(p)$ is the entropy of the target distribution (i.e., the average amount of uncertainty/information about p) and $D_{\mathrm{KL}}(p \parallel q)$ is the Kullback–Leibler divergence from q to p (a measure of the difference between the two distributions). When target distribution p is the final classified label (binary one-hot distribution), $H(p) = 0$ and cross-entropy simplifies to $D_{\mathrm{KL}}(p \parallel q)$; to optimize this function, the model simply pushes the q estimate to match p as closely as possible. We also compute cross-entropy for target distributions other than the final one-hot label; more specifically, we compute distributions for fragments at level n (q in notation above) and measure their predictive power for target distributions of closest fragments at level $n + 1$ (p in notation above). The goal of this between-level metric is to measure redundancy between distributions at different levels. To produce a more interpretable metric in Figure 8, redundancy is defined as $\exp[-H(p, q)]$. The maximum redundancy is 1, corresponding with 0 cross-entropy. A chance level can also be defined as the cross-entropy between equiprobable distributions, simplifying to a redundancy of 0.25 with four classes.

Results

Model performance on whole paintings

Model performance on whole paintings of the abstract genre is around 50% (Figure 4a), in excellent agreement with human measurements from existing literature (Lindauer, 1969; Mather, 2012). Performance also progressively improves from abstract to objects, landscapes through to portraits (Figure 4a). Qualitatively speaking, this progression seems to be related to the characteristics of possible orientation cues, such as their diversity and reliability. For example, Portraits (e.g., *Mona Lisa* in Figure 1) contain faces that are almost exclusively in the upright orientation, making for highly stereotyped and reliable cues. Genre Paintings often display people in standing position, during battles, religious ceremonies or everyday life (e.g., *The Meeting* (*Bonjour Monsieur Courbet*) in Figure 1); cues are still primarily restricted to human characters, but are less stereotyped due to different (potentially conflicting) body poses. Landscapes and Figurative genres display greater diversity of cues, more abundant but certainly less reliable: trees and clouds can be seen via water reflections and objects may not be associated with specific orientations. Along

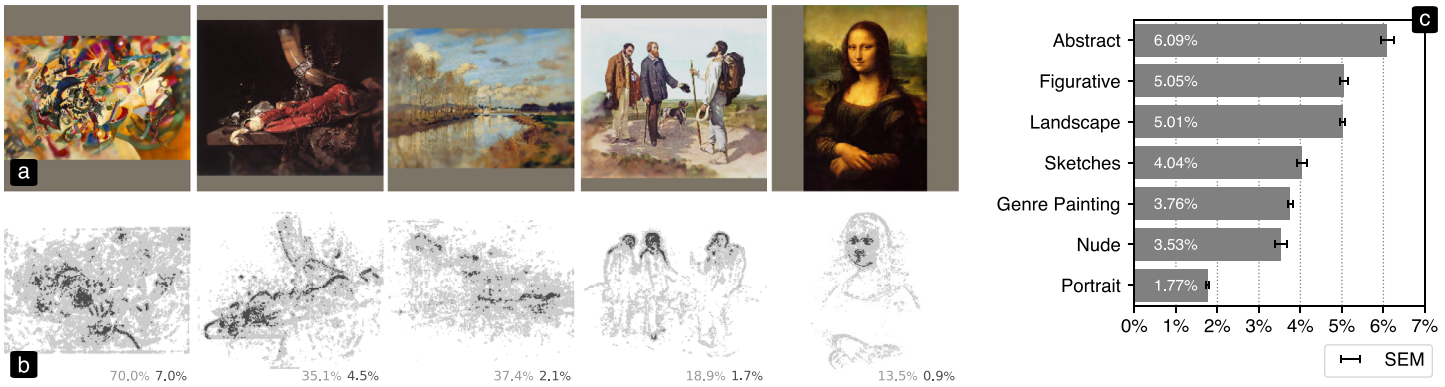

Figure 5. Network attention through guided error back-propagation (see Methods). (**a**) Five examples of original inputs for validation (*Komposition VII* by *Wassily Kandinsky* (1913), *Still-Life with Drinking-Horn* by *Willem Kalf* (1653), *Argenteuil seen from the small arm of the Seine* by *Claude Monet* (1872), *The Meeting (Bonjour Monsieur Courbet)* by *Gustave Courbet* (1854), *Mona Lisa* by *Leonardo da Vinci* (1503-1519)). (**b**) Error maps with inverted and thresholded intensity. Light gray indicates pixels where attention reaches at least 1% of its maximum (moderate attention); dark gray indicates pixels where it exceeds 10% (high attention). Numeric values report light and dark pixel percentages over the entire painting surface. (**c**) Average surface ratio of high attention, plotted separately for different genres.

this qualitative scale, Nude is perhaps the only genre that seems to be misplaced (right before Landscape in Figure 4a), because one may expect that it should be similar to Genre Painting. Looking at *Young Girls on the Edge of the Sea* in Figure 1, Nude paintings seem to explore an extended range of body poses, making body orientation a potentially unreliable cue.

To investigate this interpretation more quantitatively, we can visualize the network's error back-propagation, a technique that exposes regions where the network directs its attention during evaluation. The spatial organization of attentional deployment offers useful insight into the diversity of available cues. Consider *Mona Lisa* in Figure 5b: the most active attentional areas, indicated by dark gray pixels, are highly localized and limited to facial details. In comparison, Kandinsky's *Komposition VII* prompts the model to gather information across the entire image. For Monet's landscape and Kalf's still life, the model operates in a manner that appears to sit halfway between those two extremes, in line with the hypothesis described earlier. We attempt to quantify this trend by simply measuring the proportion of image pixels where the back-propagated attentional signal exceeds 10%. When plotted separately for the different genres (Figure 5c), this quantity is well aligned with the genre ordering of Figure 4a. If we adopt pixel area as a proxy for cue numerosity, the network model uses nearly 3.5 times more cues for Abstract paintings than Portraits. In this ranking, Nude is closer to Genre Painting, as expected from our earlier qualitative considerations. Finally, Sketches may be expected to occupy a position closer to the Abstract genre; however, the sparseness of line content over the flat canvas may explain the lower ratio reported in Figure 5c.

A related concept for ordering model performance on different art material is the reliability/interpretability of available orientation cues, which may reflect the purported importance of "meaningful" content for orientation judgements. From this perspective, painting style (rather than genre) may offer better insight into the role of image content. For example, portraits from Leonardo da Vinci and Picasso (see Figure 1) encompass different degrees of ambiguity. With this notion in mind, Figure 4b demonstrates a lawful relationship between performance and abstraction level (concreteness): from abstract style to Cubism, Symbolism, and post-Renaissance realism. Therefore, taken as a genre or a style, abstraction is in both cases the most difficult material to orientate.

