What Makes a Visualization Image Complex?

Mengdi Chu, Zefeng Qiu, Meng Ling, Shuning Jiang, Robert S. Laramee, Michael Sedlmair, and Jian Chen

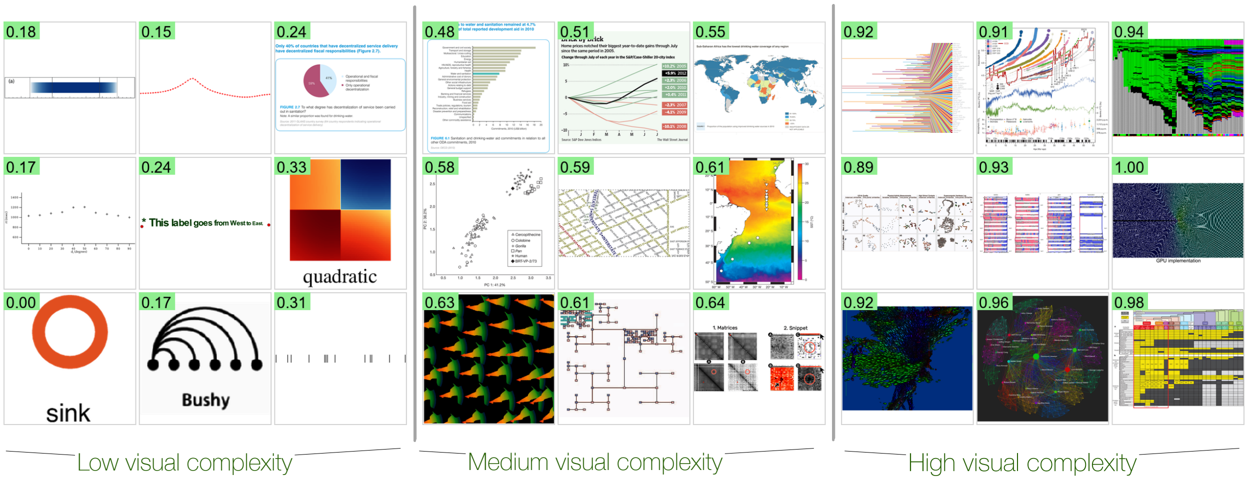


Fig. 1: Examples of low, medium, and high visual complexity (VC) images from the VisComplexity2K dataset. VC scores labeled in the upper-left corner are gathered from human observer ratings (least to most in the 9 image blocks from top left to bottom right.)

Abstract—We investigate the perceived visual complexity (VC) in data visualizations using objective image-based metrics. We collected VC scores through a large-scale crowdsourcing experiment involving 349 participants and 1,800 visualization images. We then examined how these scores align with 12 image-based metrics spanning information-theoretic, clutter, color, and our two new object-based metrics (meaningful-color-count (MeC) and text-to-ink ratio (TiR)). Our results show that both low-level pixel-based image properties and the high-level elements affect perceived VC in visualization images; The number of corners and distinct colors are robust metrics across visualizations. Second, feature congestion, a statistical information-theoretic metric capturing color and texture patterns, is the strongest predictor of perceived complexity in visualizations rich in the same color/texture stimuli; edge density effectively explains VC in node-link diagrams. Additionally, we observe a bell-curve effect for texts: increasing TiR initially reduces complexity, reaching an optimal point, beyond which further text increases VC. Our quantification model is also interpretable—enabling metric-based explanations—grounded in the VisComplexity2K dataset, bridging computational metrics with human perceptual responses at both pixel and object levels. The preregistration is available at `osf.io/5xe8a`. The dataset and analysis code can be accessed at `osf.io/bdet6`.

Index Terms—Perceived visual complexity, image-based metrics, scene-like, text-ink-ratio, meaningful-color-count

1 Introduction

Understanding the visual complexity (VC) of an image is crucial in a variety of applications. It can affect the ability of an observer in understanding an environment [48, 128] or navigate through it [45]. VC affects visibility [37], aesthetics [47, 114], clutter [2, 118, 119], and thus understandability [35] and memory [134]. VC also impacts visual search [119], design [149], and metaphorical thinking [9].

Evaluating perceived VC can also contribute to foundational research in visualization. VC has been defined as "*the amount of detail or intricacy*" [107, 128]. Most evaluations use simple images, that are less complex than those found in practical, real-world uses [155]. In contrast, practitioners foraging information need to handle in realistic visual design to achieve more generalizable outcomes [149]. This gap introduces at least two evaluation challenges. First, the vast majority of visualization images contain an array of features, patterns, and contexts. We have a limited understanding of their feature space. Even though high-level features (e. g., number of elements [86], text [49], and the scene gist [99]) and a computational equivalent (e. g., feature congestion [118]) contribute to VC, studied previously primarily for natural scenes [129], it is unclear if any of the existing metrics and theories are generalizable across domains. Second, we also lack of understanding of observers' experiences when examining these images.

Our goal is to systematically identify some of the image features that influence VC. In this work, we explore how VC can be viewed through both perceived and objective lenses. To this end, we draw inspiration from the extensive work in vision science that collects large-scale absolute VC scores [123]. Furthermore, we draw on a precedent for such metric-based explanations [107], to study what metrics can represent VC judgment. In the longer term, measuring and explaining human judgment with an extensive set of metrics can reverse engineer high-level multifaceted understanding of the underlying mechanisms [72].

Method (Figure 2). We begin with an extensive literature review on subjective VC factors in Table 1 to guide our selection of 12 objective metrics in Table 2. Some metrics, such as information-theoretic metrics, capture mathematical uncertainty, while others are sensitive to clutter

- *Mengdi Chu, Zefeng Qiu, Meng Ling, Shuning Jiang, and Jian Chen are with Ohio State University. E-mail: {chu.752, qiu.573, ling.253, jiang.2126, chen.8028}@osu.edu.*
- *Robert S. Laramee is with University of Nottingham. E-mail: R.S.Laramee@swansea.ac.uk.*
- *Michael Sedlmair is with University of Stuttgart. E-mail: michael.sedlmair@visus.uni-stuttgart.de.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

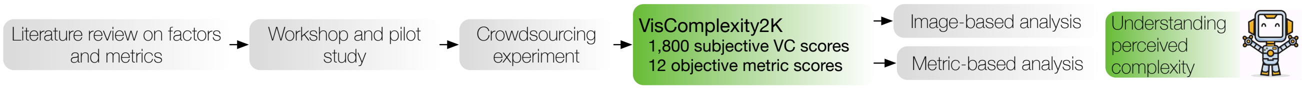


Fig. 2: **Overview of our quantification study method of perceived visual complexity.** We used objective quality metric to measure the subjective high-order perception. Activities (in gray) and outcomes (in green).

(feature congestion) [119], shape and structure, or color. We further introduce two new metrics to represent visualization elements: (1) MeC (the number of meaningful colors), to represent the number of component parts in an image—by applying color-as-object-boundaries [41]; (2) TiR (text-ink-ratio)—*the proportion of annotation in a visualization image*, to investigate how context influences VC, given its prominent role in visualization [109] and its power to assist viewers to comprehend concepts that are otherwise difficult to interpret. We collect perceived VC scores of 1,800 visualization images and then use image-based feature metrics to capture VC judgment. While a precedent for metric-based explanations of specific visualizations (e. g., networks [107]) and stimulus-specific clutter (e. g., color and texture [119]) exists within our community, measuring and explaining human perceived VC using an extensive set of metrics offers a pathway towards finding low-dimensional proxy to engineer high-dimensional VC.

