Cognition 231 (2023) 105319

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Characterising and dissecting human perception of scene complexity

Cameron Kyle-Davidson [a,*], Elizabeth Yue Zhou [b], Dirk B. Walther [c], Adrian G. Bors [a], Karla K. Evans [d]

[a] *University of York, Department of Computer Science, York, YO10 5GH, UK*
[b] *New York University, Department of Psychology, New York, NY, USA*
[c] *University of Toronto, Department of Psychology, Toronto ON, M5S 1A1, Canada*
[d] *University of York, Department of Psychology, York, YO10 5DD, UK*

ARTICLE INFO

Dataset link: VISCHEMA-COMPLEXITY (Original data)

Keywords:
Perceptual complexity
Scene complexity
Neural networks
Modelling vision
Human vision

ABSTRACT

Humans can effortlessly assess the complexity of the visual stimuli they encounter. However, our understanding of how we do this, and the relevant factors that result in our perception of scene complexity remain unclear; especially for the natural scenes in which we are constantly immersed. We introduce several new datasets to further understanding of human perception of scene complexity. Our first dataset (VISC-C) contains 800 scenes and 800 corresponding two-dimensional complexity annotations gathered from human observers, allowing exploration for how complexity perception varies across a scene. Our second dataset, (VISC-CI) consists of inverted scenes (reflection on the horizontal axis) with corresponding complexity maps, collected from human observers. Inverting images in this fashion is associated with destruction of semantic scene characteristics when viewed by humans, and hence allows analysis of the impact of semantics on perceptual complexity. We analysed perceptual complexity from both a single-score and a two-dimensional perspective, by evaluating a set of calculable and observable perceptual features based upon grounded psychological research (clutter, symmetry, entropy and openness). We considered these factors' relationship to complexity via hierarchical regressions analyses, tested the efficacy of various neural models against our datasets, and validated our perceptual features against a large and varied complexity dataset consisting of nearly 5000 images. Our results indicate that both global image properties and semantic features are important for complexity perception. We further verified this by combining identified perceptual features with the output of a neural network predictor capable of extracting semantics, and found that we could increase the amount of explained human variance in complexity beyond that of low-level measures alone. Finally, we dissect our best performing prediction network, determining that artificial neurons learn to extract both global image properties and semantic details from scenes for complexity prediction. Based on our experimental results, we propose the "dual information" framework of complexity perception, hypothesising that humans rely on both low-level image features and high-level semantic content to evaluate the complexity of images.

0. Introduction

It is readily apparent that humans are capable of implicitly determining the complexity of a given image upon perception; shown a blank canvas and an abstract painting, it is easy to identify the more complex of the two. However, it is less clear how humans perceive the everyday complexity in which they are immersed: that of the natural scene (see Fig. 1 for examples). It remains relatively unknown which mechanisms underlie complexity perception. Determining these mechanisms may lead to a better understanding of how the human visual system operates, and how it processes scene complexity. Complexity research has implications for aesthetics research (Mayer & Landwehr, 2018; Van Geert & Wagemans, 2020), and even potentially image memorability research (Saraee et al., 2020). In addition to theoretic advancements, there are also numerous practical applications for the study and measurement of perceptual complexity. These include influencing users' experience on webpages (Deng & Poole, 2012), and marketing applications such as designing brand logos (Wang et al., 2018), and car sales (Landwehr et al., 2011). It may impact psychological experiment design (you may want all your visual stimuli to be of similar complexity to exclude a confounding factor), healthcare applications (the evaluation of cognitive image processing disorders; how easily a patient can process an image of known complexity),

[*] Corresponding author.
E-mail address: cameron.kyle-davidson@york.ac.uk (C. Kyle-Davidson).

https://doi.org/10.1016/j.cognition.2022.105319
Received 4 February 2022; Received in revised form 28 October 2022; Accepted 29 October 2022
Available online 15 November 2022

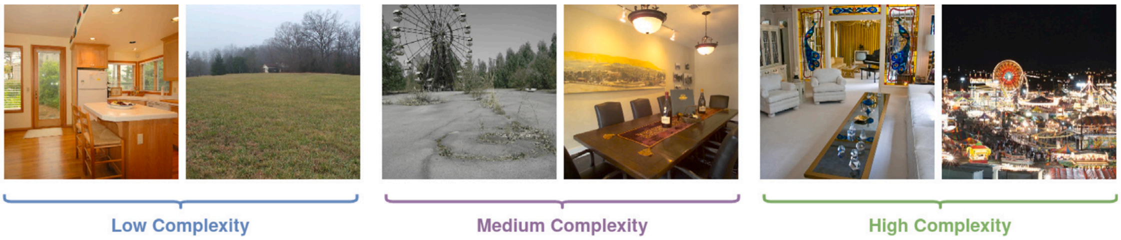

Fig. 1. Examples of low, medium, and high complexity images from the VISC-C dataset. Complexity ratings are gathered from human observer ratings.

and virtual reality environment development, where a simulated world desiring realism should be capable of matching the complexity of real environments, without appearing too simplistic, or overly complex. The development of complexity models allows the extraction of complexity values from scenes to take place automatically without requiring a human-in-the-loop for each application. Without these models, the majority of practical applications become significantly more difficult, requiring costly human intervention for every instance of the application. Such modelling offers the ability to computationally determine complexity; side-stepping the intensive data gathering process. Computational models open up new, less intensive ways to approach the understanding of these problems, and a way to move beyond theoretical, into practical applications.

The first apparent quantification of what humans might perceive as complexity appears in the early 20th century (Birkhoff, 1933), defined as the count of elements in an image. Later work redefines complexity as the intricacy or detail present in a line drawing (Snodgrass & Vanderwart, 1980), or as the degree of difficulty involved in generating a verbal description of a texture (Heaps & Handel, 1999), or evaluates complexity in the context of aesthetics (Day, 1968). Birkhoff hypothesises that aesthetic perception is in some manner based upon the ratio between the level of *order* and the level of *complexity* present in the stimuli. Later work builds upon this, finding that while order and complexity may represent two different dimensions of perception, there is likely to be some interplay between the two factors. That is, the level of order present is capable of influencing how the complexity present in the stimuli is perceived (Van Geert & Wagemans, 2020). Under this view, explaining complexity may actually require measures of *order*, in whole or in part, and models that combine various measures of complexity may be capturing part of the interplay between complexity and order. However, most research into complexity does not specifically target scene perception, with initial research on complexity perception in scenes (Olivia et al., 2004) finding evidence that clutter and mirror symmetry play a key role in visual complexity, along with openness and object organisation (e.g. factors based upon scene gist research (Oliva & Torralba, 2001), where gist represents the general semantic content of the scene). As computing systems became more powerful, and the field of information science evolved, so too have definitions of complexity.

One computational technique often applied as an analogue of visual complexity involves the calculation of the Shannon entropy of the image (Cardaci et al., 2009; Yu & Winkler, 2013), under the hypothesis that more complex images have a greater level of entropy (or disorganisation), and simpler images contain more redundant information (and hence, lower entropy). Entropy-based measures appear to be one method of operationalising visual clutter (Rosenholtz et al., 2007), as the more cluttered the image, the more disorganised the image, hence the greater entropy. Another potential information theoretic measure is Kolmogorov complexity (Kolmogorov, 1965). The Kolmogorov complexity of an output defines the shortest length of a computer program that could produce that output, and while uncomputable, can be approximated (Rigau et al., 2007). Naturally, information-theoretic measures are somewhat divorced from human perception, and the applicability of entropy measures to scene images remains relatively unclear. An image of random, coloured noise is high-entropy, yet meaningless to a human. More recent research has turned to finding combinations of metrics that predict visual complexity (Corchs et al., 2016a; Nagle & Lavie, 2020), some information theoretic, some more grounded in human perception. These models are capable of predicting human complexity scores with an accuracy greater than any single predictor alone. The most recent work focuses on developing neural models of perceptual image complexity, finding that visual complexity information arises within the feature maps of deep convolutional networks (Saraee et al., 2020), and similarly that multiple regions across the brain are involved with the representation of the complexity inherent in naturalistic stimuli (Güçlütürk et al., 2018).

Progress in understanding human perception of visual complexity, especially in the area of natural scene perception (Corchs et al., 2016a), is made more difficult by a lack of high-quality, varied scene datasets (Nagle & Lavie, 2020). Existing datasets are either small (sub-200 images) (Corchs et al., 2016b), or are object-focused, which leads participants to evaluate the complexity of the object that fills the frame rather than the image as a whole. While object complexity likely contributes to the overall perception of complexity in a given scene, in order to understand scene complexity these objects must be placed in the wider context of their surroundings. Finally, drawing from image memorability research, it is becoming more apparent that perceptual image characteristics, while often represented as a single score for a given image, are better represented as two-dimensional properties that vary across an image (Akagunduz et al., 2019). Currently, available datasets indicate that the complexity rating a human may give is based on the entire image, which ignores the local properties of complexity within that image.

Our aim is to address previous shortcomings by developing human observer based, high quality, two-dimensional scene complexity datasets, and computationally operationalising psychological measures of perceptual complexity to further understand how humans perceive complexity. We choose four different metrics: clutter, symmetry, entropy, and openness, each either hypothesised, or evidenced to have some relation to complexity in prior work, though which have not been examined in conjunction. We employ these measures together for the first time in order to develop an understanding of exactly which perceptual factors account for human perception of visual complexity, 'factorising' out the degree to which each metric helps to explain human variance in complexity ratings. Our primary dataset, which we call 'Vischema-Complexity' (VISC-C), is based upon a categorical scene dataset (known as VISCHEMA (Akagunduz et al., 2019)), and consists of 800 images with 800 complexity scores; giving a rating for each image, obtained from a human observer experiment. In addition, critically, it contains 800 'complexity maps' that capture the image regions that participants find simple or complex, and for the first time reveal the image areas that contribute to perceptions of scene complexity. We also introduce VISC-CI, a complexity dataset of complexity scores and complexity maps from human observer experiment of vertically flipped variants of our scene images. Vertical inversion results in destroying or damaging the semantic content present in an image (Epstein et al., 2006; Kelley et al., 2003; Neri, 2014; Walther et al., 2009), when perceived by a human, thus allowing the quantification of the effect of scene semantics on perceptions of image complexity.

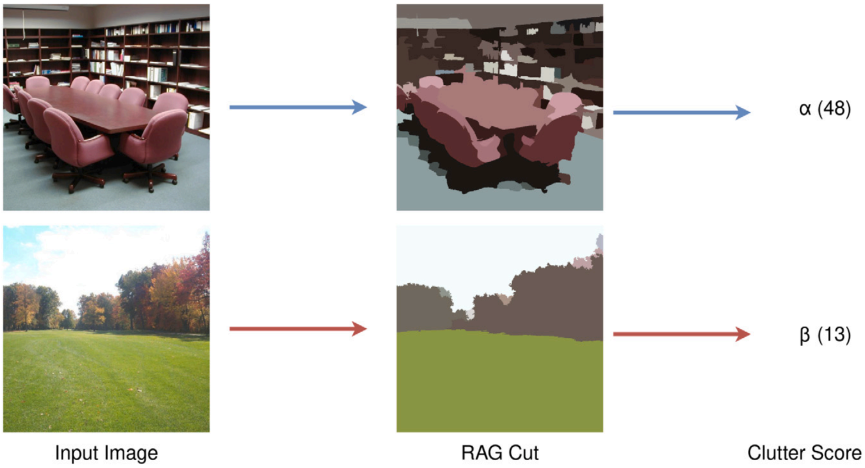

Fig. 2. Example of clutter algorithm working on a perceptually simple image and a more complex scene, as rated by humans.