These observations may be summarized by the notion that, although abstract orientation cues are widely distributed across the canvas, they seem to carry limited predictive power. By and large, these visual features are likely employed by artists regardless of their orientation; nonetheless, the associated performance in the orientation judgment task is well above chance. A recent study (Specker et al., 2020) reports that human observers share artistic judgment more effectively in relation to whole abstract artworks as opposed to isolated elements (lines and colors). Therefore, it appears that, in the absence of preferred orientation for individual elements, the only effective source of information must come from the combination of the different cues into specific arrangements that may or may not be represented at the level of the perceptual/neural process. The progressive

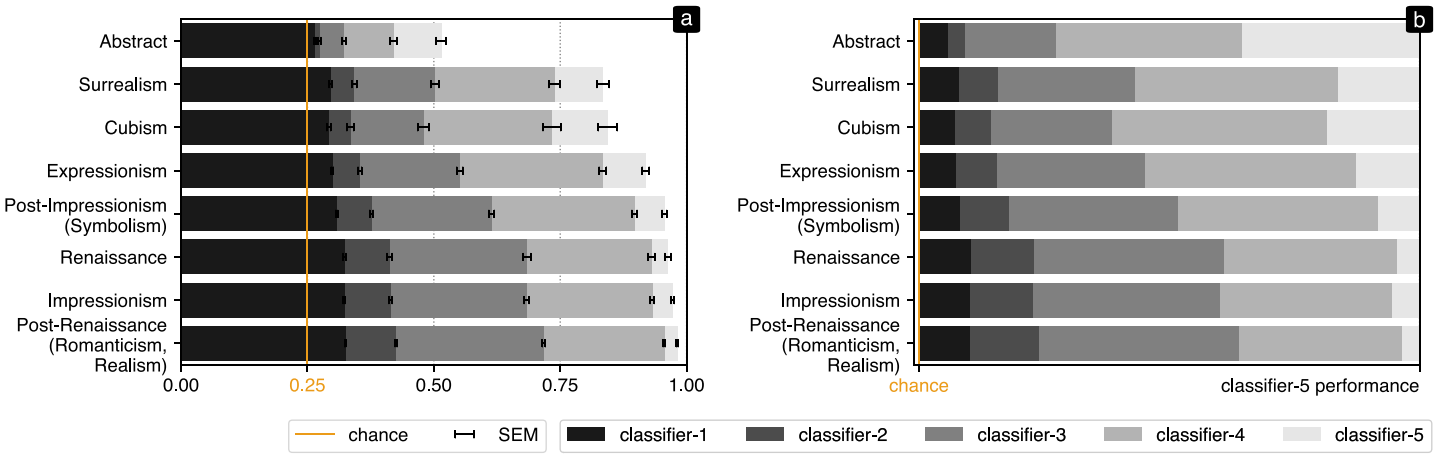

Figure 6. Model performance across classifiers. Values are grouped by style (as in Figure 4b) and displayed separately for the five distinct classifiers. (**b**) plots values from (**a**) after rescaling between chance and maximum value for given style (corresponding to performance of classifier-5).

construction/representation of compositional patterns is a phenomenon of central interest to our study, and one which we hope to understand further by examining the role of local image cues/fragments in greater detail (discussed further in the next Section).

Model performance on fragments

We are in a position to study model behavior for earlier layers via inspection of classifier-1:4 (Figure 6a). Model performance shifts toward chance as its spatial resolution is restricted to smaller receptive fields and could be rephrased as "more is better." We interpret this trend as reflective of the commonly regarded high-level nature of the orientation judgement (Neri, 2014; Valentine, 1988). Perhaps related to this observation, a recent investigation of human aesthetic judgement viewed from a neural network perspective (Iigaya et al., 2020) reports that judgments of "concreteness" become increasingly dominant with neuronal integration. We also find that, when values are normalized by the performance level associated with classifier-5 (Figure 6b), the dependence on deeper layers increases with abstraction level of the painting.

We can gain more insight into the issue of granular representation within the model by plotting predicted orientations from individual receptive field units (Figure 7). The first and most obvious characteristic of these results is that figurative paintings are more spatially redundant than abstract paintings: they offer orientation cues more uniformly spread across the image down to small scales. Further to this observation, although the results for figurative paintings at coarser scales can be roughly predicted from those at finer scales via simple integration of local cues, this rule does not seem applicable to abstract paintings: a large fragment is not reflected by simple averaging of smaller related fragments.

To quantify redundancy between adjacent classifiers, we measure how well distributions at level n describe those at level $n+1$ using rescaled cross-entropy (see Methods). This quantity is plotted in Figure 8; it ranges between chance (level $n+1$ cannot be predicted by level n) and ceiling performance (level $n+1$ can be fully predicted by level n). First, we notice that redundancy increases as we transition from the earlier layers to the later layers, meaning that redundancy increases along the processing pipeline. For example, redundancy between classifiers 1-2 and 2-3 remains near chance across all styles. As we transition to later levels (description of classifier-5 from classifier-4), figurative paintings show a strong correlation between classifiers, while abstract paintings remain close to chance.

A different (but related) way of thinking about Figure 8 is to consider the progressively expanding horizontal bars for figurative paintings as reflecting a gradual emergence of a structured representation that is largely shared across layers. Whatever properties are being represented by the network to support classification, their representation is constructed incrementally along the processing hierarchy and is therefore distributed across layers. In the case of abstract art, representation of relevant properties does not appear to emerge gradually along the pathway. Classifier-5 seems to represent a global property of abstract art that is not transparently available from earlier layers, and which we speculate may be connected with composition. It is true that earlier layers support an appreciable level of task performance (see Figure 6a), but our cross-entropy analysis indicates that

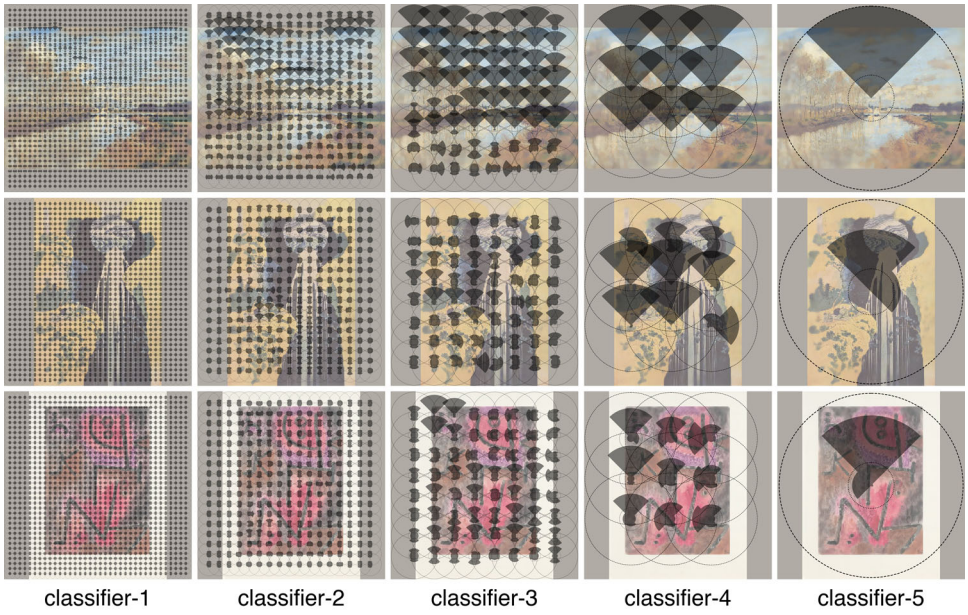

Figure 7. Predicted orientations from individual receptive field units within each classifier. Different classifiers (1–5) are plotted from left to right. Relative size of the four wedges within each circle reflects prediction strength across the four different orientations. Examples are shown for three paintings (dates given when known): *Argenteuil seen from the small arm of the Seine* by *Claude Monet* (1872), *The Waterfall of Amida behind the Kiso Road* by *Katsushika Hokusai*, *After Annealing* by *Paul Klee* (1940).