Results. We collected perceived VC scores from 1,800 images (Figure 1). Our analysis reveals that a range of metrics contribute to explaining perceived VC. Among them, edge density (ED) emerges as a dominant factor across visualization images. Object-level metrics such as color count (MeC) and text-ink-ratio (TiR) capture high-level aspects of visual complexity but have little correlation with pixel-based metrics. A notable finding is that clutter-based feature congestion (FC) [118, 119], an information-theoretic measure adapted to account for statistical scene structures of color and texture, proves particularly effective in distinguishing VC in visualizations rich in the same color and texture variations. Finally, we observe that the amount of annotation text, as measured by TiR, has a non-linear U-shaped relation with perceived VC: moderate levels of text reduce VC while excessive annotations increase it, extending the current literature.

In summary, our main **contributions** are: (1) A crowdsourcing study collecting perceived VC scores for 1,800 visualization images from 349 participants. (2) A metrics study that analyzes how 10 existing and 2 new metrics help us interpret perceived VC. (3) We provide the human and metric scores from both studies, along with the 1,800 visualization images, in a dataset called VisComplexity2K, openly available at `osf.io/bdet6`.

2 Background and Related Work

This section reviews the definitions of VC and the associated metrics. As outlined in Table 1, our literature review identifies five key categories, which serve as the foundation for organizing the metrics summarized in Table 2, largely from vision science.

2.1 Definitions of Visual Complexity

Defining visual complexity is challenging due to its subjective and multifaceted nature. VC arguably plays a key role in comprehension. For example, vision science has broadly explored human understanding of VC, and defined it as "*the degree of difficulty in providing a verbal description of an image*" [53,99], related to *recall of the (natural) scene* or to make sense of the structure of it [99], or the naming [36]. A key factor contributing to VC is the representation of structure and attributes, such as clutter. Oliva et al. [99] studied VC as both a high-level "*scene gist*" and a set of low-level features, finding that humans adapt their perception based on the task. VC can be "*principally represented by perceptions of quantifying objects, clutter, openness, symmetry, organization and variety of colors*." These structural elements influence how people later remember or forget about an environment [74].

In visualization, a core challenge our community faces is understanding how human observers interpret visualizations in terms of readability, legibility, and understanding [14, 17]. *Perceived* experience often refers

Table 1: **Five factors that affect perceived visual complexity (VC).** VC was studied along several dimensions, and reported in numerous controlled studies, mainly under natural scene settings. This is a compact summary of the Apdx. Table 3 which is a complete literature overview.

Factors	Scope	Source
Information theoretic	Information amount in a mathematical sense [126].	[22, 65, 67, 92, 99, 123, 149]
Clutter	There exists "*little room for a new feature to draw attention*" [119].	[99, 119, 123]
Color	The variety of colors (hue, saturation, and brightness), the level of continuity, intensity, and range of colors.	[27, 38, 59, 87, 92, 107, 119]
Shape and Structure	The variety and intricacy of edges and turns within an image, encompassing the contours and boundaries and Structural clarity (e.g., regularity, heterogeneity, region segmentation by color or textures).	[5, 27, 59, 77, 87, 92, 100, 107, 110, 132]
Object	The number (set-size) of unique objects (e. g., graphical elements, color and other ensemble attributes [109], and text).	[16, 51, 99, 128]

to the subjective impression or feeling evoked by a visualization, regardless of tasks or familiarity [47]. This understanding is influenced by intrinsic factors such as the nature of the data [35], the task at hand, and high-order perceptual and cognitive processes [155], all of which may be influenced by the same set of design variables that govern VCs [33]. As a result, VC has been described in a variety of ways across the literature: (1) as "*a form of visual uncertainty*" [22, 35], where data variations enable its evaluation through information-theoretical measures [22, 67]; (2) as "*a measure of the degree of difficulty for humans to interpret a visual presentation correctly*"; (3) as feature congestion (FC), framed as the antonym of *'simplicity'* [86], or "*allowing little room to put in more information*" [118]; (4) as "*the amount of detail or intricacy*" [107, 128]; and (5) as a spatial metaphor, where Borgo et al. [9] described complexity as "*a direct function of the degree of perceivable structure, variety of parts and separation of parts vs. their conceptualization as a whole*".

Our work builds on these definitions and refocuses them within the context of visualization images, adopting the definition from Purchase et al. [107], which describes VC as "the amount of detail or intricacy".

2.2 Five Factors that Govern Visual Complexity

We draw inspiration from the extensive research on characterizing complexity to select a set of metrics that represent these factors.

2.2.1 Data-driven Information Theoretic Measure

The information-theoretic measure, often computed using Shannon's entropy [126], is a purely mathematical representation of pixel data in images. In data visualization, Yang-Peläez and Flowers (Y-F) [154] extended Shannon's classic entropy by partitioning 'information' into three types: *syntactic* (marks, lines, regions, and structures thereof), *semantics* (meaning of those marks or other syntactic information), and *pragmatic* (the value or usefulness of that information). The number of *bits* is subsequently derived from *marks* by searching neighboring pixels. Every bit/pixel carries an equal weight representing the meaning of the message. Later, Chen and Jänicke [22] (C-J) conceptualize the Y-F method to take data variations in every stage of the visualization workflow, by *weighting* the neighboring pixels to compute the infor-

Table 2: Twelve objective image metrics associated with factors that may influence perceived VC. The first 10 metrics are from the literature where their influence on VC was previously reported. We define two new object-level metrics to quantify the number of elements in a visualization image.

		Metrics	Description	Equation	
Info.-theoretic (O.Info)	(1)	Shannon Image entropy (O.IE)	The amount of information or randomness in the image, calculated by applying the entropy formula to the grayscale pixel intensity values of the image.	$-\sum_{i=1}^{N} p(i)\log_2 p(i)$, where $p(i)$, the probability of occurrence of the i^{th} grayscale pixel intensity in the image, and N, the total number of different pixel intensities.	[40]
	(2)	Kolmogorov Complexity (O.KC)	Image size after removing redundant information.	$S_{\text{origin}} - S_{\text{redundancy}}$, where S_{origin}, the original size of the image, and $S_{\text{redundancy}}$, the size of redundant compressible information.	[25]
	(3)	Subband entropy (O.SE)	The entropy of the image content within subbands, calculated using the probability distribution of the the wavelet coefficients within each subband.	$-\sum_{i=1}^{N} p(i)\log_2 p(i)$, where $p(i)$, the probability distribution of intensity values or coefficients within the i^{th} subband, and the sum runs over all N subbands.	[119]
	(4)	Information gain (O.IG)	The amount of additional information needed based on the mean information gain to determine the color of an adjacent cell given the color of a particular cell.	$-\sum_{i,j} p_{ij}\log_2(p_{i\to j})$, where p_{ij}, the joint probability that a given cell has the i^{th} color and the adjacent cell has the j^{th} color, and $p_{i\to j} = p_{ij}/p_i$, the conditional probability that the adjacent cell has the j^{th} color given that a cell has the i^{th} color.	[3]
Clutter (O.CL)	(5)	Feature congestion (O.FC)	The clutter estimate in the image computed by assessing feature viability based on the entropy of color, luminance contrast, and orientation.	$-\sum_{i=1}^{N} p_i \log_2 p_i (p_i = \frac{E_i}{\sum_{i=1}^{N} E_i})$, where p_i, the proportion of energy E_i of the i^{th} feature relative to the total energy of all N features in the image.	[119]
	(6)	Heterogeneity (O.H)	A gray-Level concurrent matrix to provide insights into the texture and contrast of an image. Higher heterogeneity values indicate more contrast and variation in gray levels, suggesting a more complex or textured image.	$\sum^{L-1} i = 1 \sum^{L-1} j = 1 \frac{P(i,j)}{1+(i-j)^2}$, $P(i,j)$ represents the probability that a pixel with intensity i occurs adjacent to a pixel with intensity j, and L is the total number of gray levels in the image.	[29]
Color (O.CD)	(7)	Colorfulness (O.CF)	Overall saturation and the variety and extent of colors.	$\sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + \kappa \cdot \sqrt{\mu_{rg}^2 + \mu_{yb}^2}$, where σ_{rg} and σ_{yb}, the standard deviations of the chromaticity of red-green and yellow-blue, μ_{rg} and μ_{yb}, the mean of the chromaticity of red-green and yellow-blue, and κ, the weight coefficient.	[46]
	(8)	Color RGB entropy (O.ERGB)	The variability in the distribution of colors calculated by applying Shannon's entropy to the color histogram of an image.	$-\sum_{i=1}^{N} p_{\text{color}}(i)\log_2(p_{\text{color}})(i)$, where $p_{\text{color}}(i)$, the probability of occurrence of the i^{th} color in the histogram, and N, the total number of different colors histogram.	[40]
Shape (O.ShD)	(9)	Edge density (O.ED)	The proportion of pixels identified as edges within an image, quantified by the Canny edge detector.	$\frac{P_{\text{edge}}}{P_{\text{total}}}$, where P_{edge}, the total number of edge pixels, and P_{total}, the total number of pixels in the image.	[15]
	(10)	Feature point (O.FP)	Points where edges intersect, identified by measuring changes in image brightness.	$\frac{P_{\text{junction}}}{P_{\text{total}}}$, where $P_junction$, the total number of pixels of junction points, and P_{total}, the total number of pixels in the image.	[44]
Object (O.OD)	(11)	Text-ink ratio (O.TiR)	The proportion of text pixels in an image.	$\frac{P_{\text{text}}}{P_{\text{total}}}$, where P_{text}, the total number of text pixels in the image, and P_{total}, the total number of pixels in the image.	
	(12)	Number-of-color (O.MeC)	The number of meaningful colors count.	The curated human namable colors of Heer and Stone [50] in an image.	