Further we generalised our analysis to an existing image set, BOLD5000 (Chang et al., 2019). Finally, we develop and evaluate a neural network model capable of simultaneously predicting complexity scores and two-dimensional complexity maps. We examine which features these "black box" neural models have learned to associate with perceptual complexity by dissecting the network and examining individual artificial neurons. Across our behavioural studies and computational analysis, we find evidence that indicates that low-level features alone cannot fully explain how humans perceive the complexity of scene images. To account for this, we propose a "dual information" framework for human complexity perception, which suggests both low-level features *and* semantic information is used for the evaluation of the complexity of scenes.

1. Factorising complexity

Upon review of studies investigating human perception of visual complexity it is evident that multiple factors contribute to this perception, and in part some of these factors can be operationalised with computational measures. However, it is difficult to ground complex information-theoretic measures to human perception. As the first step in our investigation, we instead define a set of four possible complexity measures (entropy, clutter, symmetry and openness). These four metrics are selected based on their existing grounding in cognitive psychology that considers the complexity of scene images (Olivia et al., 2004). While there are many existing and varied operationalisations of "objective complexity", it is impractical to consider all of them. Instead, we aim to select measures that capture both information-theoretic (entropy), as well as more perceptual (clutter, symmetry, openness) aspects of complexity. We evaluate their success in explaining the variance inherent in human complexity perception obtained from human observer studies and recorded in the VISC-C, VISC-CI and BOLD5000 datasets. As colour has been found to both relate (Corchs et al., 2016a), and not relate (Ciocca et al., 2015) to complexity, we err on the side of caution and include colour as integral part of the examined factors, where appropriate (clutter, symmetry, openness). Each metric has its own corresponding range of values, which is discussed in each following section. However, for the purpose of analysis conducted in the study all metrics are normalised to values between 0 and 1.0 over the dataset as a whole.

1.1. Clutter

It is intuitive that the level of variation across an image would, in some fashion, be related to the complexity of that image. Prior research has revealed that human perception of clutter is one of the components that correlates with scene complexity (Olivia et al., 2004). There have been various attempts to characterise clutter, primarily through information-theoretic measures (Rosenholtz et al., 2007). Here, instead of an information-theoretic entropy-based approach, we characterise clutter as the number of separable regions computed by a normalised graph-cut of the region-adjacency graph of an image (Shi & Malik, 2000). The normalised graph cut here divides an image into a number of 'perceptually distinct' regions. This has the effect of grouping similar parts of the image together into one average-colour region. Our hypothesis here is that images that are perceived to be more complex would be decomposed into a greater number of distinct and separable regions, whereas simpler scenes are segmented into less regions, as overall they contain more 'perceptually similar' parts. The cost of dividing a graph into two disjoint regions is the summed weights of the edges that are removed in order to create the bisection. The optimal bisection of this graph is the bisection with the lowest cost (i.e, that optimally separates two perceptually distinct regions). The normalised cut of graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ into distinct sets A, B is given in Eq. (1). $Cut(A, B)$ computes the sum of edge weights removed, and $Assoc(A, V)$ is the sum of edge weights from A to all vertices in the region-adjacency graph. The normalised cut essentially finds the 'cheapest' way to divide an image into distinct regions, by finding the best 'cut' between the regions (which are groups of similar pixels). The lower limit of this measure is 1 (as all pixels in the image are similar enough to be grouped into a single region, an occurrence highly unlikely for a natural scene dataset), while the upper limit is *theoretically* unbounded, but in practice is limited by the natural level of variation in a scene image. Intuitively, this measure can be thought of as measuring the level of variation across a scene in a without being as vulnerable to high-frequency elements as information theoretic measures. This prevents frequent textural changes (i.e, those that naturally occur across a grass field) from skewing the results too heavily. This effect can be seen in Example β in Fig. 2.

$$Ncut(A, B) = \frac{Cut(A, B)}{Assoc(A, V)} + \frac{Cut(A, B)}{Assoc(B, V)} \tag{1}$$

1.2. Patch-based symmetry

There is no argument that symmetry is an important part of human perception; symmetry detection is a flexible and rapid process (Treder, 2010). When asked to perceive symmetry, humans are capable of doing so even at extremely rapid presentation times. However, even when *not* asked to specifically detect symmetry, there is evidence that symmetry is detected and used (for example, during visual search) at a preattentive level (Wagemans, 1997). Certainly, it appears that the detection of mirror symmetry enjoys some advantage over other forms

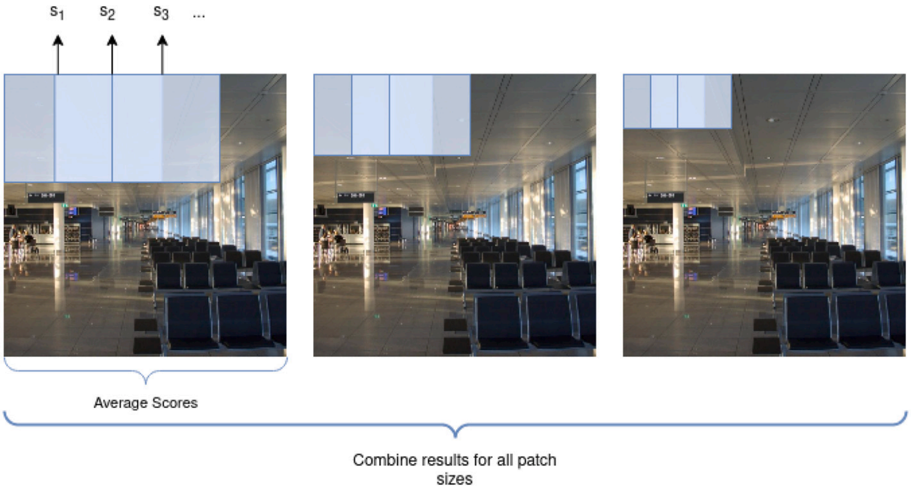

Fig. 3. Simplified diagram for symmetry computation with three different patch sizes.

of symmetry (rotational, or skewed) (Wagemans, 1995), with vertical mirror symmetry enabling better detection of symmetric objects compared to asymmetric (Machilsen et al., 2009). Much of the research into how symmetry affects the perception of images is conducted with “global symmetry”. This is defined as the difference between two regions of an image generated by bisecting the image either horizontally or vertically; equivalent to folding an image in half and determining the degree to which each half matches the other. When participants are asked to give rankings on the complexity of images, this global symmetry of the image has been found to be a significant component of those rankings (Olivia et al., 2004), and evidently relates to complexity in some manner. Computationally, most symmetry extraction methods focus on detecting the axis of symmetry of objects, or for determining where rotational symmetry appears in an image (Hauagge & Snavely, 2012; Liu et al., 2010; Patraucean et al., 2013). These methods are object-focused, and hence provide less information about the general symmetry present in a scene. While symmetry (at least global symmetry) appears important for general perception, for which reasons might it be important for complexity perception? The most obvious answer is that symmetry is a measure of redundant information, and that the more symmetries present, the less information is present in the scene. For example, given perfect global symmetry along one axis (say, vertical), the same scene could be exactly described by *half* of that scene, by mirroring the image across the symmetry axis. The same may hold for smaller, more local elements in the scene. In this sense, one might consider symmetry to be a possible representation of the level of *order* present in the scene. As order and complexity interact in complex and unclear ways, but nonetheless interact, the effect of symmetry on complexity perception may in-fact reduce down to the interaction effect that order itself has on perceived complexity. To extract local symmetries, we focus on extracting the symmetry of patches across the image, a compromise between computationally intensive methods that identify the symmetry of objects, and simpler methods that evaluate global bilateral symmetry. Our approach in particular works well for scenes, whose main semantic details are often aligned in a horizontal plane.

In this work, we define local patch symmetry as the mean of the horizontal and vertical symmetry contained within arbitrary-sized patches across the scene image. Given an image patch $N_{ij}^{h\times w\times c}$, at location (i,j), we bisect the patch vertically giving $(A^{h\times\frac{w}{2}\times c}, B^{h\times\frac{w}{2}\times c})$, where $A_{ij} = N_{i,0<j<\frac{w}{2}}$ and $B_{ij} = N_{i,\frac{w}{2}<j<w}$, defining $F_h(A)$ as the horizontal flip of A, the horizontal symmetry of the patch is simply $sym_h(N) = \sqrt{(f_h(A)-B)^2}$. The vertical case is similarly defined. Hence, $sym(N) = \frac{H_n^{sym}+V_n^{sym}}{2}$, and the overall symmetry of image I given by $sym(I) = \frac{1}{|K|}\sum_{i=0}^{I_{cols/s}-1}\sum_{j=0}^{I_{rows/s}-1} sym(N_{i\cdot s,j\cdot s}^{h\times w\times c})$ where K is the set of patches extracted, and s the stride. Intuitively, this method calculates how symmetric (both horizontally, and vertically) a small region of the scene is, and keeps track of this symmetry for a set of regions that together, cover the entire scene. This allows the symmetry of local objects to be considered when computing how ‘symmetric’ the scene is as a whole. For example, a scene may be globally lacking in symmetry, but may contain a highly symmetric region which decreases that scene’s overall perceptual complexity. A simplified diagram of this process is shown in Fig. 3. We choose patches systematically, selecting sizes to capture large, medium, and small elements of the scene. Our first patch is sized at 100 by 100 pixels, our next at 50 by 50, and our smallest by 25 by 25 pixels. Each patch was stepped across the image by half its resolution, ensuring that elements that would otherwise fall on the boundaries between patches are not missed. Smaller/larger patches, and smaller step-sizes are possible, but increase the computational requirements significantly. Our selected sizes, and step distance, balance capturing a reasonable amount of scene elements within them, with computing time required. The output of this measure ranges from 0 (no symmetry at all) to 1.0 (perfectly symmetrical in all locations). In practice, no scene images lie at these extremes.

1.3. Entropy

It is common in the literature on complexity to examine measures of entropy and the relationship between entropy and complexity. Shannon entropy is measured in ‘bits’, and represents the mathematical limit on the degree of compression to which data can be losslessly compressed (i.e, without losing information, such as degrading the quality of an image). In the case of images, those with more variation tend to contain more information, and as a result, require more bits of information to losslessly describe. For the sake of completeness, we considered the Shannon entropy of the image histogram $H = -\sum_k p_k log_2(p_k)$ with p_k representing the probability of finding a pixel of k intensity over the image. It is generally expected that entropy reflects perceived complexity in some way, at least in the case of simplistic images such as those shown in Fig. 4. Entropy can be thought of as being sensitive to high-frequency information that contains minimal redundancy, and hence aligns with the “complexity” property in the order-complexity multidimensional space. Here, we compute the entropy over each colour channel in the HSV (hue, saturation, value) space, at several different resolutions via Gaussian pyramid (resolution 1, 0.5, 0.25, and so on). This allows us to capture variation at multiple scales, while taking colour information into account. Shannon entropy is minimally bounded by zero, indicating no variation at all, and the maximal bound depends upon the input. However, for all input, the maximal bound is related to the Shannon entropy of random data (see Fig. 4).

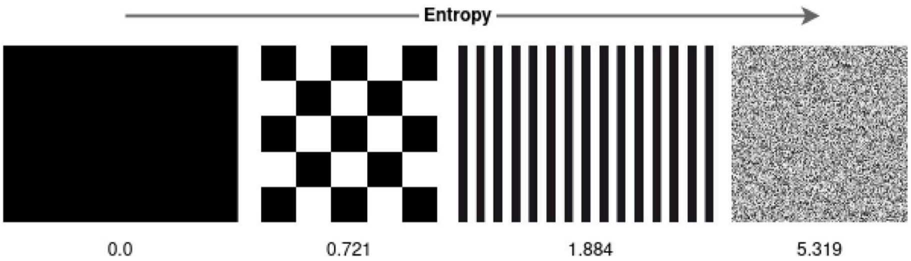

Fig. 4. Entropy values for different images.