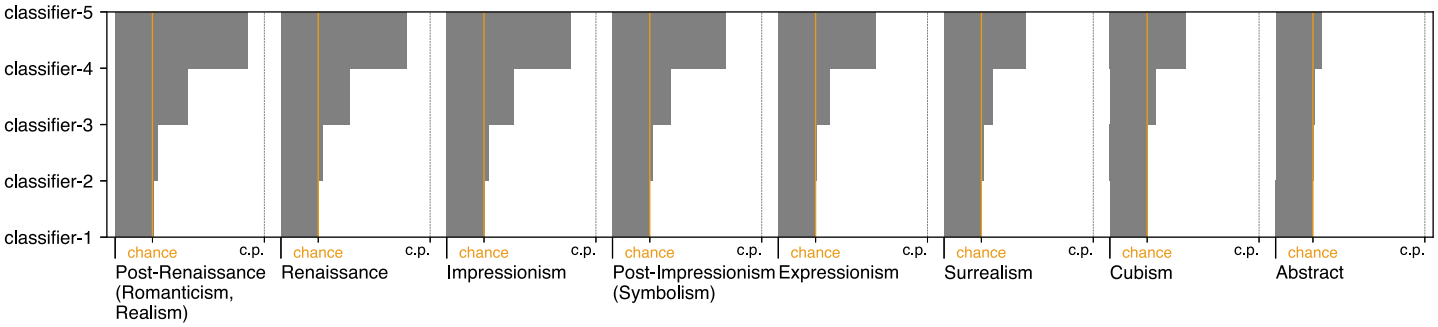

Figure 8. Redundancy between adjacent classifiers, grouped by style. This metric corresponds to rescaled cross-entropy between classifier distributions at level n and those at level $n + 1$ (see Methods). Values are averaged across fragments. Along x axis, c.p. stands for ceiling performance.

this is achieved via representation of other task-relevant properties that do not share characteristics with those represented by classifier-5.

To summarize these results, it appears that abstract art suffers from higher local variability of compositional effects, requiring spatially extended integration of orientation cues for them to cohere into a reliable orientation estimate. Deeper layers must represent emergent global properties that are not necessarily available to previous layers; these properties may be connected with Gestalt principles associated with abstract material, for which the whole is more than the sum of its parts. It is true that we measured performance levels that are relatively low (albeit well above chance), and that this observation alone prompts caution in potentially overstating the universality of this phenomenon; nonetheless, it also implies the existence of a mechanism that is clearly structured to a measurable extent (i.e. stands above chance). Partial, but systematic, neural integration of image features has also been described for other aesthetic judgments (Iigaya et al., 2020). To determine whether these findings are idiosyncratic to our model or, as we hope, they reflect real compositional mechanisms of more general relevance to cognition, we report on human behavioral experiments designed to retain the closest possible connection with the above characterization of the network model.

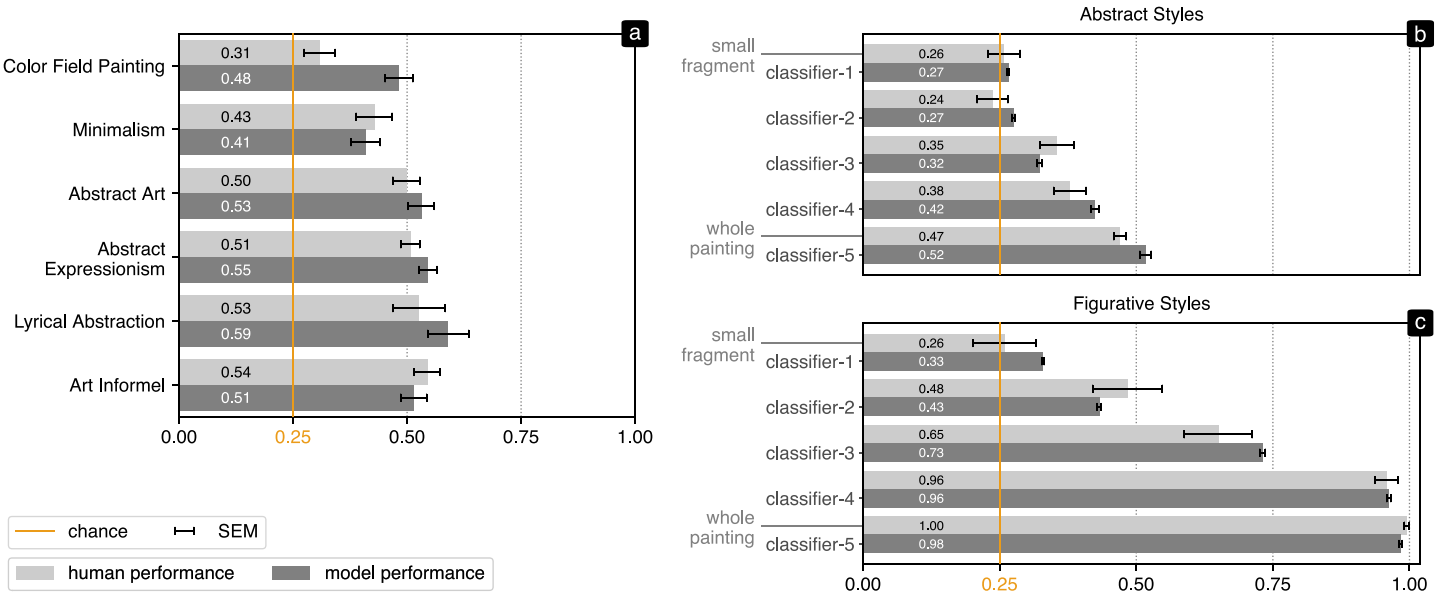

Figure 9. Human versus model performance for whole paintings and fragments. In (**a**), model performance from classifier-5 is plotted alongside human performance on whole paintings (dark versus light bars, respectively), grouped by style. In (**b-c**), model performance from different classifiers (1–5) is plotted alongside human performance on image fragments, separately for abstract (**b**) and figurative styles (**c**).

Human experiments and comparison with network model

Among abstract styles, human observers have the most difficulty determining the orientation of images from Color Field Painting (e.g., Mark Rothko) and Minimalism (e.g., Francois Morellet), as these styles provide less pictorial content and more perfect symmetries. Figure 9a demonstrates that results from the neural network are well-aligned with the corresponding human results (except for Color Field Painting).