mation capacity from data in an image. Both Y-F and C-J's methods require knowing the information flow in the image space. To automate the process to compute *data-bits*, Jänicke et al. [67, 80] use color entropy to highlight and sort flow features. Here, they treated sensory perception independent of data saliency to analyze flow features by information theoretic solutions. Furthermore, Jänicke and Scheuermann [66] computed statistical complexity in data [67] to compare, analyze, and highlight features that assist human observers [91].

What inspired us in these works is the transition from the low-level design element to key probabilistic data-specific knowledge in images, modeled under the same information theoretic framework of mathematic uncertainty. However, we were unable to directly extend their work, as there is currently no established method for defining bits in the context of general visualizations beyond flow fields. A broader idea drawn from this line of research is that more complex images tend to exhibit lower redundancy and thus higher entropy. For this reason, the compression rate of an image—aligned with Kolmogorov complexity [115]—serves as reasonable approximation of pixel-level uncertainty (Item (1) O.KC, Table 2). Building on this idea, We introduced several additional metrics (Items (2)-(4), Table 2) that adopt this entropy-based approach. Furthermore, we incorporated pixel-level RGB entropy into our metric set (Item (8) O.ERGB, Table 2).

2.2.2 Clutter-based Feature Congestion

The design of FC originated from an interesting question: whether the robust measure of the "number of unique objects" [16, 86] can be computed mathematically [118] to eliminate the need for labor-intensive manual coding. Here, congestion is characterized by the lack of available space for new features [119], and "*the more cluttered a display or scene is, the more difficult it would be to add a new item that would reliably draw attention*" [118]. Subsequently, this metric assembles the spatial frequency of Gabor texture patches and colors to simulate receptive fields, extending the subband entropy onto the *object-level* structure of color, orientation, and saliency [118]. This is grounded on the understanding that an image is considered 'simple' [99] when it is symmetric and structured. Thus, complexity is inversely correlated with perceptual grouping and regularity [53]. Empirical evaluation results also supported the FC metric (Item (5) O.FC, Table 2), subband entropy (Item (3) O.SE, Table 2) and edge density (Item (9) O.ED, Table 2) on measuring VC [117, 118]. Borgo et al. [9] further emphasized the idea of statistical attributes to characterize complexity. This line of work informs us that information amount influences VC. Both Rosenholtz in vision science and Jänicke and Chen in visualization use color as the primary channel for information-theoretic metrics [67, 117]. Given that color and textures are uniquely represented in visualization images, we modeled between continuity and complexity conditioned for continuous and discrete color and texture stimuli, or RR-ColorTexture of Continuous Heatmap and Discrete Heatmaps in images.

2.2.3 Color-based Measures

Beyond its role as a secondary attribute attached to other primary stimuli [10], color is perhaps one of the most extensively studied visual variables for direct data mapping, as evidenced by numerous reviews [127, 140, 141, 159] and real-world applications [8, 21]. The importance of color is further underscored by the fact that the transfer function in visualization is fundamentally a coloring process [106].

Color also plays a significant role in shaping perceptual experiences. For example, Li and Chen [79] suggest that using too many colors can reduce memorability due to increased clutter. Moreover, color is a leading factor in how visual groupings are perceived, supporting both segmentation and the identification of structure within complex visual scenes. In vision science, both Greene et al. [41] and Rosenholtz [117] highlight the role of color in defining object boundaries. Many studies have linked color-related features to VC. For example, Rosenholtz et al. observed that, in natural object scenes, color variability is independent of both edge density and subband entropy [118]. Both Oliva et al. [99] and Rosenholtz et al. [119] reported that perceived VC increases more with color variability (the number of colors and how different they are) than with the presence of high frequency details in color channels. Corchs [27, 55] introduced metrics such as color harmony and color coherence to show that both the interaction and uniformity of colors contribute to VC. Here, we balance our metric choices to capture both statistical color properties and object-level color counts. Specifically, we used *colorfulness* (Item (7) O.CF, Table 2), defined as "*the variety and diversity of colors present in the region*" [46], and gray-level color heterogeneity (Item (6) O.H, Table 2) to indicate overall saturation [46], in addition to the aforementioned O.ERGB.

2.2.4 Object-based Measures

Several studies have focused on object-level measures, emphasizing that VC involves higher-level cognitive processing, which requires the analysis and integration of multiple elements along with their semantic content. Semantic contents in visualization can have *two* types: *graphical elements* (e. g., legend, tick marks); and *annotations* [109]) conveying key insights and takeaways [1]. The effect of texts on VC can be mixed. For one, they can effectively guide viewing behavior [10, 13] to reduce VC. For the other and from a pixel-based perspective, text is treated as a visual structure—often dark-colored, densely arranged, and high in contrast, that can increase complexity. (Item (11) O.TiR, Table 2.)

Color plays *two* main roles in object perception: it can function as a *low-level* graphical element to encode data directly (e. g., field visualization). It also supports *high-order* ensemble perception by contributing to perceptual grouping, thereby facilitating visual search. For the later, Greene et al. [41], Oliva et al. [99], and Chai et al. [16] suggest that color helps define object boundaries, and that the number of distinct object types in a scene influences its perceived complexity. Similarly, Snodgrass et al. [128] and Donderi et al. [32] consider both the number and variety of objects in their assessments of visual complexity. Finally, it is worth noting that we are not the first to introduce color count (MeC) as a way to quantify high-order perception. Borkin et al. and Li and Chen used it to quantify memorability [10, 79]. However, a key distinction is that their MeC values were based on participant-reported perceptions of color, while our metric is derived from objective image analysis based on actual color usage. Collectively, these studies demonstrate how the presence, quantity, diversity, and relationships of visual objects contribute to the perception of VC.

Since images are not typically treated as object-based scenes, where elements are explicitly counted or statistical attributes are quantified, we explore the use of color to differentiate visual elements (Item (12) O.MeC, Table 2). Results from our workshop further suggest that color is a leading factor influencing image clarity. Together, these insights allow us to investigate how high-level visualization elements contribute to perceived VC.

3 A Crowd-Sourcing Study

Having established an understanding of the factors influencing VC through an extensive literature review, we proceeded to directly collect crowdsourced VC scores. We began with a workshop and pilot studies to inform the proper procedure for our data collection.