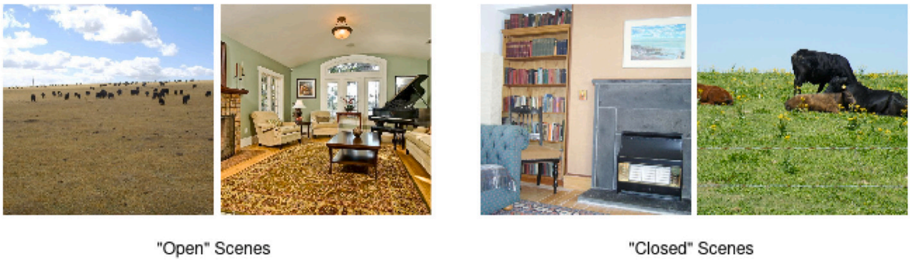

Fig. 5. Examples of open vs closed images.

1.4. Openness

Despite human behavioural evidence for the influence of scene openness (Olivia et al., 2004) on perception of complexity, this factor remains relatively unexamined in computational approaches to perceptual complexity. Images with clear horizon lines and lack of boundaries are said to be 'open' (e.g., a field), and scenes that lack these, to be closed (e.g., a photograph of a kitchen taken perpendicular to a flat surface). We computed openness following the methodology from Ross and Oliva (2010), which calculates openness based upon the spatial frequencies (Oliva & Torralba, 2001) present in the scene, and hence predicts openness scores for every image in our dataset. In Ross et al., a set of human ratings are gathered for a variety of scene images (7,138 total), including "openness" ratings. From these input images, GIST features were computed, which provide a low-dimensional description of the scene. GIST features are dependent upon the local response of frequency filters at different scales and orientations that are run over the input image. These low-dimensional features are combined with a learning algorithm using the human-ratings as a target. The fully trained learning algorithm is capable of estimating openness given arbitrary input images (assuming said images have their GIST features computed). Some examples of open vs closed images are shown in Fig. 5. While openness has been found to potentially relate to human perception of complexity (Olivia et al., 2004), it is unclear what powers this effect. Seeing as openness is a measure that depends upon the spatial layout of the scene, it may be that openness, like symmetry, is a measure of "order". More open scenes tend to be more ordered (and perhaps, globally symmetric) than closed scenes, which are more likely to be disordered. For example, a wide picture of a field may be more ordered than a closed-in photograph of a kitchen counter, containing objects strewn over it. Openness is on a scale between 0 and 1.0, with 0 indicating a fully 'closed' scene, and 1.0 indicating a fully 'open' scene. In practice, scene images tend not to lie at these extremes.

2. Study 1 — Two dimensional complexity

To evaluate which factors relate to human perception of complexity, we conducted a study on human observers, and tested our predefined computational measures against human complexity ratings. The study was designed to capture both complexity scores and two-dimensional annotations across a series of scene images. By evaluating our computational measures against the same images, we can determine which factors explain human perception of complexity. We term the dataset resulting from this study "VISC-C" for "VISCHEMA-COMPLEXITY.

2.1. Participants

A total of forty participants aged between 18 and 65, and fluent in English participated in the study. There were no other preconditions. Participants were paid for their participation and no personally identifiable information about participants was gathered or stored by the authors. Participants informed consent was obtained, and they were free to withdraw from the study at any time. The study was approved by the ethics board of the University of York, UK.

2.2. Materials

The stimuli used were images from the VISCHEMA image set: a categorical scene dataset initially gathered for the purposes of image memorability studies (Akagunduz et al., 2019). The dataset consists of 800 images with a resolution of 700 x 700 pixels. The image-set is divided into eight classes of 100 images each, with each class corresponding to a commonly encountered scene category. Available classes are: kitchen, living-room, conference-room, airport-terminal, work/home (containing images of houses/office buildings), public entertainment (amusement parks/playgrounds), populated outdoor scenes (pastures/golf courses), and isolated outdoor scenes (mountains/badlands). Example images are shown in Fig. 6.

2.3. Procedure

The study was conducted online via Prolific (Prolific: https://www.prolific.co/, 2022), an online experimentation platform. Participants were shown a continuous stream of 200 scene images and completed the task at their own pace. Each stream was built by randomly sampling from the total 800 image dataset. For each image in the stream, they were first asked to rate the complexity of the image on a scale between 0 (least complex) and 100 (most complex). Once participants gave a rating, they were then asked to annotate the image. Each participant was randomly asked to annotate either complex regions or simplistic regions in the image. The randomisation was balanced to ensure an identical amount in total of complex and simple ratings per image. In no case was any participant asked to

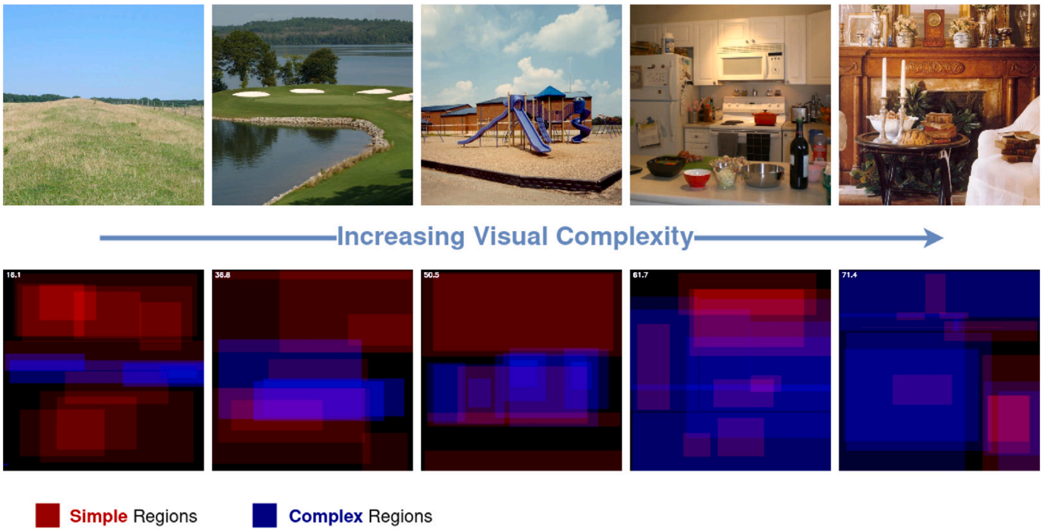

Fig. 6. A set of scenes sorted into ascending complexity, as rated by a group of human observers. The images below the arrow reveal the regions that humans labelled as complex (in blue) or simple (in red). Regions labelled as simple often contain textural variation (e.g., grass in image 1, or the sky/clouds in image 3), yet are labelled simple nonetheless. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

annotate both the simple and complex regions of the same scene image. In this manner we acquire independent annotations of both simple and complex regions for each of the images in the dataset.

In our study, every image stream shown to a participant was first randomised to minimise the sequential context effect and avoid potential biasing issues (García-Pérez & Alcalá-Quintana, 2011; Ulrich & Vorberg, 2009) that may arise in 2-Alternative Forced Choice style complexity studies; hence no two participants saw the same stream of images, and the average complexity score and annotations for the image can be considered independent of the context of the other images in the stream. We obtained 10 score ratings, and 10 annotations for each of the 800 images (five complex annotations and five simple annotations). A participant had to label at least one, and at most three, rectangular regions in an image before continuing on to the next image. Participants were free to choose the size of the annotation.

2.4. Data analysis

We employed a hierarchical regression analysis (HRA) to analyse the contribution of each potential computational factor to perceived complexity, considering the contribution of the previous factors. We based our initial ordering of the factors on the order of their singular degree of correlation with human complexity ratings. Manipulating the order in which the factors were entered into the HRA, did not have any significant effect on the result. Notably, we tested whether entropy or clutter as the first factor results in decreased explanatory power of whichever factor is added second, and find that this does not change the outcome of the analysis. Hence, we start with clutter, then in turn add entropy, patch-based symmetry, and openness. The complexity score of any given image is defined as an average of scores from participants who saw that image.

The two-dimensional data are analysed separately. We concatenated all the per-image annotations into a singular two-channel ‘complexity map’, which captures complexity in one channel, and simplicity in the other. Annotation coverage is calculated by combining all of the simple or complex annotations of that image, and determining, as a percentage, how much of the image is covered by those annotations. Annotation intensity refers to the per-pixel intensity of the complex or simple channel from the complexity map image, and is related to how many participants’ annotations capture a particular location in that image. Our hierarchical regression analysis is only concerned with the one-dimension complexity scores, a more advanced method is necessary for predicting complexity maps (see Section 5).

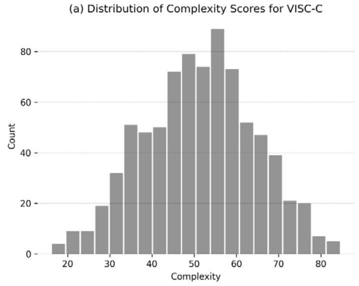

Fig. 7. The distributions of human decided complexity scores for the VISC-C dataset.

2.5. Results

Fig. 7 shows that human complexity ratings follow a Gaussian distribution across images, a property reasonably expected for a scene dataset. Most images are unlikely to be either minimally or maximally complex. The mean complexity score for the images was 51.25, and the standard deviation was 13.14. We know from prior work that complexity ratings given by humans for images are consistent. However, there is little data on the consistency of complexity ratings purely for scene images, and whether participants agree that the same regions of the scene are simple or complex.

We evaluated both the consistency of participant scores and the consistency of our two-dimensional complexity annotations. Participant consistency was measured by dividing the participant data into two splits, and computing both complexity maps and scores from each half of the data. We compared the scores from each split via the Spearman’s correlation, and the two-dimensional maps via the Pearson 2D Correlation (P^{2D}), following prior literature (Akagunduz et al., 2019). We evaluated 100 splits for the scores, and 25 splits for the complexity maps. Participants show a strong agreement in their complexity scores ($r = 0.72$). Correcting for split-half reliability (and hence estimating for the entire dataset) via the Spearman–Brown (S–B) correction, this increases to ($r = 0.84$). They also saw a good agreement on the complex regions of an image ($P^{2D} = 0.41$), and to a lesser extent, on the simple regions of the image ($P^{2D} = 0.27$). From the score consistency data, we can say that, on average, a random symmetrical split of human complexity ratings can explain 70% (based on corrected reliability) of the variance of the other splits ratings, the other 30% is surmised

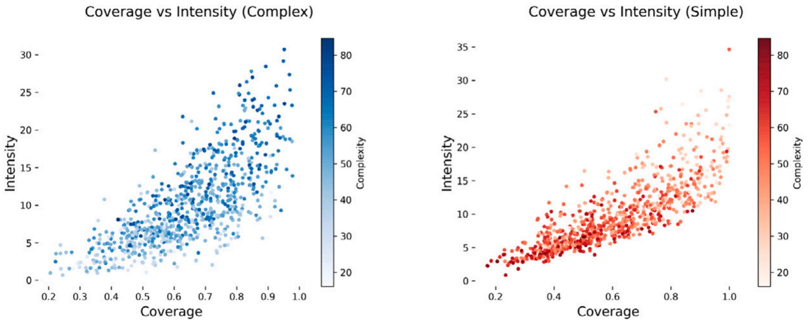

Fig. 8. Relationship between annotation coverage, intensity, and complexity for scenes. As coverage and intensity of the complex channel increases, so does the human complexity score ratings, and vice-versa for simplicity.