The model-human correspondence also exists on a per-classifier basis (Figures 9b, c). For this analysis, we compare model performance from different layers with human performance for different fragment sizes. We emphasize that values for the model are not obtained by presenting the model with fragments (as for example in Rodriguez et al. (2018)): here the model is always presented with full-size images. Different values refer to different classifiers at different depths. We can establish a one-to-one pairing between network layer and fragment size because, when selecting fragment size in the human experiments, the different sizes were tailored to the receptive-field size of different layers within the model. Other than that, there is no obvious connection between model and human results, meaning that it is not trivially expected that values obtained from different network depths should mirror those obtained from human measurements at different fragment sizes.

We find good correspondence between the two sets of results: abstract and figurative styles show the same progression of performance across different fragment/receptive-field sizes ($r^2 = 0.976$ with $p < 0.001$). One implication of this result is that, if we assume that the network model represents an acceptable approximation to the human visual pathway (Kriegeskorte, 2015; Yamins & DiCarlo, 2016), we should be able to probe activity at different levels within the pathway by simply restricting fragment size in a behavioral experiment. Although this result may seem trivial on the surface, it is not to be taken for granted when the output metric is a relatively complex perceptual judgment (see Discussion for more in-depth consideration of these issues). Further experiments using different behavioral tasks would be necessary to confirm/disprove the generality of this result.

Our proposed model is not only able to replicate the extent to which humans produce correct responses, but also specific patterns according to which humans produce incorrect responses. Figure 10 plots normalized frequency of incorrect predictions (three orientations other than the upright orientation of reference) across classifiers (for model in a and b) and fragment size (for humans in c). It is evident that, when incorrect responses are produced, there is a tendency on the part of both model and humans to select the orientation 180° away from the orientation of reference (painting in upside-down configuration) more often than those orthogonal to it. This anisotropic effect applies to all styles for the model (Figure 10a) and is particularly

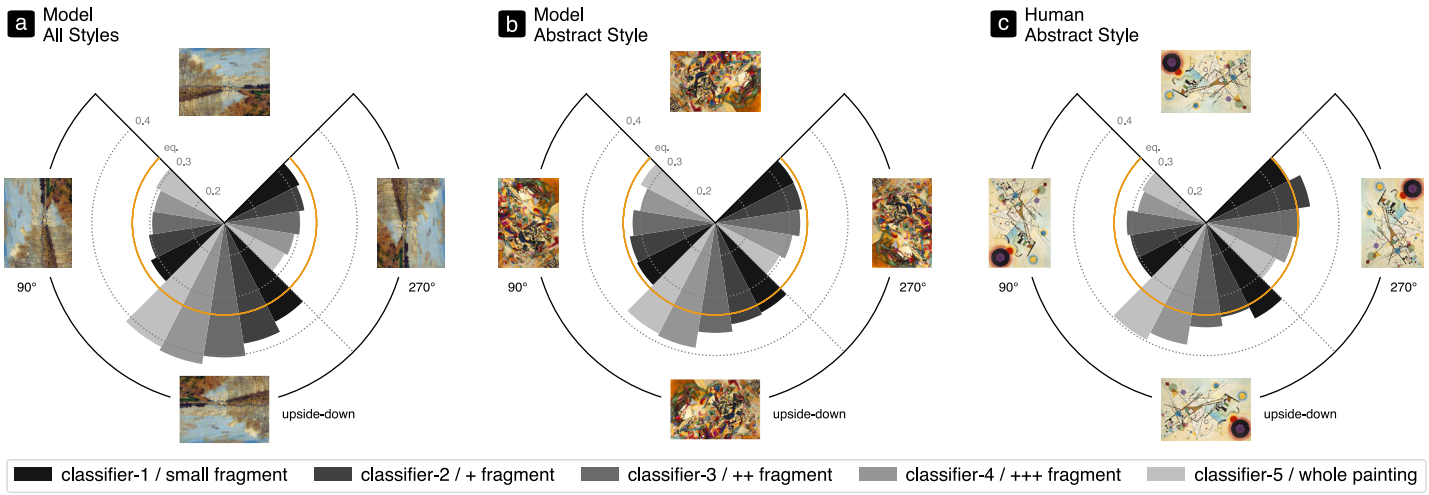

Figure 10. Normalized frequency of incorrectly predicted orientations across classifiers for model, all styles (**a**) and abstract style (**b**); across fragment size for humans, abstract styles (**c**). *eq.* stands for equi-frequency. Examples are shown for three paintings: *Argenteuil seen from the small arm of the Seine* by *Claude Monet* (1872), *Komposition VII* by *Wassily Kandinsky* (1913), *Komposition VIII* by *Wassily Kandinsky* (1923).

pronounced when analysis is restricted to abstract paintings (Figures 10b, c); for this type of art material, model and human behavior are well-correlated across classifiers and fragment sizes ($r^2 = 0.745$ with $p < 0.001$).

The correspondence between human and model behavior for incorrect responses indicates that, in both cases, some image features present horizontal/vertical compositional cues that support alignment of the image along either horizontal or vertical axes, without providing useful information for determining how the image should be mirror-flipped around the chosen axis. Consider, for example, an image containing a mountain reflected against a lake in front of the mountain; clearly, a human observer is able to orient this image so that one mountain is above, and the other one is below. However, if the observer were asked to determine which mountain should be on top and which below, he or she may be unable to make such a determination (in the assumption that the lake produces a nearly perfect reflection of the mountain above it). Similarly, if the observer were asked to determine whether the image should be flipped left-right or not, he or she may be unable to produce an informed answer. Our results indicate that cues of this kind are available from the image database we constructed, and that both model and human are able to exploit them in similar fashion. In Figure 10, the upside-down confusion also seems to be more pronounced for later/larger layers/fragments, suggesting that the horizontal/vertical opposition emerges as a consequence of spatially broad cue integration. On abstract material, across classifiers and fragment sizes, a Cuzick's test (Cuzick, 1985) confirms this trend with $p = 0.012$.

Human/model comparison on a per-painting basis

So far, we have considered the behavior of humans and model without referring to individual paintings. For example, when we say that model performance matches human performance for orienting abstract art, we mean that out of 100 abstract paintings, the model responds on average as correctly as the human observers. This finding does not mean that model and human responses match at the level of individual paintings: the model may be correct for 50 out of 100 paintings, and so may be the human observer, but the 50 paintings for which the model is correct may be those 50 for which the human is incorrect. To address this possibility, below we consider model versus human responses on a per-painting basis.

Figure 11 plots the density distribution of joint orientation choices generated by model and humans for individual abstract paintings. If model and humans were to agree on the orientation of every painting, modulations would only be present within the diagonal bins; all other values should be zero. Because all values must sum to 1 in each plot, we can take the sum of the diagonal values as an indication of model–human agreement (the sum is 1 when model and humans fully agree, 0 when they consistently disagree). The diagonal sum is significantly different from the null prediction only for whole paintings (Figure 11e); when data are plotted for humans orienting smaller fragments and model responses from more superficial layers (Figures 11a–d), agreement decreases to around chance. But how do we assess significance in relation to the statements above?