3.1 Image Data Collection and Characteristics

We prioritize diverse stimuli and therefore used both VIS30K [18] and MASSVIS [10] as primary data sources. They both come from real-world applications thus let us examine practical VC. Most importantly, they function differently: VIS30K contains scholarly images texting formal IEEE publications, often presents scientific discoveries or new design, where annotation is rare and the corresponding description often appears in the captions or main text. In contrast, MASSVIS images are used mainly for communication, therefore key insights and trends are sometimes annotated, which further serve our goal of understanding non-graphical elements, e. g., text in images. Also, VIS30K contains glyphs and vector field images not found in MASSVIS.

3.2 A Workshop and A Pilot Study

The goal of the workshop was mainly to choose a reliable method to collect a large number of VC scores. Extensive detail of the activities and outcomes are provided in the Apdx. subsection III.2. Here we give a high-level overview and decisions influenced by these activities.

We first adapted a recent and well-accepted method used in BeauVis for a direct score assignment [47] and found it difficult to align scores above several 10s of images (He et al. [47] used 15 images). Next we conducted an in-person exploratory workshop with 49 participants, akin to the hierarchical division method of Oliva et al. [99]. Participants ranked images and captured factors influencing VC. We found that it was well-suited for collecting comments and factors, however, it also did not scale (Oliva et al. [99] used 100 images). We also found that not a single participant used a single criterion in their VC assignment.

The third method we tried was an active sampling solutions that could provide us absolute VC scores via pair-wise 2AFC ranking, successfully used in the computer vision community to collect a large set of image VC scores (e. g., [97, 123]). The idea was similar to how game competitions assigned competing players to achieve a global score for each player. Our piloting experiments with 90 and 200 images supported that the single-stage Bradley-Terry ranking [12] and Microsoft's TrueSkill™ [52] were both reliable. However, scaling up this approach to 1,000 in the global space using only one-time sampling would require many more trials, not very feasible. We finally piloted Mikhailiuk et al. [90]'s multi-stage active sampling algorithm (Apdx. Figure 11), which could achieve the same level of reliability as the single-stage active sampling with 10% of total trials. In each phase, the algorithm optimized the information obtained from all pairwise comparisons to compute reliable VC scores, while limiting the total number of comparisons. Our pilot testing using visualization images obtained the same reliability as the authors described and saved running cost by 9 fold.

In the pilot study, we also discussed what *not* to collect. For example, we excluded other high-level subjective metrics, such as aesthetics, understandability, and readability, even though they might correlate with VC [112, 116, 138] because these factors are also inherently subjective, potentially relying on the same set of objective metrics. Instead, we focused on objective image metrics, aiming to develop an understanding that remains independent of individuals' backgrounds, experience and expertise, for quantifying the crowdsourcing perceived VC scores. Images were static free of measuring viewers' ability to navigate.

3.3 Participants and Procedure

At this stage, we had clear ideas on the image data and the method to collect the perceived VC scores. Now we report our crowd-sourcing experiment via the online experiment platform Prolific [103] to capture the absolute perceived VC scores for the 1,800 images.

We instructed participants on the experiment procedure and provided them the high-level definition of VC always shown on the screen "*the amount of detail or intricacy in images*" (Apdx. Figure 10) as used in Purchase et al. [107, 128]. They were instructed not to infer knowledge and not to use information other than those presented in the images, following the instruction of Oliva et al [63]. We also had two attention test trials to flag inattentive participants out of the study. In each stage, about 20-25 volunteers compared 79 pairs of images for us to determine the VC for that stage, viewed images recommended by the active sampling algorithm. Participants viewing image pairs recommended by the algorithm. At each stage, an absolute complexity score for each image was updated from a probabilistic process to generate a global complexity score described in more detail next.

Participants' browser size and their operating system were automatically detected and their screen view must be at least 1028×764 on a desktop operating system to be eligible. Before this data collection,

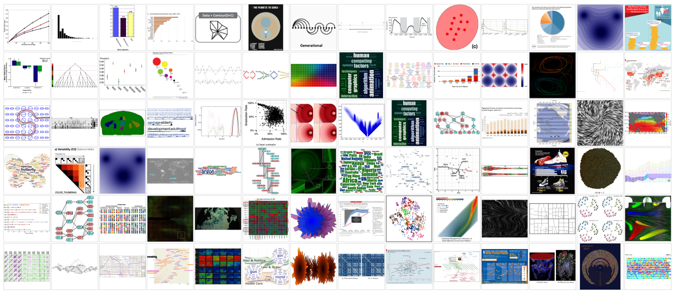


Fig. 3: **Resulting Visual Complexity Scored Image Examples** (least to most from top left to bottom right).

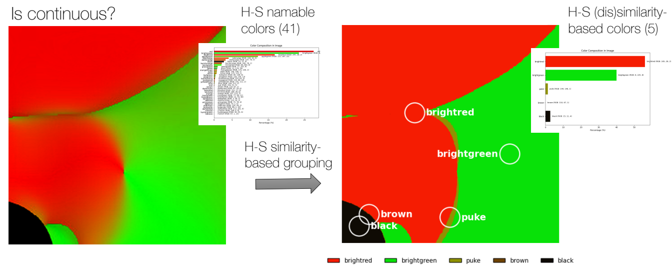


Fig. 4: **An O.MeC calculation pipeline for an image with a continuous color-map.** Left: the original image with 41 measured H-S [50] namable colors; Right: 5 colors grouped by H-S similarity to produce O.MeC=5.

we first collected participants' demographic information and ensured accountability by recruiting only those with a success rate above 95% on Prolific. Additionally in the post-questionnaire, we gathered participants' written reasoning from two randomly selected trials.

In total, 349 participants (aged 18 to 25 (111), 26 to 35 (143), 36 to 45 (51), 46 to 55 (27) and over 55 (17)), with education levels: graduate school (134), college (130), professional school (55) and some high school (30). The gender distribution is nearly balanced: 168 women, 176 men, and 5 who did not disclose their gender. The participants were compensated €9.5 / hour for their participation.

3.4 Crowdsourcing Results

The experiment was conducted over 14 stages for approximately 34 hours, resulting in a total of 27,571 trials, with an average of 1,969 trials per stage. Each trial lasted an average of 4.34 seconds (95% CI: [4.17, 4.50]). Each visualization image was viewed approximately 234 times (95% CI: [202, 266]).

VC Score Calculation. We applied the Mikhailiuk et al.'s approach [90] to calculate the VC scores for all 1,800 images. The approach first constructed a comparison matrix containing the outcomes of all pairwise comparisons, where each entry recorded how often an image was selected as more visually complex when compared against others, and then applied the TrueSkill™ algorithm [52], to infer relative complexity from the comparisons. TrueSkill™ is a Bayesian ranking algorithm that models each image's VC as a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with a mean (μ) representing the estimated VC score and a standard deviation (σ) representing the uncertainty of that estimate in our context. During each comparison between two images i and j, TrueSkill™ estimates the probability that i is more visually complex than j. Specifically, if image i is judged more complex than j, the mean μ_i is increased and μ_j is decreased; both variances σ_i^2 and σ_j^2 are reduced as the model gains confidence through increased comparisons. After many pairwise comparisons, the distributions converge.

For our analyses, the final VC score of each image is taken as the posterior mean which reflects the inferred likelihood that it would be judged more complex than a randomly chosen image from the dataset (Additional validation in Apdx. Figure 11.) Figure 3 shows a subset of 84 images, sorted in ascending order of their VC scores from left to right, then top to bottom. (Apdx. Figure 12 provides the full set of 1,800 VC images and scores.)

4 Metrics-Based VC Attribution Analysis

This section presents our two new objective metrics and observations after applying these 12 metrics to our image dataset. Figure 5 shows some example outputs after applying the metrics to visualization images.