Table 1
Bivariate correlation table for low-level features (VISC-C/VISC-CI).

	Clutter	Entropy	Openness	Symmetry
Clutter	–	0.38	−0.38	−0.71
Entropy	–	–	−0.21	−0.44
Openness	–	–	–	0.47

to be due to individual differences between participants. In comparison, aesthetics judgements can only explain 19% of a symmetrical split (Brielmann & Pelli, 2019), leaving much more up to individual differences than in the case of complexity ratings.

To evaluate the two-dimensional annotations, we considered two properties: annotation coverage, and average simple or complex intensity. Intuitively, we assumed that a more complex image should contain more complex annotations, and a simple image should contain more simple annotations. The more intense these annotations in the complexity map, the more agreement exists between participants that the indicated region is of consequence, and the more complex (or simple) the region. We find that both annotation coverage and annotation intensity are strongly related to the complexity scores given by the participants. Annotation coverage and intensity are predictive of complexity score (multiple linear regression, $R^2 = 0.6$, also see Fig. 8) indicating the participants are labelling the images in-line with their scores. These results indicate that our two-dimensional annotation maps are indeed capturing both complexity and simplicity, and are strongly associated with "single-score" measures of complexity.

However, it is still not clear which low-level image feature could help explain the single-score complexity ratings that humans attribute to scene images. To investigate this, we employ a hierarchical regression analysis (HRA) to determine how much additional variance each factor contributes to the overall score rating. Results for the HRA are provided in Table 2, and we show correlation coefficients for each low-level feature in Table 1. Symmetry and clutter are naturally negatively correlated, explaining 50% of the variance of each other. Intuitively, this makes sense, as the presence of more clutter reduces the likelihood of local symmetry occurring, and perhaps reflects some interplay between complexity and order. The rest of the unexplained 50% reveals why symmetry helps explain complexity beyond clutter alone. Our HRA reveals that the computational complexity factors explain approximately 36% of the variance inherent in human complexity ratings. Generally, we can say that any measure that approaches or exceeds this 'target score' of 70% captures complexity to the same degree with which two disjoint groups of humans will agree with each other on the complexity of an image. Thus, it appears that we can capture just over half of the potential variance with low-level measures. Lastly, the results indicate that human complexity ratings are well explained by both clutter, and patch-based symmetry, and that entropy and openness contribute little. Visual clutter explains the most variance in complexity scores, followed by local symmetry. It is intuitive that the more cluttered the scene, the more complex the scene. Conversely, the more locally symmetrical features exist in the scene, the less complex the scene is rated, as there is less locally novel information to be processed. Entropy appears to have minimal explanatory power for perceptual scene complexity, as does openness. For completeness, we also compare against a prior complexity measure that has seen success in aesthetics research. Specifically, this preexisting metric, based upon image compression, has been shown to be relevant to processing fluency theory (Mayer & Landwehr, 2018), and is associated with aesthetic liking of images. The 'imagefluency' complexity metric calculates the ratio of the compressed file-size to the uncompressed file-size (via the ZIP/deflate algorithm), with the ratio indicative of the level of complexity present in the image. We computed this ratio over our dataset, and find that it explains 18% of the variance of our human complexity ratings. This is approximately half that of our low-level measures combined, and about 10% less than clutter as a predictor alone. This may suggest that compression-based measures are not sufficient for explaining perceptual complexity in scene-based datasets.

3. Study 2 — The effect of semantics

The aim of our second study was to investigate the role scene semantics play in perception of scene complexity. We ask to what degree the participants' complexity ratings depend upon the semantic content of the scene. It is well accepted that inversion of stimuli results in increased difficulty processing the content of the stimuli. In the case of scenes, Epstein (Epstein et al., 2006) finds a behavioural penalty for scene inversion, showing that inversion causes a reduction in specific-scene processing ability (a reduction in PPA response), but a greater response in the generic object-processing parts of the brain (LO). In order to investigate this, we invert our scene dataset by rotating each image 180 degrees, disrupting the processing of semantic content for human observers. As with Study 1, we evaluated how our computational factors explain the perception of complexity of inverted scenes by human observers. These factors do not extract any semantic information from the scene. If they explain a considerable amount of variance inherent in inverted complexity scores, then perceived complexity for inverted images is more likely to be bottom-up driven and independent of semantic meanings. We term the dataset resulting from this study "VISC-CI" for "VISCHEMA-COMPLEXITY INVERTED".

3.1. Participants

A new group of 40 participants, aged between 18 and 65 years of age were recruited for the second study. Participants were made aware they would be viewing inverted images and their informed consent to participate was obtained prior to completing the study. This study was approved by the ethics board of the University of York.

Table 2
Results of a hierarchical regression analysis showing the contribution of each potential complexity factor towards explaining variance (coefficient of determination, R^2) in complexity ratings for our VISC-C dataset. Together, clutter and symmetry explain 36% of human complexity (disjoint sets of human ratings explain 70% corrected of each other's variance). Entries in bold indicate significant increase in variance explained. Standard error of each linear model (Lm. Std.) and residual sum of squares (RSS) are reported for completeness, and is already incorporated into reported R^2.

Model	RSS	Adjusted R^2	ΔR^2	Lm. std. error	Significance (p)
(constant)	29.35	–	–	0.19	–
Clutter (C)	**20.57**	**0.2983**	**0.2983**	**0.16**	<**0.001**
C, Entropy (E)	20.55	0.2983	0	0.16	>0.05
C, E, Symmetry (S)	**18.84**	**0.3557**	**0.0574**	**0.15**	<**0.001**
C, E, S, Openness	18.82	0.3557	0	0.15	>0.05

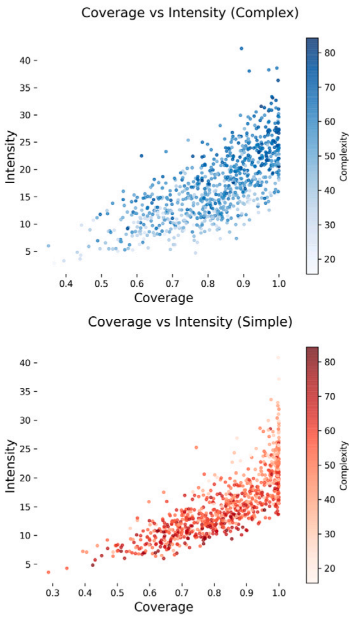

Fig. 9. Relationship between inverted scene complexity and 2d annotation metrics.

3.2. *Materials & procedure*

The images and the procedure in this study was identical to Study 1 except that the presented images were rotationally flipped, producing an inverted variant of the scene. The data analysis employed in the study was the same as reported in Study 1. We also conducted one-way ANOVA to compare the results here and in Study 1.

3.3. *Results*

Complexity scores for inverted images show a mild skew towards being rated as more complex (mean = 53.20, standard deviation = 13.31) compared to upright images. After applying a Mann–Whitney test of directionality we find this effect to be significant ($p < 0.001$), indicating inverted images are viewed as significantly more complex than their upright counterparts. There was also a lower degree of agreement in complexity scores among observers compared to upright images ($r = 0.60$, $r = 0.75$ S–B corrected). Despite these variations, complexity scores between upright scenes and inverted scenes correlate strongly together (r = 0.77), which suggests even when inverted (semantic structure disrupted) participants are still able to determine the complexity of the image (though with a lower degree of inter-participant agreement). While participants agreed to the same degree on the complex regions of inverted images as they did for upright ($P^{2D} = 0.39$), there was a lower degree of agreement between participants for the simplistic regions ($P^{2D} = 0.17$). This reflects the increased difficulty of the task, and is an initial indication that destruction of semantic structure affects complexity perception, especially in the case of determining what is simple. Participant consistency is decreased compared to Study 1, with a split of human data explaining 54% (S–B corrected) of the variance of its corresponding half. The difference between upright and inverted complexity scores is significant (ANOVA $p < 0.01$ 95% CI [0.66, 3.25]), as is the difference in consistency between upright and inverted simple annotations (ANOVA, $p < 0.001$ 95% CI [0.08, 0.094]). However, there is no difference in the consistency of the *complexity* annotations of the participants (ANOVA, $p > 0.05$ 95% CI [-0.001, 0.003]).

The two-dimensional annotation properties of the delineated regions (annotation size and coverage) correlate strongly with given complexity scores ($r = 0.66$, Fig. 9). Interestingly, we find that by inverting the scene images we caused a significant change in annotation coverage (Fig. 10), with a greater percentage of the image being indicated as complex or simple. This suggests that participants find it more difficult to identify and localise exactly what within the image is complex, or simple; defaulting to a global view of complexity for the entire image. These results imply that for images with degraded semantic information, humans fall back to lower-level, global features when perceiving complexity, but do make use of semantic content where it is present.

To evaluate the impact of low-level features on inverted complexity perception, we again make use of a hierarchical regression analysis. We find our complexity factors explain 38% of the variance in complexity scores of inverted scenes (Table 3), approaching the average human consistency of 54%. In this case, low-level classical features appear to explain more of the variance in the human ratings. Given that inverting the scene damages the semantic information present in the image, we can hypothesise that the remaining 30% of variance not captured in the case of upright scenes we observed in Study 1 is due, in part, to the semantic structure of the scene images shown. So far, these explanatory regression analyses lack any notion of semantics. As we hypothesise the *inverted* complexity scores involve a lower degree of semantics, we can ask the question of how well the final linear model drawn from the Study 1 HRA, can predict these inverted scores. If our hypothesis is correct, the linear predictor with no semantic features should be able to reasonably predict the human ratings for the inverted image set. Conducting this analysis, we find that scores predicted via the Study 1 model correlate strongly with participant scores for the inverted images ($r = 0.614$). The predicted scores explain 37.78% of the variance in the actual inverted scores, explaining a significant chunk of the 54% human variance, but based on low-level features alone. In the inverted case, the remaining 16% is likely down to individual differences, though it is possible that semantic information strong enough to survive the inversion process, contributes here. Nonetheless, much more variance is explainable by low-level features in the inverted case, than in the upright case.

3.4. *A note on semantics*

Throughout this section, and in the rest of the paper, we make several references to the "semantics" of the scene; both when discussing how humans perceive scene information, and in reference to what neural networks are extracting from images. However, exactly what is meant by semantics can vary by discipline; and is further complicated

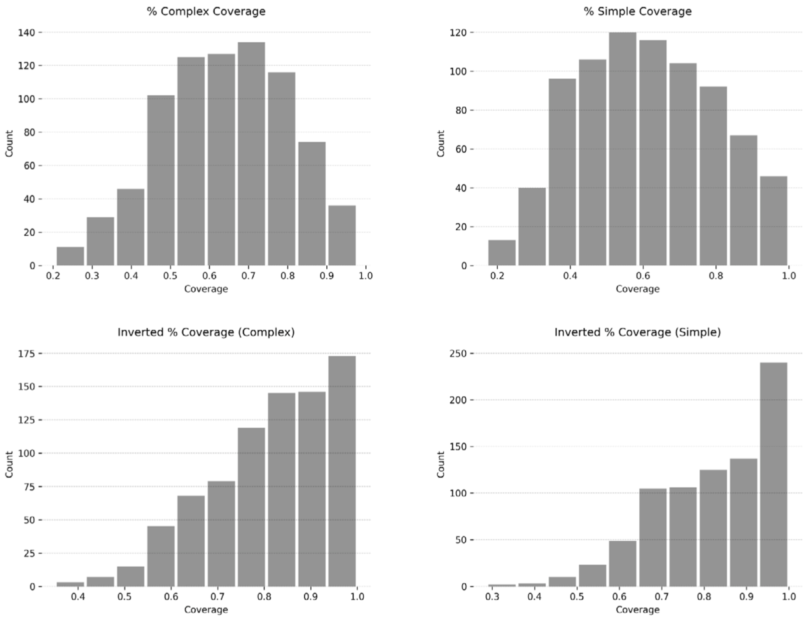

Fig. 10. Complex/simple annotation coverage for upright (VISC-C, top) and inverted (VISC-CI, bottom) scenes. Coverage shows that much more of the image is indicated as complex or simple when inverted, despite low-level textural properties remaining the same.