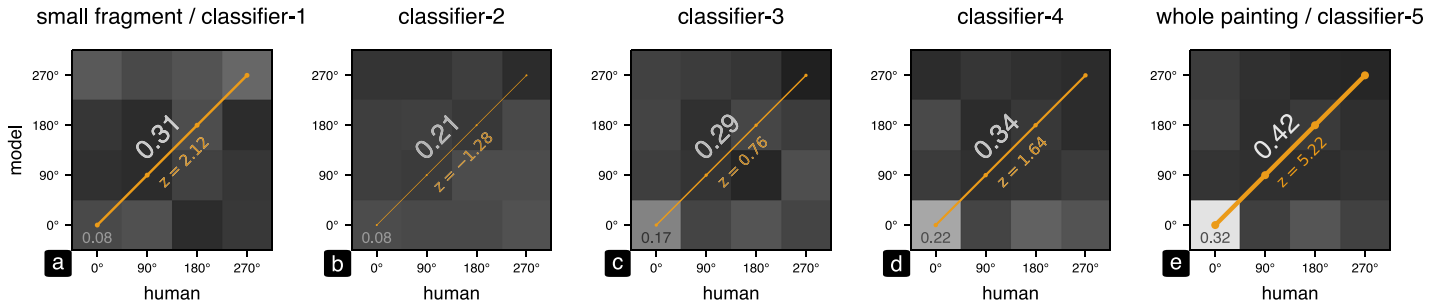

Figure 11. Density distribution of joint orientation choices generated by model and humans for individual abstract paintings, computed separately for different fragment-size/classifier from small/early (**a**) to large/late (**e**). Diagonal values correspond with matching responses (humans and model generate the same response); the diagonal sum (indicated by large white digits) is therefore termed "mutual agreement." Its value is z-scored against the null hypothesis of human/model independence of choices (see main text for clarification). Intensity of white digits and thickness of diagonal orange line scale with corresponding z score. Bottom-left value reports agreement on target orientation.

There are at least two different ways of defining a null hypothesis against which to test significance of the agreement value. The simplest approach is to define the null hypothesis as one where both humans and network respond randomly; in this case, the expected value for each pixel in the 4×4 surface plots of Figure 11 is simply 1/16, and the expected sum across the diagonal is 1/4. Although this approach may be appropriate for evaluating whether humans/models perform above chance, we find that it is inadequate for the purpose of addressing the specific issue we formulated at the beginning of this section. Consider, for example, a scenario in which humans and the model are always correct, regardless of the specific painting that is presented to them; clearly, they are also always in agreement with each other, merely as a consequence of being correct: the diagonal sum would be 1 and, when tested against the null hypothesis as outlined, it would be highly significant. We would then incorrectly conclude that humans and model behave similarly on a per-painting basis. A similar issue arises if, for example, humans and model are always incorrect by consistently reporting the upside-down orientation: again, they will be 100% in agreement, but this outcome does not carry any specificity for distinct paintings. More generally, this problem applies to any non-random pattern of responses on the part of humans/model, including less extreme versions of the scenarios outlined above; that is, ones where a given response is not certain but has an associated probability different than chance. Our goal is to define the null hypothesis in relation to this class of scenarios.

To establish a baseline level for agreement, we calculate expected agreement under the hypothesis that humans and model act independently with relation to specific paintings: on any given trial, we assume that humans produce the four possible responses with probabilities $\{p_\uparrow, p_\rightarrow, p_\downarrow, p_\leftarrow\}$ regardless of the specific painting that is presented, and the model produces those responses with probabilities $\{q_\uparrow, q_\rightarrow, q_\downarrow, q_\leftarrow\}$; using the empirical estimates for these quantities, we calculate the expected value for their agreement a_0 and its standard deviation σ_0 on a per-painting basis. We then assess the experimentally measured agreement value $\hat{a}$ in relation to this baseline via $(\hat{a} - a_0)/\sigma_0$ (z-score); that is, we determine how far the observed agreement values score over and above their expected level under the hypothesis that humans and model present no per-painting association. When we apply this calculation, we find that the agreement value associated with the whole-painting/classifier-5 dataset (Figure 11e) returns a large z-score (>5), whereas the z-scores associated with the other four datasets (Figure 11a-d) barely reject the null hypothesis of independence. We therefore conclude that, although humans and model perform similarly on average across the entire database for all fragment-size-versus-classifier comparisons (see Figure 9), their strategies may differ on a per-painting basis. More specifically, when humans have access to fragmentary information about a specific painting, and the network is restricted to early classifiers, humans adopt a decision strategy that bears little resemblance to the strategy adopted by the network. In contrast, when the whole painting is available to human observers and the network has access to classifier-5 information, their strategies present similarities that are specific to the given painting and extend to both correct and incorrect classifications.

We propose the following explanation for these results. Earlier classifiers (corresponding to smaller receptive fields) only have access to fragment-like information during training; this constraint may steer the classifiers towards discovering local statistical regularities for the purpose of identifying the overall

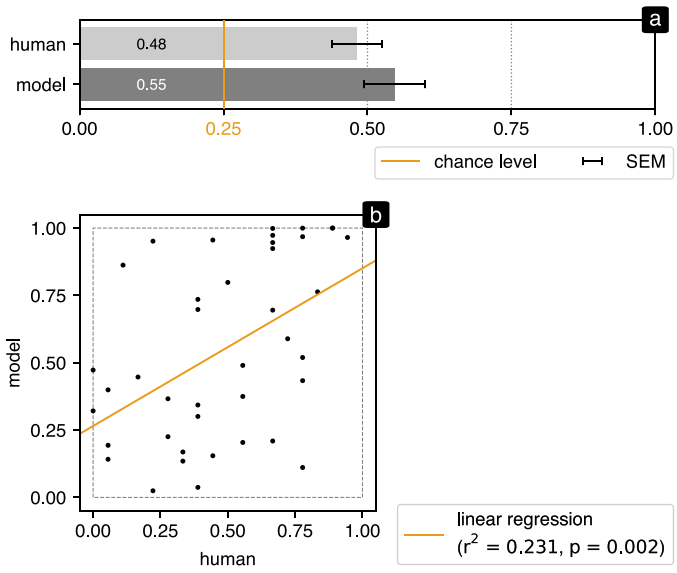

Figure 12. Comparison between our model and the results reported by Mather (2012). (**a**) The average human and model performance. The original article reports human mean performance per painting. This quantity is not directly comparable to top-1 accuracy of the model, because the latter does not reflect the level of uncertainty for each painting. We have therefore chosen to plot the raw prediction value for the correct orientation as the model metric to plot against human performance (**b**).

orientation of the painting. The resulting strategy may differ from the way in which humans approach fragments of Abstract art: the human tendency is to consider sub-parts of an abstract painting as a new complete painting, rather than as a fragment. An additional factor that may be relevant in this context is the well-documented inconsistency of aesthetic judgments across observers, especially for abstract material (Leder et al., 2016; Schepman et al., 2015; Specker et al., 2020; Vessel, 2010; Vessel et al., 2018). Although the network model does not suffer from subjective variability in the human sense, it is affected by the stochastic nature of the training protocol. Therefore, it is possible to quantify and compare internal noise between model and humans (Neri, 2010), an endeavor which we hope to pursue in future research.