4.1 Our Two Object-based Image Quality Metrics

We define two metrics related to the semantic information of a scene, building on the concept of graphical elements, where objects are recognized quickly, sometimes in a single glance [101]. While our community has traditionally relied on edges to define boundaries, we focus here on high-level perception of using colors. Here, we drew on insights from vision science, including the role of color in delineating element boundaries and the preattentive nature of coloring in visualizations [21]. Also, we saw that many images contained categorical colors in the legend. These perspectives informed our decision to use color as a means of defining boundaries between data elements, alongside text, which can be recognized and quantified with relative ease.

Text-ink ratio (O.TiR) represents the proportion of text-area pixels in an image. Here, "text" refers to contextual information, including annotations, axis labels, or titles. Any text used to explain, clarify, or define image content is considered "annotated" [109]. We excluded text-based visualizations where text was used to show data. There are many methods to compute O.TiR. In our case, we used an OCR (Optical Character Recognition) package PP-OCR [34] to compute text regions and then use the VIS30K bounding box labeling tool [18] (Apdx. Figure 16) to curate annotations, legends, etc.

Color count (O.MeC) defines the number of distinct colors/objects in a visualization image. Several considerations guided how we computed the MeC, due to the difficulty to name the groups as Rosenholtze et al. pointed out [119]. For example, would a data sample in a scatterplot be considered an item or a cluster of these an item? Here, we assume that designers would have distinguished 'items' by color—if the goal is to see the cluster, they would be shown as color-based clusters. As a result, for discrete data, we simply count the number of items by looking at the legend without concerning items within that cluster, since legend captures human attention first [13] and color supports fast pre-attentive processing [146, 157]. For these images using discrete colors, we count the number of unique colors present in the legend, along with any additional text color.

Counting colors in continuous colormaps is also not straightforward. Our implementation prioritizes semantically meaningful colors by Heer and Stone (H-S) [50], as these are more easily and rapidly perceived by humans. Using the H-S color dictionary, we map each unique pixel value to the closest color name in the H-S vocabulary. For example, this step shows 41 namable colors for the image in Figure 4. We next merge the similar colors using the H-S's similarity dictionary, where we use the nearest-centroid color of the colors in the same similarity group. For the same figure, this step trimmed the 41 to 5 colors, shown in the white circles. The name of these colors are shown next to their locations. For this figure, we used "5" as the final O.MeC.

One may argue that luminance areas were not consistently counted in our analysis. In practice, we addressed this on a case-by-case basis. For texture-based pattern images such as LIC, we included luminance in the color count. Since the H-S naming system is not uniformly distributed in the L*A*B space, some color names are perceptually closer than others. In cases like this, we applied a color distance threshold: colors with a Δ E $\leq$ 14 were merged (Apdx. Figure 26). This value is slightly higher than the typical human perceptual threshold ($\Delta E > 10$ [151]), allowing us to account for pixel-level noise and small color variations. This threshold also helped reduce salt-and-pepper noise in the image. Two co-authors curated the final answers.

4.2 Observing 12 Objective Image Quality Metrics

Information-theoretic-based metrics. We used four metrics: (1) the original *Shannon Image Entropy (O.IE)* [126]; (2) *Kolmogorov Complexity (O.KC)*, which estimates an upper bound of an image's information capacity using the zstd compression method [25], where higher complexity indicates greater internal redundancy, resulting in larger compressed sizes; (3) *Subband entropy (O.SE)* [118], measuring the

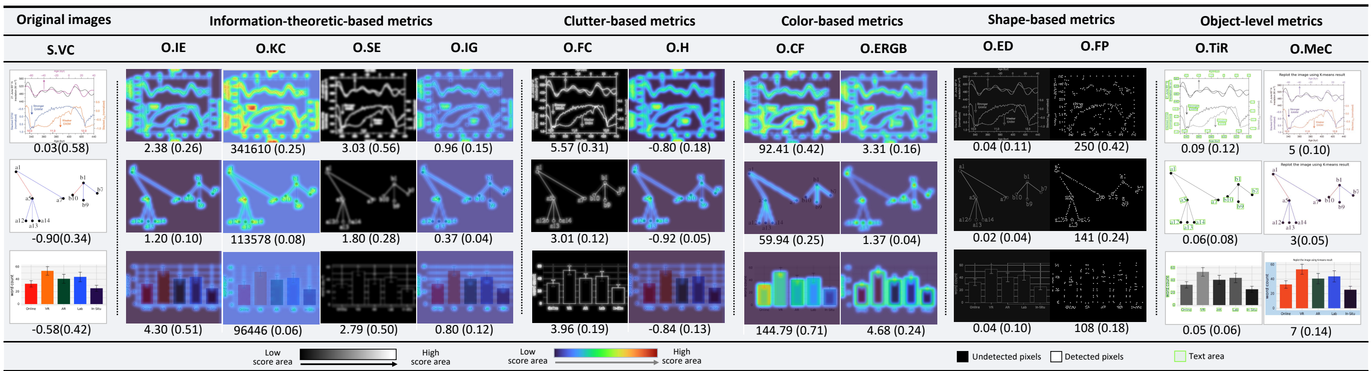


Fig. 5: **Objective metric scores in response to the visualization image input.** Each row shows the original image followed by visual representations of the 12 objective metrics, with corresponding metric scores shown as raw (normalized) values below. The heatmaps highlight image regions where each metric responds most strongly—brighter and redder areas indicate higher values or more visually complex regions.

randomness of information within different frequency components of an image after wavelet decomposition; and (4) *information gain (O.IG)* [3], which examines neighboring pixel probabilities and sensitive to local textures, edges, and details within the image. We observed that these four information-theoretic metrics were highly correlated emphasizing similar regions of an image (Figure 5 and Apdx. Figure 23).

Clutter-based Metrics include (5) feature congestion (*O.FC*) [118] and (6) heterogeneity (*O.H*) [29], considering arrangement and overall pattern distribution. O.FC quantifies how "cluttered" or "busy" an image appears by analyzing structural variations of color and luminance contrast. Similarly, higher *O.H* indicates greater contrast and variation in gray levels, which suggest a more complex or textured image. Figure 5 shows that areas with a high density pattern and structure variations were emphasized by these two metrics.

Color-based metrics. For *O.ERGB*, higher entropy indicates greater color variation. We can see that smaller regions with many color changes tend to have higher ERGB, capturing the edges between the colored bars. In contrast, grey-scale images exhibit lower scores due to reduced color variability. *O.ERGB* also successfully highlighted nodes in node-link diagrams. On the other hand, *O.CF* tends to miss dense node areas but causes empty regions to stand out more prominently.

Shape-based metrics. We used (9) edge density (*O.ED*) studied in [86, 118, 119], representing the ratio of edge pixels with discontinuities in image intensity to total pixels in an image and (10) feature points (*O.FP*) [44] represent the corner points with significant variation from surrounding pixels. They both emphasize areas with highly dense lines and significant variations in color values.

Multi-colinearity analysis. We also performed the co-linearity analysis (see Apdx. Figure 23 for detailed analysis). The main take away message was that we observed within category correlation especially in the pixel-based information theoretic and between shape-based metrics. In contrast, the object-based metrics, *O.TiR* and *O.MeC* showed little to no correlation with other metrics.

4.3 Metric-informed Factors that influence Perceived VC

We now examine how the 12 objective image quality metrics relate to perceived VC, as collected through our crowdsourcing experiment described in section 3. We aimed to obtain the most informative subset of these 12 input metrics to model VC and make it interpretable.

4.3.1 Analysis Method

We employed Partial Least Squares Regression (PLS) to model perceived VC using the 12 objective image quality metrics as predictors. PLS was chosen for its ability to handle multicollinearity, i. e., the high correlations among predictors, shown in our data in Apdx. Figure 23. PLS leverages the correlations to extract latent components that summarize the variance in both predictors and the response to identify a few variables with strong influence on the outcome (perceived VC) [89].