Table 3
Results of a hierarchical regression analysis run on human complexity ratings from the VISC-CI dataset (inverted scene images). The main contributors are clutter and symmetry (38%), with minor contribution from openness. Entries in bold indicate significant difference in variance explained. Std. Error is reported for completeness, and is already incorporated into given R^2.

Model	RSS	Adjusted R^2	ΔR^2	Lm. std. error	Significance (p)
(constant)	30.05	–	–	0.19	–
Clutter (C)	**20.59**	**0.314**	**0.314**	**0.16**	**<0.001**
C, Entropy (E)	20.59	0.313	−0.001	0.16	>0.05
C, E, Symmetry (S)	**18.80**	**0.372**	**0.059**	**0.15**	**<0.001**
C, E, S, Openness	**18.633**	**0.378**	**0.006**	**0.15**	**<0.01**

by the fact that “scene semantics” is an umbrella term that encompasses several related, yet distinct, types of scene information. Hence, to avoid confusion, here we will describe the type of semantic information relevant to complexity perception.

A useful taxonomy of different scene semantics has been put forth by Wu et al. (2014), who categorises the types of scene semantics into “gist”, scene–object relationships", and object–object relationships (so-occurrences & spatial dependency)". In this work, we define semantics as information extracted from sets of scene elements, and the associations between them, as well as the associations between these sets and the scene itself. We define the rather ambiguous term “scene elements” as shorthand for visual items present in a scene, for example: everyday objects, surfaces, or meaningful textural compositions (such as foliage, or beach pebbles). These elements are often constrained by some form of layout, or are unified in some manner (for example, individual leaves can be “unified” as foliage). Importantly, when we talk about semantics we do not mean the semantics of the individual scene elements; but collections of these elements together and their corresponding relationships. Specifically, how these collections relate to both the complexity of the scene as a whole, and the complexity of the region of the scene that these collections reside in. For example, the object–object relation includes the relationship between a bed and a floor segment, or a lamp and a cupboard face (“a surface”). Given that Wu *et al.’s* use of “object” is similar to our use of “scene element”, our definition of semantic information fits firmly into both the “object–object relations” category, as well as the “scene–object relations” category of types of semantic information, and is similar to that discussed by Hayes and Henderson (2021), in which they define semantics, in part, as “the relationships between the scene category and the objects it contains”.

4. Study 3 — Generalising to a different dataset

Study 3 examines how well our computational factors generalise to another existing more varied, image set: BOLD5000. BOLD 5000 is a dataset of 4914 images for which there is accompanying neuroimaging data primarily used for training and testing computer vision models (Chang et al., 2019). The images in the BOLD5000 dataset are an amalgamation of images drawn from several different source image sets: COCO, depicting objects (Lin et al., 2014), ImageNet, depicting diverse content of objects and scenes (Deng et al., 2009), and scene images based on categories from SUN (Xiao et al., 2010).

4.1. Participants

We recruited 1118 participants from Amazon Mechanical Turk and compensated them for their participation. We only recruited participants in Canada or the USA who had approval rates greater than or equal to 75%. The study was approved by the University of Toronto Research Ethics Board.

4.2. Materials

For the purpose of our study, we used 4914 images from BOLD5000. Each image in the dataset has a resolution of 375 x 375 pixels. These images consist of 1999 images from COCO, 1915 images from ImageNet, and 1000 scene images based on categories from the SUN image set. Images from COCO were collected to depict objects and images from ImageNet either depict objects or scenes.

Table 4
Bivariate correlation table for low-level features (BOLD5000).

	Clutter	Entropy	Openness	Symmetry
Clutter	–	0.5	−0.4	−0.57
Entropy	–	–	−0.41	−0.48
Openness	–	–	–	0.3

4.3. Procedure

For the study we collected complexity ratings from participants for selected images from the BOLD5000 image set. The study ran on each participant's computer using the Inquisit (Inquisit: https://www.millisecond.com, 0000) software. The images were pseudo-randomly assigned into groups such that each image received ratings from 50 participants. The images were presented sequentially in a random order to each participant and each participant viewed and rated 252 images. Participants provide three different ratings for each image on a 5-point Likert scale. Question 1 was: "How symmetric do you think this image is?"; Question 2: "How simple or complex is this image?"; the response options were 1 = "very simple", 2 = "simple", 3 = "neutral", 4 = "complex" and 5 = "very complex". Lastly, Question 3: "How much do you enjoy looking at this image?". Participants needed to respond to all three ratings in sequence before the next image appeared.

4.4. Data analysis

We established exclusion criteria to ensure high data quality. To detect participants always giving the same response we computed the variance of responses for each participant in a 15-rating sliding window. We excluded participants with a variance less than or equal to 0.2 on average. We also excluded participants with average variance between 0.2 and 0.5 and mean reaction time shorter than or equal to 250 ms. Data from participants who did not finish the entire study were also discarded. These criteria resulted in the exclusion of the data of 143 participants (12.8%). We continued data collection until we had 50 valid ratings per image. For purposes of this study, we examined only the complexity ratings for each of the images. Ratings for complexity were converted to z-scores separately for each participant by subtracting the mean of their responses and dividing by the standard deviation. Z-scored ratings for each image were averaged over participants for further use, giving a standardised continuous score from the original discrete ratings. We conducted a hierarchical regression analysis similar to that in Studies 1 and 2, with complexity ratings as the dependent variable and our computational factors as independent variables.

4.5. Results

First, we evaluated the consistency in human complexity ratings over the BOLD5000 dataset. On average, a random split explains 27.56% (corrected) of the variance in the other split after normalisation within each participant. The consistency in human ratings is lower in this study than in Study 1 or 2. This is most likely caused by both the high diversity in the image set, and due to the dataset being primarily object-focused, all of which might result in lower consistency across participants compared to a scene dataset that consists of commonly encountered real natural scenes. A post-hoc analysis showed that the rating consistency is higher in a subset of images that consists only of scenes, than the other, more object-focused categories. On average, a random split in COCO images explains 16.71% of the variance in the other split, a random split in ImageNet images explain 21.12% of the variance in the other split, and a random split in scene images based on SUN explains 24.81% of the variance in the other split. This may be because the scenes contain a greater degree of information than object-focused images.

A correlation table of all measures is given in Table 4, and the results of the hierarchical regression analysis are shown in Table 5. Our computational complexity factors explain approximately 11% of the variance inherent in human complexity ratings, which is close to half of the rating consistency across participants. Interestingly, this result closely matches that found in Study 1. The hierarchical regression analysis shows that all four computational factors contribute to explaining variance in human ratings. Compared with Studies 1 and 2, more of the role of entropy and openness for ratings in this image set might be explained by the higher diversity of images from BOLD5000. Nonetheless, the results here confirm results in Study 1, that our computational factors are able to explain a reasonable proportion of variance in human complexity ratings. It indicates that our analysis is generalisable to a larger and more diverged set of images. If we compute our computational metrics purely for the scene focused part of the BOLD5000 dataset, we find that we can explain 15.14% of the variance (human variance, 24.81%). Interestingly, in this case, exactly as in Study 1, only clutter and symmetry are significant contributors to the explanation of variance. The remaining 9% may be due to uncaptured semantic information.

Re-using our approach from Study 2 and employing the Study 1 linear predictor to predict BOLD5000 complexity scores, we find this model can explain 5.34% of the variance of the entire BOLD5000 dataset. This lower degree of explanation is likely due to both the Study 1 model not making use of entropy or openness, and due to the lack of semantics. Considering the bivariate correlations, for the more varied dataset, the relationship between clutter and symmetry appears to decouple, with symmetry only explaining 32% of the variance of clutter and vice-versa.

5. Modelling complexity with neural networks

In Studies 1 and 2 we established that annotation statistics for simple and complex regions extracted by human observers are strongly associated with overall global image complexity score, and that low-level computational measures explain a large proportion of variance inherent in complexity ratings. We now examine the efficacy of employing deep neural networks to predict both scene complexity scores and complexity maps. A neural network is a type of machine learning algorithm composed of an ensemble of 'artificial neurons' and the connections between them. Based originally on models of biological learning in the brain, the neural network is 'taught' the relationship between the input and output. Given a certain input and desired output, the difference in the neural networks actual output and the target output serves as an error with which the 'weights' between artificial neurons are updated. It is these weights that encode the information the neural network has learned; much like the strength of inter-neuronal connections in the brain. We select a *convolutional* neural network, a type of network well suited to image understanding, with neurons that can efficiently learn from visual input. In the context of neural networks, 'deep' ('deep neural network') refers to a network with many layers. Neural networks have seen state-of-the-art results for object detection (He et al., 2016a, 2016b), scene understanding (Gu et al., 2019), segmentation tasks (He et al., 2017), and image generation (Patashnik et al., 2021). Further we ask whether neural networks are capable of capturing the semantic component of image complexity, given their general success in semantic extraction (Saxe et al., 2019; Zhou et al., 2014). Of interest is discovering whether deep neural networks learn features which can be used in conjunction with classical clutter and symmetry features to explain human perception of image complexity. Hence to explore this, we develop a neural complexity model that can predict 2D complexity maps and scores simultaneously.

Table 5
Hierarchical Regression Results for the BOLD5000 dataset. Best explanatory model uses all factors, likely an effect of the more varied dataset, explaining 11.32% of variance in complexity ratings. These factors come close to explaining half the variance of humans over the dataset (27.56% corrected).

Model	RSS	Adjusted R^2	ΔR^2	Lm. std. error	Significance (p)
(constant)	90.994	–	–	0.14	–
Clutter (C)	88.674	0.0253	0.0253	0.13	<0.001
C, Entropy (E)	84.238	0.0739	0.0486	0.13	<0.001
C, E, Symmetry (S)	82.176	0.0964	0.0225	0.13	<0.001
C, E, S, Openness	80.625	0.1132	0.0168	0.13	<0.001

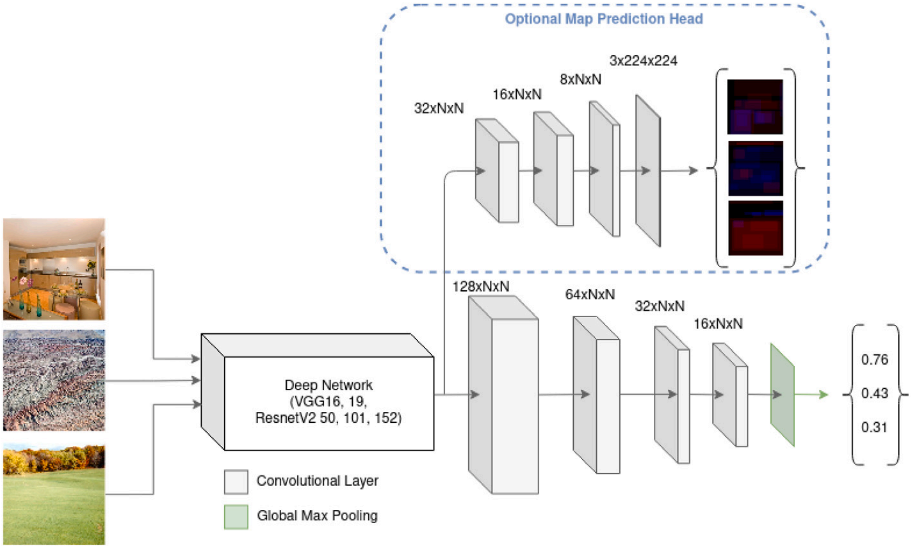

Fig. 11. Basic Complexity Prediction Architecture, with optional complexity map prediction head.