We find similar results with human data collected by others. Our model is better than human observers for the selection of paintings adopted in Mather (2012) (Figure 12a), similar to the small difference we observe for our own data (Figure 9b). When we plot model-versus-human responses to individual paintings from this prior study (Figure 12b), we find a measurable trend ($p = 0.002$), but the magnitude of the correlation is relatively small ($r^2 = 0.231$) (see Dodge & Karam, 2017 for related results). Clearly, the detailed behavior

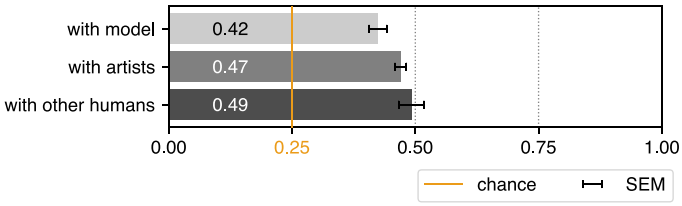

Figure 13. Painting-by-painting human agreement with network model (top), the artists who painted the images used in our study (middle), and other humans from our sample of participants (bottom). This analysis was restricted to abstract material.

of our model on a per-painting basis presents some limitations that will require further investigation.

Human agreement with model/artist/other humans for abstract paintings

Figure 13 reports human overall agreement with the model (classifier-5), the artist (whose choice is used as correct reference above) and other humans, for judgments made on whole paintings (data from fragments is excluded from this analysis). The model–human agreement is 42%. Agreement with artists is the same as human performance (already reported in Figure 9b) and it is slightly higher at 47%. Because model performance is close to human performance, this difference in agreement is due to the discrepancy already highlighted in Figure 11. Finally, for paintings that have been evaluated multiple times by humans, we can compute mutual agreement via average agreement of all possible pairs of judgements per painting. Defined this way, inter-human agreement reaches 49%, amusingly suggesting that artists themselves may not be the most reliable reference on this task or, more likely, that some artists deliberately choose non-optimal orientations (insofar as optimality is defined with reference to the orientation considered most appropriate by an average human observer).

Discussion

Relations to art composition

Despite its rich history, the study of pictorial composition has been hampered by the inherent combinatorial complexity of how graphical elements interact on canvas. Our goal was to determine whether modern computational tools, in particular deep learning, may help to tackle this difficult problem and advance our understanding of abstract art composition. Engaging with a research programme of this kind

brings up an immediate problem: how do we go about quantifying the perception of art composition in humans? It goes without saying that the cognitive phenomena underlying composition, both those exploited by the creating artist and those engaged by the observing spectators, reach far beyond the remit of one scientific study. Moreover, we would like to draw a distinction between "objective" description and "metric" description: a metric description does not necessarily imply an objective description. Our goal is not to objectify the meaning or the interpretation of the composition, neither defining rightness of compositions, but to build a metric of composition based on data, that is, existing paintings from a specified limited spectrum. Objectiveness of the metric is then transparently constrained by the range of the dataset itself, rather than its fundamental correctness. We want to organize composition around measurable dimensions that are relevant to human perception, so that perceptual processes may 1) serve as a guide in the identification of important dimensions for candidate metric(s) and 2) enable quantitative measurements of how pertinent those metrics are to composition. As a consequence, we do not have a definitive answer concerning whether and how the chosen metric is connected with the notion of "objective" description. Possibly, we will never have such an answer because the very concept of objective description may not exist. This consideration has forced us to focus on a relatively simple, yet critical metric of perceptual judgment relating to art material: determining the overall orientation of the picture (Lindauer, 1969, 1987; Liu et al., 2017; Mather, 2012). We view this as a first humble step in the direction of answering the question laid out in this article, and therefore recognize that a satisfactory account of art composition will require further research. Notwithstanding the simplicity of this behavioral metric, we discuss below its merits and its connection with existing literature in vision science.

Anecdotal evidence from the art world provides some relevant points of contact with the judgment task used in this study. Upon returning home, still lost in his thoughts, Kandinsky once noted: "I suddenly saw a painting of indescribable beauty, impregnated with great inner ardor. I was at first dumbfounded, then I quickly reached this mysterious painting on which I only saw shapes and colors and whose subject was incomprehensible." (Kandinsky, 2014). As a matter of fact, he was looking at one of his own paintings, but set out in unfamiliar orientation. A mere change of orientation in the picture was sufficient to spark a perceptual reaction that would conjure up a novel composition, serving as a cursory indication that image orientation and art composition are somehow connected, albeit in ways that we (or even the artist) may not fully understand. If we accept that this connection may be present, we must then ask whether orientation judgments of art material are supported by perceptual mechanisms that overlap with those studied by visual psychophysics; in other words, is vision science an appropriate tool for understanding this problem at any meaningful level (Mamassian, 2008)? There is evidence to support this additional connection: portrait artists, for example, are more efficient at certain visual discrimination tasks than non-artists; however, they are equally subject to the well-known face inversion effect (Devue & Barsics, 2016), a phenomenon intimately linked with the perception of overall image orientation. This brings us to the last connecting element between art composition and vision science: if we accept that global orientation judgments are relevant to art composition, and if we accept that judgments of this kind may engage similar mechanisms to those operating in other visual skills, we then ask whether this task is also important for understanding vision in general. Existing literature provides clear answers to this question.

Relations to existing literature in vision science

Prior studies offer numerous demonstrations of perceptual inversion effects in relation to meaningful visual material, such as faces (Valentine, 1988) or moving bodies (Chang & Troje, 2009; Neri et al., 2006, 2007). In these demonstrations, flipping the stimulus upside-down generally disrupts perceptual analysis by biological observers (human as well as non-human Vallortigara et al., 2005; Vallortigara & Regolin, 2006), even though it is not expected that this manipulation should impact an artificial system for which up and down do not necessarily carry any meaning (unless the system has learnt about gravity). The impact of stimulus inversion is characterized by a distinct developmental trajectory (Zhao et al., 2014) and has been associated with specific regions of visual cortex (Grossman & Blake, 2002). In short, at least within the context of contemporary thinking about higher-level vision, there is no doubt that stimulus orientation represents a valid topic of enquiry for understanding visual perception. More specifically, inversion effects are intimately associated with the notion of holistic processing, often summarized as "the whole is more than the sum of its parts," a concept that has played a significant role in the study of higher-level vision (Ullman, 1996). Inversion effects have been exploited to selectively probe holistic processes in a number of applications, ranging from natural scene perception (Neri, 2014) to action processing (Taubert et al., 2011; Cusack et al., 2015).

Furthermore, and in direct connection with the present study, previous authors have argued that deep neural networks should prove useful for the study of perceptual inversion effects (VanRullen, 2017). In our

study, perhaps the most pertinent demonstration of the profitability afforded by this computational tool is the stratification of relevant effects across layers (Figure 6a); indeed, it is difficult to imagine how this type of analysis would have been possible using more conventional modeling tools. Collectively, our results indicate that abstract art, more than other styles, relies on global compositional principles that emerge deeper into the network (Figure 6b), and that may bear on the concept of holistic processing outlined above. The term "global" may not encompass overly complex cognitive phenomena, and may to some extent overlap with the notion of 'spatially extended' as deeper layers possess larger receptive fields. Nevertheless, we have also shown that there is no simple/naive integration of orientation cues that would explain the observed patterns in our data (Figure 8). The issue of granularity remains largely unanswered at this stage, although we do make some progress in this respect.