Various methods can measure multiple variables for model explainability in data analysis. We did not use linear [107] or least absolute shrinkage and selection operator regression (LASSO) [133], as they either require manual intervention to address multicollinearity or shrink less important coefficients to zero, which may introduce subjectivity. We also did not use factor analysis methods such as item response theory (IRT) [6], because it focuses on uncovering latent constructs within the predictor (x_i) space. PLS constructs latent components optimized to *explain* the outcome variable (perceived VC). While factor analysis or IRT are useful for understanding underlying factor structures or modeling latent traits, they do not provide direct predictive modeling frameworks aligned with regression objectives. PLS bridges this gap by estimating both the underlying structure and a modest number of explanatory variables in relation to the response.

We computed PLS regression coefficients to estimate the relative contributions of each metric to perceived VC. We also used bootstrap resampling to estimate the variability of each coefficient and compute p-values, thereby identifying which metrics are *more significant* and *stable* across resamples (details in Apdx. subsection III.3). We used the resulting PLS regression coefficients to estimate the relative contribution of each metric to perceived VC. Interpreting these coefficients, however, requires care. Unlike in standard regression, the coefficients in PLS reflect the contribution of a variable conditional on its shared variance with other predictors, which means the following: (1) variables with larger absolute coefficient values are interpreted as having greater influence, while the sign of the coefficient indicated whether the effect was positive or negative; (2) small coefficients may not indicate that a metric is unimportant, but rather that its influence overlaps with other variables or is too unstable to predict VC.

4.3.2 Results: Overall Factorizing Perceived VC

Figure 6 presents the resulting PLS coefficients, with the metrics grouped from left to right in shaded zones by the five factor categories defined in Table 2. Solid-colored bars represent significant metrics ($p < 0.05$), while those outlined in gray wireframes are not significant. The number of significant metrics indicates that all five categories originally derived from natural scene perception are applicable to visualization contexts. While only one metric is significant in each of the information-theoretic, clutter, and color categories, all metrics in the shape-based and object-based categories are significant. The R^2 value reflects the proportion of variance in perceived VC explained by the model—the higher the value, the better the model fit. In this case, R^2=0.41 indicates an overall modest fit across all images.

(1) The influence of generic pixel-based information-theoretic metrics and color metrics on perceived VC is minimal. The pixel-based information-theoretic metrics, O.SE ($coe = -0.08, p = 0.02$), is the only significant coefficient at the 5% level. O.IE ($coe = -0.02, p = 0.87$), O.KC ($coe = 0.05, p = 0.21$), and O.IG ($coe = -0.03, p = 0.45$) are much less significant in the prediction of VC.

(2) Information-theoretic metrics based on specific stimulus and attributes are significant factors. Specifically, O.FC in the clutter category, adapted to statistical distributions of color and texture, and O.ERGB in the color category, the original Shannon entropy on color alone, were both significant main effects with $coe = 0.08$ ($p = 0.01$)

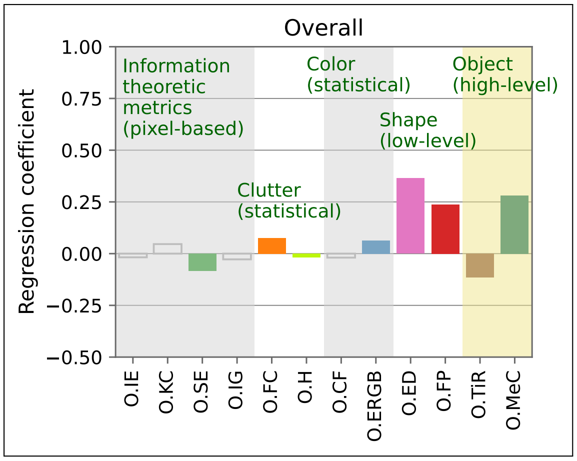


Fig. 6: **Factoring visual complexity.** The relative contribution of each image metric is represented by the magnitude of its corresponding regression coefficients, modeled using PLS, with the significant ones highlighted in solid color (metrics significant at the 0.05 level).

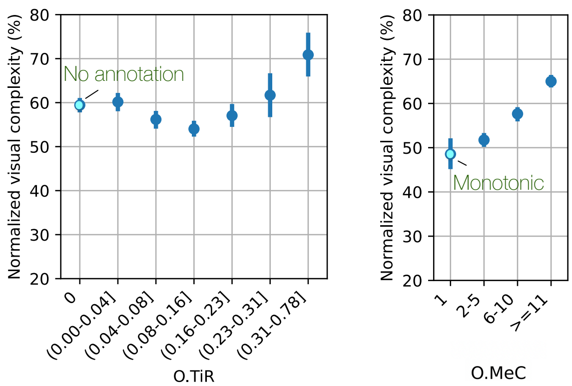


Fig. 7: Mean VC scores for visualizations based on the number of colors O.MeC and the Text-ink-Ratio O.TiR it contains. Error bars represent 95% confidence intervals.

and $coe = 0.06$ ($p < 0.001$), respectively. The other metric in the clutter category, O.H, that computed pixel-level heterogeneity of an image, was also significant due to a small variation in data, although its coefficient was small ($coe = -0.02, p = 0.001$). The metric that measured the variety and diversity of colors (O.CF), however, did not show a significant contribution to perceived VC prediction.

(3) Shape and object metrics contribute most to complexity. Metrics capturing shape and object quantity were the strongest predictors of perceived VC. O.ED and O.FP, categorized as shape-based metrics (Table 2), reflect low-level features following Oliva et al. [99], while O.MeC represents a high-level feature that estimates the number of objects based on color count [41]. Overall, metrics that quantify the total number of visual elements, whether shape- or object-based, proved to be significant. Specifically, O.ED ($coe = 0.36, p < 0.001$), O.FP ($coe = 0.24, p < 0.001$), and O.MeC ($coe = 0.28, p < 0.001$) were the top three positive contributors to perceived VC.

(4) Text-ink-ratio reduces perceived visual complexity. O.TiR, which measured the proportion of annotated text within an image, was found to have a negative coefficient (-0.12, p=0.005), suggesting that the presence of annotation may in general reduce rather than increase perceived VC. As discussed earlier, annotated text can enhance clarity and aid in interpretation, thereby lowering perceived complexity. To explore this further, we analyzed the trend between O.TiR and perceived VC and observed a non-linear, bell-shaped relationship (Figure 7). Specifically, VC initially decreases as O.TiR increases, reaching a minimum around the $0.08-0.16$ range, and rises again beyond a threshold of approximately $0.23-0.31$. This difference in complexity is statistically significant ($F_{6,1770} = 9.8, p < 0.001$), indicating a potential optimal range of O.TiR, where a moderate amount of text can reduce perceived complexity by improving clarity, while too little or too much text may have the opposite effect.

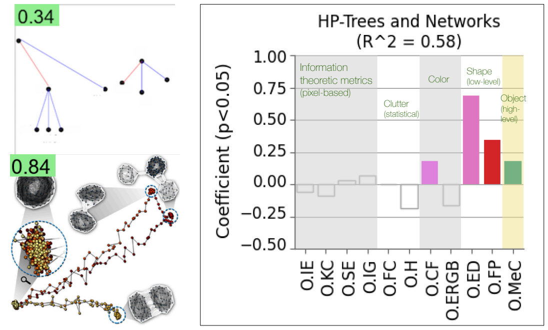


Fig. 8: **Factoring 189 visual complexity of trees and network only** (in Purchase [107] style of direct node-link diagrams). The contribution of each image metric is represented by the magnitude of its corresponding regression coefficients, modeled using PLS. Only metrics significant at the 0.05 level are shown in solid color.