5.1. Predicting complexity scores & maps

We performed transfer learning upon five architectures: VGG16, VGG19 (Simonyan & Zisserman, 2014), and ResnetV2 (He et al., 2016a, 2016b) with 50, 101 and 152 layers, to develop a different variants of complexity prediction network, which we term 'ComplexityNet'. Each network has its classification head removed, and a four-layer convolutional regression head attached at a selected point in the network (as shown in Fig. 11). In this context, transfer learning refers to taking an already-trained neural network, and re-using the final output of these already-trained networks for some other purpose. Here, the networks are trained for object detection, but we extract the intermediate (but high-dimensional) features (which serve as an encoding of the objects and other relevant elements in the scene), and repurpose these with additional neural layers that can re-use these features for complexity detection. While there has been work towards the artificial prediction of memorability maps (Kyle-Davidson et al., 2019), two-dimensional map prediction remains unexplored for complexity prediction. To resolve this, we include an optional fully-convolutional complexity map prediction head, tasked to generate complexity maps for the input images. To evaluate the effect of network depth on complexity prediction, we systematically attach the regression head after each major processing block in each target network (results shown in Fig. 12). Each ComplexityNet variant is then trained for 100 epochs with RMSProp (learning rate: 0.0001), and cross-validated on 8 splits of the data. In cross-validation, the data is divided into a training set of 700 images, and a test set of 100 images. This is done eight times, to ensure that all data is either used for training, or testing. The data is randomised prior to being divided into train/test sets, to ensure there is no specific category weighting in either the train or test set. This cross-validation ensures the network is not trained on the images on which it computes predictions, preventing the network from simply memorising the human complexity scores for the test split of the data. From this cross-validation, we obtain predictions for every image in the VISC-C dataset. When the complexity map prediction head is enabled, the network is trained simultaneously with both scores and maps as inputs. We use the standard mean squared error for both score and map regression, and use ReLU activation functions throughout the network, aside from each output, which terminates with a sigmoid activation. The training process takes approximately six hours on a single NVidia Tesla V100 GPU.

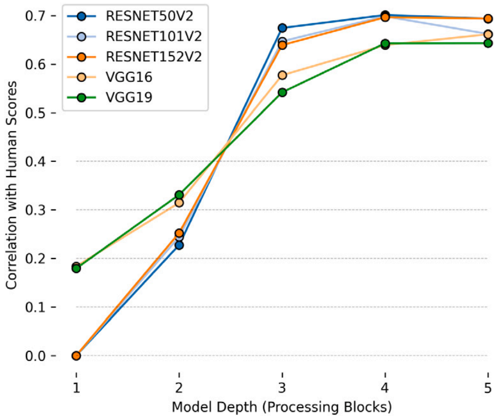

Fig. 12. Effect of network depth on complexity score prediction performance. Performance peaks in the penultimate processing block of each model, then plateaus.

Our complexity prediction model performs well at predicting complexity scores for scene images. When considering complexity map prediction, ComplexityNet achieves good performance for both scores and maps, with the best-performing model (when considering both

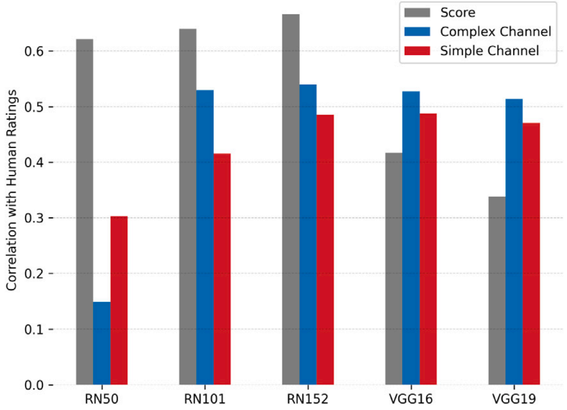

Fig. 13. Correlation with human ratings for both scores and complexity maps for different base network architectures.

scores and maps) achieving a Spearmans correlation of $\rho = 0.67$ with human scores, and generating complexity maps that correlate with human complexity maps (complex annotations: $P^{2D} = 0.54$, simple annotations: $P^{2D} = 0.49$). Samples of predicted complexity maps and their human observer-based maps can be seen in Fig. 14, and prediction results from all tested architectures in Fig. 13.

5.2. What neural networks learn about complexity

Do neural networks learn mostly low-level features, or do they extract semantic features with relations to complexity? A neural network will only learn to extract features that are relevant to the task that it is being trained to accomplish. This means that if low-level features are important to predicting complexity scores, the network is highly likely learn to extract said features. Likewise, if semantic elements are also required, the neural network will learn semantic feature extractors. It is generally understood that earlier layers in the network learn more basic, lower-level features, while more 'semantic' features are learned by the upper layers of the network. If only lower-level features were required, we would expect prediction performance to not increase as network depth increases. However, we clearly see a drastic increase in performance between early, and later, DNN layers (Fig. 12). This suggests that it is not *just* the lower-level features the network has learned to extract. To investigate this further, and determine whether neural networks learn features orthogonal to low-level computational complexity measures, we combine the previous results of our hierarchical regression analyses with the predicted score outputs from our best performing ComplexityNet (based on RESNETv2-152). If the neural network adds little additional variance explained, we assume that the neural prediction is based primarily on low level features. On the other hand, if the network can explain more variance inherent in complexity in addition to low-level features, this implies the network is learning semantic features that relate to complexity. In order to investigate this question we employed two different approaches. In one we examined the network by dissecting it (Bau et al., 2020). In the other, we combine the prediction scores of the network with our computational factors to examine how much of human complexity ratings can be explained. Network dissection allows us to 'take apart' our neural model. This allows us to determine which image features are important for complexity prediction, and to examine which image features the network is considering when predicting complexity scores for scene images. We dissect our best-performing ComplexityNet model, and examine each neuron from the final convolutional layer of the complexity prediction head, 16 neurons in total. Each neuron is assigned a set of images that best activate that neuron. Tens of thousands of varied images are presented to the network, and each neurons activation is recorded for each image. The images that cause an activation in that neuron, and the regions in that image the neuron responds to, are recorded and sorted by strength of activation.

This neural network dissection reveals the images that activate a sample of the trained neurons (shown in Fig. 15). Each set of images represent the activations of a single neuron. With this dissection we make our "black-box model" transparent, and can analyse the features that each neuron in the output layer of the network searches for in the input scene image. We find in the network an emergence of both neurons that can detect low-level repeated features, such as checkerboard patterns or lined surfaces, as well as neurons focused on semantic structures such as skies, architectural elements, and road surfaces. This is indicative of the importance of semantic information in complexity perception. Additionally, we find the emergence of neurons that appear to detect object clutter and activate strongly for images which contain large amounts of assorted objects. This reinforces prior literature suggesting that visual clutter influences perception of complexity, even inside neural networks modelled after human perception. To further explore the neural networks semantic extraction faculties, we run the same analysis as in Study 1, but with our predicted scores rather than the actual human scores. In this case, clutter and symmetry together explain 35% of the variance of the neural network scores — close to the same amount explained by those measures in the actual human scores. As with the human data, entropy and openness are not significant contributors. As clutter and symmetry can explain some of the variance of the neural network scores, this reflects the "lower-level" feature extractors learned by the neural network. However, as the low-level measures cannot explain a much *larger* proportion of the variance in the neural network scores, this is suggestive that the predicted scores additionally contain the contribution of semantic feature extractors — which lower-level features cannot capture.

By combining our computational factors with the predicted score outputs from our best performing ComplexityNet (based on RESNETv2-152) in a hierarchical regression analysis, we can explain an additional 17% of complexity score variance (on top of the 35% explained by low-level features), orthogonal to global image features such as clutter and symmetry. In total, this explains a total of 52% of the variance inherent in human complexity, more closely approaching the variance that can be explained in one set of human ratings by another randomly chosen set of human ratings (70%). This suggests that complexity perception functions as a combination of both global image features (clutter, symmetry) and semantic information (architectural details, presence of roads, or skies), and that to predict complexity accurately, both are necessary. If the neural network did not require semantic features to help predict complexity, those features would not have been learned by the network. Equally, lower-level features the network has evolved are likely to be explained by our lower-level features. Certainly, elements such as checkerboarding/lined surfaces can be detected by the entropy feature extractor, and would cause effects in both clutter and symmetry detectors. For this reason we suspect the contribution to variance explained from the neural network comes from the semantic feature extractor neurons; elements we cannot capture with our low-level measures (e.g., "there is a building present in this scene" or "there is an arrangement of objects"). It is these object–scene and object–object relations that both appear to contribute to complexity perception, and are readily extractable by a neural network. The network can certainly learn that the detection of certain objects together raises the chance of a given region to be labelled as complex; just as it can learn that other objects or elements lead to a decrease in perceptual complexity. The network is also capable of determining the relationship between object and scene: it has been proven that in scene detection networks, neural object detectors arise automatically (Zhou et al., 2014) *even if the network is not trained to find objects*. Evidently, these objects reveal classification clues about the scene. Given this, it is highly likely that the relation between scene and object can be co-opted for complexity prediction also.

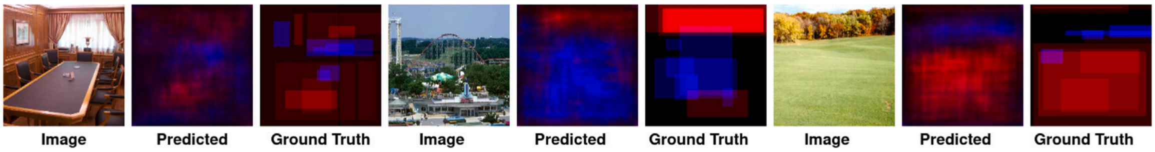

Fig. 14. Examples of predicted complexity maps and their ground-truth counterparts.

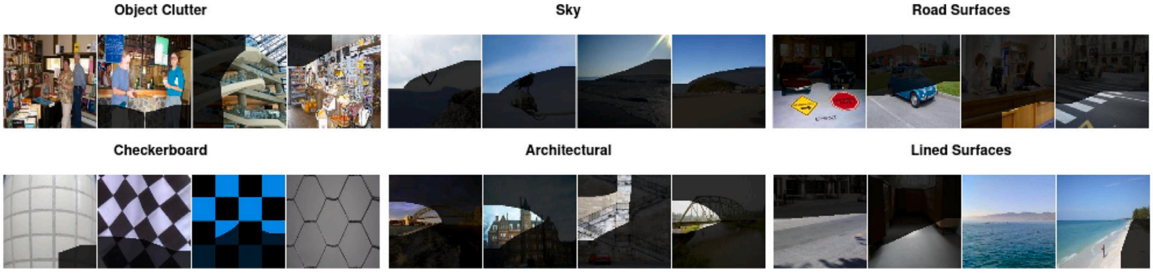

Fig. 15. Images which activate a sample of neurons from the final layer of a complexity prediction network. The network appears to learn both low-level and semantic features. Each set of images represent the category of images that best activate a singular neuron. The darkened areas in each image do not cause a strong activation for that neuron.