Granularity and receptive field structure in human versus network architectures

By breaking paintings into fragments, our goal was to venture beyond prior studies and begin to consider composition as dynamic interaction of image subelements. As outlined, we find that local features of abstract art are integrated into a global representation that remains hidden from transparent explanation. This may, or may not, conform to artistic intuition. On the one hand, Abstract art explores pictorial composition on a level that is not bound by conventional relationships of experiential space, so it may be expected that the underlying structure should not be available at the level of simple spatial integration. On the other hand, it is often the case that Abstract art seems to be redundant across space (e.g., some applications of action painting), so that it would seem that little should be gained from incorporating more spatially extended information. Furthermore, Figurative art often presents complex spatial relationships on a large scale; indeed, natural scene perception is by no means a phenomenon that can be easily reduced to naive spatial integration of local cues (DiCarlo et al., 2012). We conclude that our demonstration of emergent global encoding at deeper layers for abstract art is not trivially expected based on either conventional ideas about art material, nor on mainstream considerations about receptive-field structure in hierarchical models. We discuss the latter issue further below.

The notion that visual cortex is organized along a hierarchical pathway of visual areas with progressively increasing receptive-field size is established (Dumoulin & Wandell, 2008; Yamins & DiCarlo, 2016); however, it is not at all understood how information is combined from one area to the next. At this stage, we are perhaps nearing adequate characterization and computational understanding of the transition from V1 to V2 (Freeman et al., 2013), but subsequent transformations remain poorly understood. This picture is further complicated by the known presence of feedback processes (Lamme et al., 1998), which are not implemented in any form within our model. With this in mind, it is somewhat surprising that our model is able to capture some properties of human orientation judgments for isolated fragments by simply restricting its access to more superficial layers. On the face of it, this result indicates that, by designing experiments with tailored fragmented stimuli, we may be in a position to probe human perceptual mechanisms corresponding to different layers in the model and possibly different visual areas along the processing hierarchy. We contend that this result is not trivial, both in consideration of the unresolved issues associated with inter-aerial transformations outlined above, and also in light of the fact that the connection between the notion of receptive/perceptive field on the one hand, and final behavioral response on the other hand, is far from being as straightforward as is often tacitly assumed (Neri & Levi, 2006; Spillmann, 1971). In humans, we cannot simply read out of earlier visual areas using experimental tools; what we can perhaps do is force observers to rely on signals from those earlier areas for the production of behavior (which we can measure). That we may achieve this by tailoring fragment size is not trivially expected, particularly in relation to a behavioral judgment that is not explicitly connected with global integration and that involves higher-level cognitive processes. We do not know whether the same result would be obtained for other perceptual judgments, an issue we hope to address in future research.

Notwithstanding the correspondence between human observers and model responses as discussed, we do find conspicuous differences between the behavior exhibited by the network and that measured from humans. Interestingly, those differences become particularly evident when we consider fragments, less so with whole paintings (Figure 11). We propose that this result should be interpreted in light of the considerations discussed above. As we have already noted, our model is purely feed-forward, that is, its architecture fails to incorporate important recurrent computations that are known to operate in cortex. It is conceivable that related perceptual processes are engaged by humans in our task, possibly contributing to the discrepancy we observe with respect to the model (see also Doerig et al., 2020 for related considerations). Furthermore, the nature of the discrepancy may be specific to the task/protocol we selected for this study and/or to the resolution of our measurements. We do

not have definite answers to these and other related questions, some of which we have highlighted in this Discussion. At this stage, we view our contribution as a starting point for more in-depth studies of art composition adopting a similar framework, namely the integrated application of deep learning models, data-driven extraction of regularities and psychophysical validation in human observers. Our results demonstrate that this approach is feasible and capable of generating non-trivial insights and predictions into the mechanisms underlying art composition in humans.

Keywords: machine learning, psychophysics, receptive field, pictorial composition, inversion effect

Acknowledgments

Supported by grants ANR-16-CE28-0016, ANR-17-EURE-0017, ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL from Agence Nationale de la Recherche.

Commercial relationships: none.
Corresponding author: Pierre Lelièvre.
Email: contact@plelievre.com.
Address: Laboratoire des systémes perceptifs, Département d'études cognitives Science Arts Création Recherche (EA 7410), Paris, France.

References

Arnheim, R. (2004). *Art and Visual Perception – A Psychology of the Creative Eye (2nd edition, 50th Anniversary)*. Berkeley: University of California Press. (Original work published 1954).

Chang, D. H., & Troje, N. F. (2009). Acceleration carries the local inversion effect in biological motion perception. *Journal of Vision, 9*(1), 1–17.

Cusack, J. P., Williams, J. H., & Neri, P. (2015). Action perception is intact in autism spectrum disorder. *Journal of Neuroscience, 35*(5), 1849–1857.

Cuzick, J. (1985). A wilcoxon-type test for trend. *Statistics in Medicine, 4*(1), 87–90, https://doi.org/10.1002/sim.4780040112.

Devue, C., & Barsics, C. (2016). Outlining face processing skills of portrait artists: Perceptual experience with faces predicts performance. *Vision Research, 127*, 92–103.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415–434.

Dodge, S., & Karam, L. (2017). A study and comparison of human and deep learning recognition performance under visual distortions. *2017 26th International Conference on Computer Communications and Networks, ICCCN 2017*, 1–7, https://doi. org/10.1109/ ICCCN.2017. 8038465.

Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2020). Crowding reveals fundamental differences in local vs. global processing in humans and machines. *Vision Research, 167*, 39–45.

Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *Neuroimage, 39*, 647–660.

Elgammal, A. M., Mazzone, M., Liu, B., Kim, D., & Elhoseiny, M. (2018). *The shape of art history in the eyes of the machine*. ArXiv:1801.07729 [Cs, AI], http://arxiv.org/abs/1801.07729.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience, 16*(7), 974–981.

Gaspar, C. M., Bennett, P. J., & Sekuler, A. B. (2008). The effects of face inversion and contrast-reversal on efficiency and internal noise. *Vision Research, 48*, 1084–1095.

Gong, M., Xuan, Y., Smart, L. J., & Olzak, L. A. (2018). The extraction of natural scene gist in visual crowding. *Scientific Report, 8*(1), 14073.

Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron, 35*(6), 1167–1175.

Iigaya, K., Yi, S., Wahle, I. A., Tanwisuth, K., & O'Doherty, J. P. (2020). Aesthetic preference for art emerges from a weighted integration over hierarchically structured visual features in the brain. BioRxiv 2020.02.09.940353, https://doi.org/10.1101/2020.02.09.940353.

Kandinsky, W. (1989). *Du spirituel dans l'art, et dans la peinture en particulier* (P. Sers, N. Debrand, & B. Du Crest, Trans.). Denoël: Gallimard. (Original work published 1912).