We also observed that increasing O.MeC was associated with higher perceived VC (Figure 7). Among the dataset, we found 96 black-and-white or monochrome images with no color variation ($O.MeC = 1$). While natural scenes are almost always colorful, visualization images can achieve similar perceptual differentiation through other cues such as texture and shape variation. In general, grayscale or low-color images were perceived as less complex than colored ones, a difference that was statistically significant ($F_{3,1773} = 125.0, p < 0.001$). For the images with higher O.MeC, we found that color often served as a boundary-defining cue, helping to visually separate elements. This boundary-forming function of color contributed to increased perceived complexity, especially when the colors clearly delineated objects or categories. Notably, this effect was consistent regardless of whether the image included textual annotations ($p < 0.01$ for both cases). Among these color-rich images, some used discrete colormaps with clearly defined edges, while others employed continuous colormaps such as gradient-style heatmaps. In some cases, these color patterns also reinforced coherent textural structures, further adding to visual complexity. Due to these observations, we next separate color/texture images by continuity to further investigate their relationship with perceived VC.

Summary. Our analysis shows that all variable groups contribute to higher perceived visual complexity: information-theoretic metrics, greater edge density, more feature points, and a larger number of distinct objects. This result replicates the findings in vision science that visual complexity is likely to be a high-dimensional phenomenon influenced by many factors. As shown in the three panels of Figure 1, the simplest and most complex visualizations differ notably in visual content. Similar patterns are observed in Apdx. Figure 13 when comparing samples from MASSVIS and VIS30K. Some visualizations, particularly Matrix views, appear more complex due to their use of many small points and a wider range of colors. In contrast, simpler images tend to contain fewer elements and use minimal colors.

5 Comparison to Previous Works

Having established the factors that influence perceived VC through quantitative, metric-based analysis, we applied the same approach to replicate whether these metrics can also account for factors previously studied in the context of node-link diagrams [107] and clutter [118].

5.1 Network-based Visual Complexity: Compared to Purchase essential stimuli [107]

There are 189 direct node-link diagrams in our VisComplexity2K dataset (Figure 8). We thus studied and compared our factorization analysis to that of Purchase et al. [107] to see if we could replicate results in that study. Purchase et al. [107] studied network aesthetics by considering four variables: color (or individual RGB components, similar to O.ERGB), object edges (which define boundaries, similar

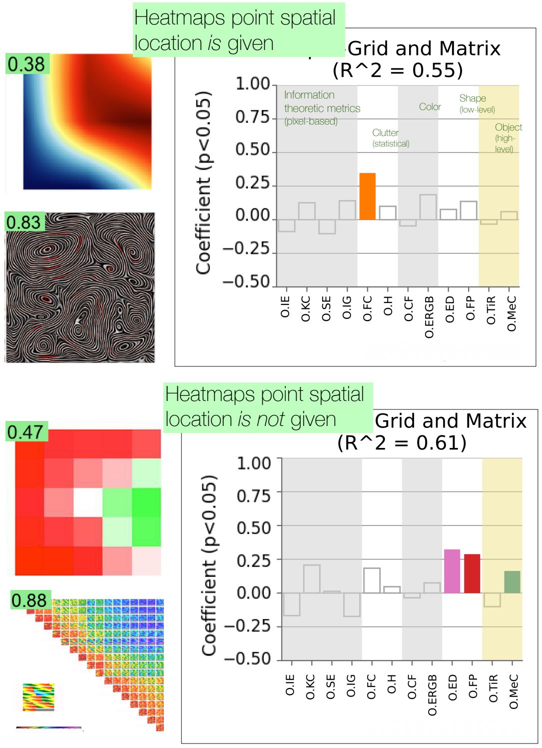


Fig. 9: **Factoring visual complexity**: Significant metrics and their contributions to VC quantification vary by whether the visualization is continuous. *Top*, pixel positions are given and surface are continuous smooth color and texture. O.FC is the only metric significant at the 0.01 level ($p = 0.003$). *Bottom*, the pixel position is discrete, artificially designed and clear boundaries are drawn. Three metrics are significant at the 0.01 level: O.ED ($p = 0.003$), O.FP ($p < 0.001$), and O.MeC ($p = 0.006$).

to O.ED), intensity variations (similar to O.CF), and file size (similar to O.KC). Their model explained only 25% of the variance in network visualizations. In contrast, our study employed the same edge detection method but found that O.ED was a statistically significant main effect, suggesting a stronger role for edge-based features in perceived VC.

To replicate and understand this extended large dataset, and compare results with the Purchase et al. study on network aesthetics, we curated a subset of our dataset containing network-style visualizations (primarily node-link diagrams, as shown in Apdx. Figure 25). To better match the original study design, we manually (1) filtered for direct node-link diagram (such that the matrix views were removed); (2) Removed all text and annotations using automated masking followed by a manual cleanup to isolate pure graphical elements given that these texts may activate the edge detector. We refer to this cleaned subset as HP-Trees and Networks. Using our 11 remaining metrics (excluding O.TiR), our PLS regression achieved an R^2 of 0.58 indicating a strong fit—substantially higher than the 0.25 reported in Purchase et al. [107]. Notably, O.ED (edge density, $p < 0.001$), O.FP (feature points, $p < 0.001$), O.MeC ($p = 0.003$), and O.CF ($p = 0.016$) were key contributors to perceived complexity in this subset (Figure 8). We would credit the results to the addition of O.MeC and O.FP, which emerge as notable contributors to perceived VC in our dataset.

5.2 Information Theoretic Measure of Color and Texture on Clutter Effect: Compared to Rosenholtz et al. [119]

Rosenholtz et al. designed this clutter-based metric to capture color and texture distributions in a natural scene. Our dataset contained 311 ‘Grid and Matrix’ or ‘heatmaps’ [10], of which 185 are spatial, continuous, where spatial location ‘is given’ [96], making them more closely aligned with Rosenholtz’s definition of structured scenes. The rest are discrete, with pixel locations artificially generated—where structured grids and matrix introduce corners and edges. This contrast is evident in our dataset in these two sets of heatmaps (Figure 9). We also removed all text before modeling these two sets of visualization images to avoid edges and corners confounded by the presence of text.

Figure 9 shows the results. When spatial location *is* given, we found that for visualizations characterized by color and texture, O.FC emerge as the single most informative metric for explaining perceived VC. In contrast, when location is *not* explicitly encoded, shape-based metrics (O.ED and O.FP) and the object-based metric O.MeC are the primary contributors to perceived VC. This result shows that in these discrete grids, edges also act as visual separators, effectively representing the number of visual elements present.

In summary, the continuity-specific results largely mirror the overall findings in terms of dominant contributing metrics, with some variation based on the stimulus type the metric models. What makes this particularly intriguing is that O.FC is highly effective in explaining complexity for continuous heatmap images but had limited explanatory power for artificially generated heatmaps. Given that the coefficients are higher by individual visualization types and continuity, we recommend that type categorization for VC must take into account continuity.

6 General Discussion

This section discusses theoretical and design implications of our results. A central problem for the science of visualization is related to perceptual experience as it influences people’ understanding about data. One of such high-level perceptual experiences is visual complexity. Our present work contributes to understanding the alignment between human reading and metric measurement.

6.1 Comparisons to Natural Scene Understanding in terms of Item-Count and Feature Congestion

Multiscale structures emerge in our study. Number of elments operates on object-associated O.MeC and O.TiR, and low-level features such as edges O.ED and corners O.FP, contributing significantly to VC across overall dataset. This result indicates that human perception does not operate on a single scale [150]. This high-level perception seems to align to Turing’s distinct hypothesis: patterned structures can self-organize through local competitive interactions [135]. As a result, perception can be fast and minimizes the need for generic instruction to perform pattern findings. Turing described that people’s perceptual computing in this competing mechanism appeared to be more robust to noise, where global structural emergence exhibits topological robustness of insensitivity to variations and noises.