Discussion

Prior work has shown that human observers are capable of evaluating the complexity of images in general and show good agreement in this evaluation. However, most complexity datasets are either small (Corchs et al., 2016b), consist of simple images designed to investigate low-level processing (polygons, line drawings) (Heaps & Handel, 1999; Snodgrass & Vanderwart, 1980), or contain a mixture of images with only a small scene component (Saraee et al., 2020). It remains unclear exactly which features contribute to human perception and subsequent evaluation of the complexity of scenes. What is more, prior work tends to treat image complexity as a single rating for the whole image, which may obscure details on how complexity varies across a scene. In these cases, complexity is often defined in terms of the count of elements in a stimulus, variance in element size and element type, self-similarity, or even the number of bends in a polygon. These attempts to define an item’s visual complexity are complicated further by the introduction of an element of “order" drawn from aesthetics research, which describes the degree of organisation or disorganisation in a stimulus. Order and complexity are not hypothesised to lay on the same singular dimension, with order at one end, and complexity at the other. Instead, each is thought to lie on their own separate descriptive axis, yet with some degree of interplay between them (Van Geert & Wagemans, 2020). It remains unclear whether these definitions carry over to scene images with the same explanatory power. In our studies, we have the benefit of two dimensional complexity information. Empirically we can find examples of complex regions containing many objects, and complex regions containing textural variations. Even simple images can contain complex regions, suggesting that for scenes, it is not quite as simple as element count or variation. Nonetheless, as we have shown in Study 1, when asked to simply rate complexity, humans are highly consistent, requiring no instruction on how to evaluate complexity. This implies a commonality in how complexity is perceived: the same visual and mental machinery is being employed across the human observers. And since this commonality largely exceeds the level of agreement in aesthetic judgement (Brielmann & Pelli, 2019), understanding the common mechanism in perceived complexity grants us the potential access to quantify aesthetics.

In turn, this implies that an image to a certain extent has its own intrinsic level of complexity, and that, on average, most people will perceive the complexity of that image in a similar fashion. Our data suggests a two-dimensional space that defines human perception of complexity. An image’s complexity is perceived via a combination of low-level image features that capture textural elements (e.g., varied patterns, such as foliage), and high-level semantic elements (objects upon a desk). This allows for perception of complexity for both semantically impoverished stimuli (textures), and allows semantic understanding to influence what is perceived as complex. Additionally, we suggest that image complexity is not a “global property” of a scene, but localised to individual regions within the scene which contribute to the overall rating given by participants. This implies that an image rated as simple is not necessarily simple everywhere, there are regions of complexity within simple images, and vice-versa. Thus, while more complex scenes may contain a higher count of elements, it is not necessarily all the elements that contribute to the complexity.

In this paper we began by considering how effective the psychologically grounded lower-level features explain the variance in human complexity perception for scene images. In Study 1 we find that participant complexity ratings are both highly consistent, and correlate strongly with the two-dimensional annotations given by the participants. We selected four potential measures that either have been shown, or hypothesised, to relate to human perception of complexity in prior research. We employed computational methods for operationalising clutter and openness based on work from Olivia et al. (2004), and entropy based on its frequent appearance in the literature, and as it can capture informational content. Finally, we computationally operationalise local symmetry, under the hypothesis that symmetry can reveal locally redundant information, and hence may be related to complexity perception. Our findings indicate that we can describe a portion of human variance in perceived scene complexity (36%) with our selected low-level measures; in the case of upright images, clutter and symmetry. It appears that while low-level features can explain a reasonable portion of variance in scene complexity, they do not capture all of it, leaving a significant portion unexplained. This results in a notable difference between human performance, and the performance of image-statistic based measures. It appears evident that scene complexity is not determined purely by the low-level statistics present in the scene. However, upon scene inversion (Study 2), this difference between human performance and that of low-level statistics significantly diminishes. Given that low-level statistics tend to remain the same when the scene is inverted (for example, the number of edges does not change, nor the overall Shannon entropy, or our patch symmetry measure), it is reasonable to hypothesise that some other factor is affecting the participants perception of scene complexity. As image inversion is associated with inducing difficulties in processing the semantics of that image we conclude that there is another dimension, or level, at which complexity is perceived: a more semantic dimension. Therefore, we propose that scene complexity perception is based upon the results of processing of two different levels of information (“dual information”), that flexibly work together to allow for judgements of complexity to be made on a wide variety of different stimuli. These levels are: 1. Extraction of the low-level features present in the image (clutter and symmetry in this work) and 2. Extraction of the semantic elements and meaning present in the scene.

For example, a scene containing a laptop computer may be considered complex both due to the texture and frequency variation present, and because a laptop computer may be considered a “structured element” that serves as a unit of perception. Upon scene inversion, the laptop computer becomes more difficult to process as said “structured element”, and hence the semantics of said object become harder to extract, and hence has a reduced impact on the perception of complexity. Between the upright and inverted images the low-level features that are relevant to complexity perception appear to be shared. However, the question remains why these features relate to complexity perception. From the aesthetics viewpoint, one can reasonably frame clutter as a measure of complexity, and symmetry as a measure of order. In the upright case, both these measures contribute towards explaining the variance of human scores. This suggests that, as aesthetics literature proposes (Van Geert & Wagemans, 2020), that there is indeed an interrelation between “order” and “complexity”, and that estimations of both are required to approximate human complexity perception. However, as these are both low-level measures, this suggests that this complex interplay occurs in the lower-level features of the image, at least for scene images. Additionally, it is hypothesised that order is not equivalent to simplicity, yet we find that higher levels of symmetry, a possible measure of order, is associated with a scene being viewed as more simplistic; and is directly (though not fully) opposed to clutter, a measure of complexity. Indeed, clutter can explain roughly 50% of the variance in symmetry and vice-versa (naturally, more clutter generally means less symmetry), revealing that while there is certainly an interrelationship between these potential measures of order and complexity, they also contribute to complexity perception in their own separate fashions. This interplay may result from our choice of local symmetry, rather than global mirror symmetry. An image that is more globally symmetric may be more ordered, but at the level of local symmetry both global disorder and local order is possible (e.g. randomly thrown paperclips are still locally symmetric about some axis). How this extrapolates to the more complex objects that make up scenes remains unknown, but there may be a need to consider how order and complexity contribute to each other at the local level, and determine what effect this has on the global image. Certainly we find in the case of two-dimensional complexity annotations that there is strong variation across a scene with regards to what is considered complex or simple.

Does this notion of order and complexity hold for more varied stimuli? To answer this question, in Study 3 we investigate a large dataset of varied images (BOLD5000), only a small proportion of which are scenes. While this dataset lacks two-dimensional annotations, it does have complexity ratings for each image. The results for BOLD5000 mirror those from Study 1, showing that low-level features work well for explaining the variance in human complexity ratings across this dataset, though a portion remains uncaptured. However, the overall participant consistency is significantly lower than that of our pure scene dataset. This may be because the dataset itself contains a much broader set of image types, including a vast array of different object photographs, and while the scene component consists of a broad variety of categories; each category consists of roughly four images (VISC-C contains 100 scenes per category). Additionally, VISC-C was specifically chosen to not include objects, whereas the BOLD5000 dataset does include object-focused images. We use these datasets together to evaluate our complexity measures: VISC-C provides depth and two-dimensional data, BOLD5000 has less depth, but significant variety in image types. It is hence not surprising that we find all four of our computational factors become significant over the BOLD5000 dataset, due to the variety in image types (for example, entropy appears relevant for objects, but not scenes). Furthermore, if we compute our measures only on the scene part of the dataset, we find we can explain 15.14% of the variance, with only clutter and symmetry contributing significantly. This is a proportion of, but not all of, the human variance of (corrected) 24.81%; the remaining proportion may be explainable by semantics. A natural consequence of entertaining the “dual information”-based complexity perception is that it helps to explain why humans can consistently evaluate the complexity of different forms of stimuli, and yet each stimuli appears to require different objective complexity measures. In the case of simple line/polygon stimuli, and texture images, only low-level features are necessary; and exactly which low-level features are needed varies, depending upon the type of stimuli. For more complex stimuli, such as object and scene images, semantics begins to play a role in participant judgements of complexity. This may additionally explain why participants are less consistent for inverted scenes, as there is less information that they can draw upon in order to generate complexity ratings, due to the missing semantic element.

The dual information framework also explains why two-dimensional complexity annotations are significantly larger in the inverted case than in the upright case. Generally, regions containing low-level features are being annotated, rather than specific semantic details. However, it is unlikely that these two processing streams are completely separate, and there is likely to be some degree of interplay between them: a scene image with a lot of semantic content may also contain more low-level features. This does not necessarily work the other way around; an image with a lot of textural variation does not need to contain more semantic content (e.g., an image of tropical rainforest canopy or a Jackson Polock paining). This proposed framework is supported both by our own findings, and additionally appears to have supporting evidence from neuro-imaging studies. The work of Güçlütürk et al. (2018) finds that representations of stimulus complexity appear both in the early visual cortex, and in later areas such as the PPA, suggestive that complexity is processed both at early and later stages throughout the visual stream.

Given we hypothesise that semantics play a role, how might said semantics be extracted for potential analysis? Notably, classic image processing techniques lack any concept of “semantics”, instead focusing on low-level image statistics. Older work in computer vision has explored scene classification and object recognition in a similar fashion to current perceptual complexity work, through the use of combinations of hand-picked lower-level features. Invariably, these methods have been far surpassed by the introduction of deep neural networks, machines which learn a highly powerful, compressed representation of the input. This representation is often considered to have extracted the “semantics” – or meaning – of the input, lending the network its powerful scene classification abilities. However, there are types of semantic information a neural network can capture, and types that it cannot. A deep learning model cannot capture the singular meaning of an object present in the scene, if there is no relevant training data that it can learn from to ascribe that meaning. For example, upon detecting a coffee cup, it does not “know” that a coffee cup can be used for the action of drinking. This is a natural consequence of the training data; drinking is an action, and our network is to be trained upon static images of scenes with no action labelling. To extract this kind of semantic information requires a different task (and different architecture) altogether. However, the network can certainly learn that the presence of the “hoop” and the “cylinder” along with a hand grasping, or facial proximity, or the fact it is simply placed on a coaster, makes the detected object more likely to be a coffee cup. This suggests the network is capable of learning an understanding of the “object–object relation” type of semantic information. Neural networks can also learn to employ generic scene cues; an ambiguous metal and glass structure on a road is more likely to be declared a car — but against a blue backdrop; a plane. Interestingly enough, this kind of scene–object linking can cause the network to misclassify; objects in unusual contexts are harder for the network to correctly detect. A neural network can learn to extract object–object and scene–object relationships by learning to detect which features, in concert together, promote or impede scene complexity as perceived by humans. In this fashion, the network will have learned to extract relevant semantic information from the scene. It is this information, the relationship between collections of scene elements, that contributes to the "high-level" features relevant to complexity perception. This extracted semantics is what we hypothesise as being part of the reason for a specific region being labelled as "complex" or "simple" by a human, and which is naturally unextractable by low-level computer vision techniques. Research exploring the internal representations of neural networks (Bau et al., 2020; Zhou et al., 2014) have shown that it has learned to individually extract semantically relevant features from the scenes it is shown: from picking out wings and beaks to identify a bird, to detecting chairs and tables to identify a living room. Notably, the network learns that it is the combinations of these elements (i.e, the association between scene elements) that identifies the bird, or the living room. This representation has also been shown to correlate with fMRI activations from both animal and human visual cortices (Cichy et al., 2016; Horikawa et al., 2019; Khaligh-Razavi & Kriegeskorte, 2014), and relate to human semantical development (Saxe et al., 2019). Hence, we trained and evaluated ComplexityNet, and combined its "semantic features" with our global image features, and found this significantly boosts the level of variance in human complexity ratings we can explain. Dissecting the network (Bau et al., 2020) revealed both low-level (checkerboards, lined surfaces) and semantic features (sky, architectural details) extractors arise in the neurons of such a model.