Kandinsky, W. (2014). *Regards sur le passé: Et autres textes, 1912-1922* (J.-P. Bouillon, Ed.). Hermann. (Original work published 1974).

Kandinsky, W. (1991). *Point et ligne sur plan: Contribution à l'analyse des éléments picturaux* (P. Sers, Ed.; S. Leppien & J. Leppien, Trans.). Gallimard. (Original work published 1926).

Kelley, T. A., Chun, M. M., & Chua, K. P. (2003). Effects of scene inversion on change detection of targets matched for visual salience. *Journal of Vision, 3*(1), 1–5.

Klee, P. (1961). *Notebooks, Volume 1: The thinking eye* (J. Spiller, Ed.; R. Manheim, C. Weidler, & J. Wittenborn, Trans.). Lund Humphries.

Klee, P. (1973). *Notebooks, Volume 2: The nature of nature* (J. Spiller, Ed.; H. Norden & J. Wittenborn, Trans.). Lund Humphries.

Klee, P. (1998). *Théorie de l'art moderne* (P.-H. Gonthier, Ed. & Trans.). Denoël: Gallimard. (Original work published 1924)

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1*, 417–446.

Krizhevsky, A., Sutskever, I., Hinton, E., & G. (2012). ImageNet classification with deep convolutional neural networks. *Neural Information Processing Systems, 25*, https://doi.org/10.1145/3065386.

Lamme, V. A., Super, H., & Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology, 8*(4), 529–535.

Leder, H., Goller, J., Rigotti, T., & Forster, M. (2016). Private and shared taste in art and face appreciation. *Frontiers in Human Neuroscience, 10*, 155, https://doi.org/10.3389/fnhum.2016.00155.

Lindauer, M. S. (1969). The orientation of form in abstract art. *Proceedings of the Annual Convention of the American Psychological Association, 4*(1), 475–476.

Lindauer, M. S. (1987). Perceived and preferred orientations of abstract art. *Empirical Studies of the Arts, 5*(1), 47–58, https://doi.org/10.2190/K1X2-X4VJ-6YN9-BKD8.

Liu, J., Dong, W., Zhang, X., & Jiang, Z. (2017). Orientation judgment for abstract paintings. *Multimedia tools and applications, 76*(1), 1017–1036, https://doi.org/10.1007/s11042-015-3104-5.

Locher, P. J., Stappers, P. J., & Overbeeke, K. (1999). An empirical evaluation of the visual rightness theory of pictorial composition. *Acta Psychologica, 103*(3), 261–280, https://doi.org/10.1016/S0001-6918(99)00044-X.

Mamassian, P. (2008). Ambiguities and conventions in the perception of visual art. *Vision Research, 48*(20), 2143–2153.

Mather, G. (2012). Aesthetic judgement of orientation in modern art. *i-Perception, 3*(1), 18–24, https://doi.org/10.1068/i0447aap.

McManus, I. C., Cheema, B., & Stoker, J. (1993). The aesthetics of composition: A study of Mondrian. *Empirical Studies of the Arts, 11*(2), 83–94, https://doi.org/10.2190/HXR4-VU9A-P5D9-BPQQ.

Neri, P. (2010). How inherently noisy is human sensory processing? *Psychonomic Bulletin & Review, 17*, 802–808.

Neri, P. (2014). Semantic control of feature extraction from natural scenes. *Journal of Neuroscience, 34*, 2374–2388.

Neri, P., & Levi, D. M. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision Research, 46*, 2465–2474.

Neri, P., Luu, J. Y., & Levi, D. M. (2006). Meaningful interactions can enhance visual discrimination of human agents. *Nature Neuroscience, 9*, 1186–1192.

Neri, P., Luu, J. Y., & Levi, D. M. (2007). Sensitivity to biological motion drops by approximately 1/2 log-unit with inversion, and is unaffected by amblyopia. *Vision Research, 47*, 1209–1214.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155*, 23–36.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Chintala, S., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., . . . Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlche-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc, http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Rodriguez, C. S., Lech, M., & Pirogova, E. (2018). Classification of style in fine-art paintings using transfer learning and weighted image patches. *2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 1–7, https://doi.org/10.1109/ICSPCS.2018.8631731.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., . . . Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. arXiv: 1409.0575 [cs]. http://arxiv.org/abs/1409.0575.

Schepman, A., Rodway, P., Pullen, S. J., & Kirkham, J. (2015). Shared liking and association valence for representational art but not abstract art. *Journal of Vision, 15*(5), 11, https://doi.org/10.1167/15.5.11.

Schwabe, K., Menzel, C., Mullin, C., Wagemans, J., & Redies, C. (2018). Gist perception of image composition in abstract artworks. *i-Perception, 9*, 204166951878079, https://doi.org/10.1177/2041669518780797.

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science, 5*, 399–426.

Simonyan, K., & Zisserman, A. (2014, September 4). Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556 [cs], http://arxiv.org/abs/1409.1556.

Specker, E., Forster, M., Brinkmann, H., Boddy, J., Immelmann, B., Goller, J., … Leder, H. (2020). Warm, lively, rough? Assessing agreement on aesthetic effects of artworks (R. T. H. Ho, Ed.). *PLoS One, 15*(5), e0232083, https://doi.org/10.1371/journal.pone.0232083.

Spillmann, L. (1971). Foveal perceptive fields in the human visual system measured with simultaneous contrast in grids and bars. *Pflugers Archiv (European Journal of Physiology), 326*, 281–299.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015, April 13). Striving for simplicity: The all convolutional net. arXiv: 1412.6806 [cs], http://arxiv.org/abs/1412.6806.

Taubert, J., Apthorp, D., Aagten-Murphy, D., & Alais, D. (2011). The role of holistic processing in face perception: Evidence from the face inversion effect. *Vision Research, 51*(11), 1273–1278.

Ullman, S. (1996). *High-level vision*. Cambridge, MA: MIT Press.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology, 79*(Pt 4), 471–491.

Vallortigara, G., & Regolin, L. (2006). Gravity bias in the interpretation of biological motion by inexperienced chicks. *Current Biology, 16*, R279–280.

Vallortigara, G., Regolin, L., & Marconato, F. (2005). Visually inexperienced chicks exhibit spontaneous preference for biological motion patterns. *PLoS Biology, 3*(7), e208.

VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology, 8*, 142.

Vessel, E. A. (2010). Beauty and the beholder: Highly individual taste for abstract, but not real-world images. *Journal of Vision, 10*(2), 1–14, https://doi.org/10.1167/10.2.18.

Vessel, E. A., Maurer, N., Denker, A. H., & Starr, G. G. (2018). Stronger shared taste for natural aesthetic domains than for artifacts of human culture. *Cognition, 179*, 121–131, https://doi.org/10.1016/j.cognition.2018.06.009.

WikiArt. (n.d.). WikiArt.org - Visual Art Encyclopedia, https://www.wikiart.org/.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356–365.

Yovel, G., & Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Current Biology, 15*, 2256–2262.

Zhao, J., Wang, L., Wang, Y., Weng, X., Li, S., & Jiang, Y. (2014). Developmental tuning of reflexive attentional effect to biological motion cues. *Scientific Report, 4*, 5558.