These results are consistent with what we observe in Study 1-3, that both low-level image features and semantic structure appear necessary to model human complexity perception. It is for this reason we suggest that our neural network model is adding the trait of "semantics" to our analysis, and hence the reason that the network performs so well at predicting human complexity evaluation for scenes. We observe that while the neural network does learn low level feature extractors, indicating their usefulness for complexity perception, it also learns semantic feature extractors. Neural networks are unlikely to learn to extract features that are not useful for the task they are given (given they tend to take the simplest possible approach towards solving a problem). If semantic information was not valuable to the network, feature extractors that capture it would not have been learned. It would additionally be highly surprising if the network found semantic information useful, but the human visual system does not; given that the general robustness, capabilities, and complexity of the visual system far surpass the average neural network. There is the additional caveat that neural networks can extract some types of semantic information (object–object, and object–scene relations), but certainly not every type of semantic content. This implies that the actual semantic information humans make use of for complexity perception is highly likely to be richer than the portion we have captured here via neural network. While neural networks are unable to extract a semantic representation as rich as a humans, the information they can extract appears powerful enough to improve complexity prediction performance beyond that which can be explained by low-level features.

There are many potential avenues for further research. We primarily focus on a collection of 800 scenes, with a brief foray into investigating a more object-focused dataset. There is plenty of opportunity to explore larger and different datasets, especially in a two-dimensional fashion. The relation between complexity and other image properties, notably aesthetics, remains unclear — investigating whether aesthetically pleasing image regions are complex, or simple, or both, may reveal additional clues as to how the visual system processes stimuli. The same question could be asked about image memorability: are memorable images more complex? So far, we have found evidence indicative of a dual information complexity framework only for scene images, though it seems unlikely that semantics only play a role in estimating the complexity of scenes. Additionally, we did not explore the "flexible interplay" between semantics and lower-level features; it is unknown exactly how one trades off against the other. A texturally complex image may lack semantics, but a semantically rich image is unlikely to lack low-level features, making it difficult to gauge the impact of semantics alone. Does the presence of more semantics lead to, and enable, less use of lower-level features for complexity perception? Or is the information additive, building first from low-level, and then semantics?

In conclusion, in this work we specifically set out to both investigate scene complexity as a property that may not be constant across a scene, and to determine how scene complexity itself can be explained; including evaluating the potential effect of semantics upon complexity. Our findings reveal evidence for a potential dual information framework that describes how humans perceive complexity, based both on extracting low-level features, and semantic meaning. Where semantic information is impaired, lower-level features present in the image enable the evaluation of that image's complexity, and in the cases where semantic information is present, humans are capable of drawing on that information in addition to the lower-level image statistics.

CRediT authorship contribution statement

Cameron Kyle-Davidson: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Visualisation, Writing – original draft. **Elizabeth Yue Zhou:** Validation, Investigation, Formal analysis, Resources, Data curation, Writing – review & editing. **Dirk B. Walther:** Investigation, Data curation, Writing – review & editing, Supervision. **Adrian G. Bors:** Software, Writing – review & editing. **Karla K. Evans:** Conceptualization, Methodology, Writing – original draft, Writing– review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset link was given in the "attach files" stage. Code is planned to be released on github.

VISCHEMA-COMPLEXITY (Original data) (Mendeley Data)

Acknowledgements

This work was supported by a fellowship from the Engineering and Physical Sciences Research Council to C.K.D and a grant from CRUK & EPSRC to K.K.E. All authors approved the version of the manuscript to be published.

Supplementary material

The VISCHEMA-COMPLEXITY (VISC-C) dataset can be found at the following repository: https://data.mendeley.com/datasets/7943zgtsr7/1.

This dataset contains upright (Study 1) and inverted (Study 2) scene images, their corresponding two-dimensional complexity maps, and a complexity score for each image.

References

Akagunduz, E., Bors, A. G., & Evans, K. K. (2019). Defining image memorability using the visual memory schema. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(9), 2165–2178.

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., & Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, http://dx.doi.org/10.1073/pnas.1907375117.

Birkhoff, G. D. (1933). *Aesthetic measure*. Harvard University Press.

Brielmann, A. A., & Pelli, D. G. (2019). Intense beauty requires intense pleasure. *Frontiers in Psychology*, *10*, 2420.

Cardaci, M., Di Gesù, V., Petrou, M., & Tabacchi, M. E. (2009). A fuzzy approach to the evaluation of image complexity. *Fuzzy Sets and Systems*, *160*(10), 1474–1484.

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, *6*(1), 1–18.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 1–13.

Ciocca, G., Corchs, S., Gasparini, F., Bricolo, E., & Tebano, R. (2015). Does color influence image complexity perception? In *International workshop on computational color imaging* (pp. 139–148). Springer.

Corchs, S. E., Ciocca, G., Bricolo, E., & Gasparini, F. (2016). Predicting complexity perception of real world images. *PLoS One*, *11*(6), Article e0157986.

Corchs, S., Ciocca, G., & Gasparini, F. (2016). Human perception of image complexity: Real scenes versus texture patches. *Journal of Alzheimer's Disease*, *53*, s51. http://dx.doi.org/10.3233/JAD-169002, Abstracts for the Second International Meeting of the Milan Center for Neuroscience (Neuromi): Prediction and Prevention of Dementia: New Hope (Milan, July 6–8, 2016).

Day, H. (1968). The importance of symmetry and complexity in the evaluation of complexity, interest and pleasingness. *Psychonomic Science*, *10*(10), 339–340.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Deng, L., & Poole, M. S. (2012). Aesthetic design of e-commerce web pages–Webpage complexity, order and preference. *Electronic Commerce Research and Applications*, *11*(4), 420–440.

Epstein, R. A., Higgins, J. S., Parker, W., Aguirre, G. K., & Cooperman, S. (2006). Cortical correlates of face and scene inversion: A comparison. *Neuropsychologia*, *44*(7), 1145–1158.

García-Pérez, M. A., & Alcalá-Quintana, R. (2011). Interval bias in 2AFC detection tasks: Sorting out the artifacts. *Attention, Perception, & Psychophysics*, *73*(7), 2332–2352.

Gu, Y., Wang, Y., & Li, Y. (2019). A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection. *Applied Sciences*, *9*(10), 2110.

Güçlütürk, Y., Güçlü, U., van Gerven, M., & van Lier, R. (2018). Representations of naturalistic stimulus complexity in early and associative visual and auditory cortices. *Scientific Reports*, *8*(1), 1–16.

Hauagge, D. C., & Snavely, N. (2012). Image matching using local symmetry features. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 206–213). IEEE.

Hayes, T. R., & Henderson, J. M. (2021). Looking for semantic similarity: What a vector-space model of semantics can tell us about attention in real-world scenes. *Psychological Science*, *32*(8), 1262–1270.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969).

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).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In *European conference on computer vision (vol. 9908)* (pp. 630–645). Springer.

Heaps, C., & Handel, S. (1999). Similarity and features of natural textures. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(2), 299.

Horikawa, T., Aoki, S. C., Tsukamoto, M., & Kamitani, Y. (2019). Characterization of deep neural network features by decodability from human brain activity. *Scientific Data*, *6*(1), 1–12.

Inquisit: https://www. millisecond. com 0000. https://www.millisecond.com.

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.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), Article e1003915.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition ofinformation'. *Problems of Information Transmission*, *1*(1), 1–7.

Kyle-Davidson, C., Bors, A., & Evans, K. (2019). Predicting visual memory schemas with variational autoencoders. In *British machine vision conference*.

Landwehr, J. R., Labroo, A. A., & Herrmann, A. (2011). Gut liking for the ordinary: Incorporating design fluency improves automobile sales forecasts. *Marketing Science*, *30*(3), 416–429.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European conference on computer vision (vol. 8693)* (pp. 740–755). Springer.

Liu, Y., Hel-Or, H., & Kaplan, C. S. (2010). *Computational symmetry in computer vision and computer graphics*. Now publishers Inc.

Machilsen, B., Pauwels, M., & Wagemans, J. (2009). The role of vertical mirror symmetry in visual shape detection. *Journal of Vision*, *9*(12), 11.

Mayer, S., & Landwehr, J. R. (2018). Quantifying visual aesthetics based on processing fluency theory: Four algorithmic measures for antecedents of aesthetic preferences. *Psychology of Aesthetics, Creativity, and the Arts*, *12*(4), 399.

Nagle, F., & Lavie, N. (2020). Predicting human complexity perception of real-world scenes. *Royal Society Open Science*, *7*(5), Article 191487.

Neri, P. (2014). Semantic control of feature extraction from natural scenes. *Journal of Neuroscience*, *34*(6), 2374–2388.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Olivia, A., Mack, M. L., Shrestha, M., & Peeper, A. (2004). Identifying the perceptual dimensions of visual complexity of scenes. In *Proceedings of the annual meeting of the cognitive science society (vol. 26)*.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., & Lischinski, D. (2021). Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2085–2094).

Patraucean, V., Grompone von Gioi, R., & Ovsjanikov, M. (2013). Detection of mirror-symmetric image patches. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 211–216).

Prolific: https://www. prolific. co/ 2022. https://www.prolific.co/.

Rigau, J., Feixas, M., & Sbert, M. (2007). Conceptualizing birkhoff's aesthetic measure using shannon entropy and kolmogorov complexity. In *Computational aesthetics* (pp. 105–112).

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, *7*(2), 17.

Ross, M. G., & Oliva, A. (2010). Estimating perception of scene layout properties from global image features. *Journal of Vision*, *10*(1), 2.

Saraee, E., Jalal, M., & Betke, M. (2020). Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding*, *195*, Article 102949.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, *6*(2), 174.

Treder, M. S. (2010). Behind the looking-glass: A review on human symmetry perception. *Symmetry*, *2*(3), 1510–1543.

Ulrich, R., & Vorberg, D. (2009). Estimating the difference limen in 2AFC tasks: Pitfalls and improved estimators. *Attention, Perception, & Psychophysics*, *71*(6), 1219–1227.

Van Geert, E., & Wagemans, J. (2020). Order, complexity, and aesthetic appreciation. *Psychology of Aesthetics, Creativity, and the Arts*, *14*(2), 135.

Wagemans, J. (1995). Detection of visual symmetries. In *Spatial vision 9* (pp. 9–32). Psychology Press.

Wagemans, J. (1997). Characteristics and models of human symmetry detection. *Trends in Cognitive Sciences*, *1*(9), 346–352.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of Neuroscience*, *29*(34), 10573–10581.

Wang, Z., Duff, B. R., & Clayton, R. B. (2018). Establishing a factor model for aesthetic preference for visual complexity of brand logo. *Journal of Current Issues & Research in Advertising*, *39*(1), 83–100.

Wu, C.-C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*, *5*, 54.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 3485–3492). IEEE.

Yu, H., & Winkler, S. (2013). Image complexity and spatial information. In *2013 fifth international workshop on quality of multimedia experience* (pp. 12–17). IEEE.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014). Object detectors emerge in deep scene cnns. arXiv preprint arXiv:1412.6856.