In visualization contexts, global structures formed by color and pattern may support visual Gestalt principles, as captured by the two object-based metrics: O.TiR (related to text) and O.MeC (related to color and object boundaries). These size-scale representations showed that low-level elements such as points and edges are important, but the features also emerge at the level of higher-order textures and objects. This is particularly evident in heatmaps where spatial location is given and aligns with the descriptor attribute O.FC, as demonstrated in our experiments (Figure 9). In this context, O.FC represents a meaningful step forward in quantifying human perceptual experiences by treating visualizations as statistical, scene-based representations.

This result may also indirectly support two perspectives. First, while Borkin et al. [10] treated Grid and Matrix as a single visualization type to characterize perceptual experiment (memorability), our findings may support that continuity can be considered as a distinguishing factor when we define type. Second, the prolific investigations of information-theoretic measures in data visualizations [22,65,67,154] may contribute to humans’ perceptual experiences—particularly when the underlying information bits are well defined. In our case, the color/texture dimensions in O.FC can be interpreted as encoding structured visual patterns. This result further confirms that continuous color/texture patterns in visualizations is understood differently from discrete grid and matrix.

6.2 Graphic or Text? The Influence of Semantic Information on Perceived Visual Complexity

Our findings suggest that the semantic information conveyed by text in visualizations impacts perceived VC differently than graphic elements.

We conjectured that explanatory text may reduce perceived VC by enhancing understandability, e. g., Hearst et al. also highlighted the significance of text, considering it as equally important as the visualizations themselves [49]. We found a positive effect of text on VC to a certain extent. In general, explanatory text reduces perceived VC due to its contribution to understandability. However, when the amount of text becomes excessive, it may increase complexity, shifting from aiding comprehension to overwhelming the visual features.

Analyzing variations in text layout and related attributes may reveal new ways to understand perceived VCs. Previous eye-tracking behavior data for capturing memorability suggested that people can quickly distinguish text and titles. These text elements always occupied separate areas on the screen, making them easy for the human visual system to decode. In previous studies on three-dimensional scenes, where texts overlapped with spatial objects, the list-based approach was often preferred [20], highlighting the role of spatial organization in effective information presentation. Other non-data elements beyond annotations may also augment human perceptual experience, although their interaction with graphical elements can be complex. For example, adding a background grid [131], incorporating chart-junks [57], or using subtle contextual cues [147] have all been shown to augment human visualization perceptual experience [37]. The question of how much text is sufficient, and how non-data elements, graphical features, and annotations should appear may need to be personalized to align with the observers' reasoning goals [105] and the narrative structure of the visualization.

6.3 Considerations of Image-based Method to Capture Task-Free Perceptual Experiences

In this work, we showed a quantitative approach that combines human subjective readings and objective metric measurement to study the mechanism underlying complex perceptual experiences. We aimed to create a representative and relatively large dataset with a broad coverage, while also minimizing human expert curation. Here, we have demonstrated how current data can be used to study "unobserved" complexity through the modeling of limited metrics and quantitative method. While big data has historically been used to address challenges in other disciplines, we can now leverage our own (relatively) big data to tackle computational problems by working with visualization-specific inquiries—the answer is always to serve and better understand human perception—while quantifying important underlying mechanisms.

Because this approach relies on data-driven significance and interpretive assessments across many factors, it helps identify a set of potential interpretable predictors. Our method measures VC at both the low-level metrics and the object-levels, allowing for the capturing of VCs that might be lost by visual design factors alone. The use of large, real-world image datasets enables us to capture a wide range of design variations, establishing a robust baseline for understanding cognitive processing in visualization, better than previously described at least for the node-link diagrams. When carefully chosen, these metrics can effectively reflect perceptual judgments [72]. As a result, models built on these features can offer a principled foundation for extracting high-level perceptual experiences from visualization inputs, and for future measuring how these experiences change in response to visual design variations.

We recommend that progress towards establishing mechanisms underlying VC through quantitative analysis should address three key challenges: *first*, the development of common and comprehensive image databases to support the assessment of perceptual experiences [113]; *second*, the design of experiments that directly capture high-level, perceptually driven experiences, such as meaningfulness, understandability, etc; *third*, the collection and validation of objective metrics capable of quantifying those perceptual experiences.

Our results revealed that human high-level understanding is supported not only by design-level stimuli but also by efficient object-level representations. The implication is to extend the current visualization representation units—whether derived from visual design or cognitive perception—by showing how they are formed by segmenting an image into understandable subsets. This reduction process is generally efficient and this understanding process is likely driven by visual complexity, which balances representational cost and memory load. By correlating complexity scores with model performance on visualization understanding tasks, one could also automatically flag overly complex figures for downstream processing, extending the impact of our work into the era of AI-driven data analysis.

6.4 Limitations and Future Work

Despite the broader range of metrics and assessments included in this study, limitations remain. Experiments like ours are often limited by the datasets used. For example, one could also validate our outcomes using a small dataset through item response theory for diagnostic efficiency, similar to approaches used in literacy testing [104]. Here we prioritize explainability based on large data and objective metrics. In addition, our study used images free of interaction or animation, viewed by the general public rather than domain experts. Also, our evaluation is task-free. One can use our method to incorporate different modalities. Finally, an exciting direction for future research would be to extend feature congestion modeling [117] to account for stimulus-specific complexity beyond continuous color-texture stimulus. In principle, metric-based modeling offers the potential for greater precision in capturing a large set of variables of how complexity is perceived.

7 Conclusion

Our empirical study systematically examines the relationship between human perception and objective image quality metrics. The go.osu.edu/viscomplexitydatacontains visualization stimulus types, human-rated VC scores, and 12 metric measurement scores. Our analyses reveal that the number of elements, whether defined by low-level features like edges and points or high-level boundaries formed by color and text, is a key factor in perceived visual complexity and shapes whether a visualization is perceived as scene-like.

Mathematical information-theoretic metrics contribute to perceived VC only when they incorporate high-level structural elements. The strong discriminative power of the feature congestion model is particularly intriguing, as it suggests that probabilistic characteristics may coexist meaningfully within visualizations. If these characteristics can be effectively integrated with scene-structure and information-theoretic metrics, it may be possible to extend feature congestion modeling to a broader range of stimulus types. As a result, integrating stimulus features, such as set-size, with statistical information-theoretic methods presents a promising pathway for systematically modeling variations of perceived visual complexity by human observers.

Ultimately, such an approach could provide a principled basis for interpreting high-level perceptual experiences across diverse visualization inputs, thereby enhancing the scientific and design potential of complexity quantification. Many aspects of visualization experience and metric behavior extend beyond what is observable in controlled lab-based experiments. This underscores the need to study real-world visualizations used in everyday contexts. To that end, the methodology used in this study, collecting subjective human ratings through large-scale crowdsourcing, offers a scalable and ecologically valid way to capture holistic human perception in real-world settings. The complementary insights gained from combining data-driven modeling and experimental methods have the potential to drive transformative advances - an opportunity that the visualization research community is well positioned to embrace.

Image Copyrights

Acknowledgments

The work was partially supported by the "Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2120/1 – 390831618" through a Visiting

Professorship Position, made to Prof. Dr. Jian Chen, via the Cluster of Excellent IntCDC, University of Stuttgart, Germany. The crowdsourcing experiment was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161, with The Ohio State University IRB approval ID: 2022B0363. R.S. Laramee was supported, in part, by funding from the EPSRC (EPSRC UKRI157, APP17227). We also wish to thank Dr. Ruth Rosenholtz at MIT for sharing the Matlab code for computing Feature Congestion, and colleagues at the Visual Attention Lab at Harvard University, as well as reviewers for their comments.

REFERENCES

[1] E. Adar and E. Lee. Communicative Visualizations as a Learning Problem. *IEEE TVCG*, 27(2):946–956, 2020. doi: 10/ghgt48 4

[2] K. Ajani, E. Lee, C. Xiong, C. N. Knaflic, W. Kemper, and S. Franconeri. Declutter and Focus: Empirically Evaluating Design Guidelines for Effective Data Communication. *IEEE TVCG*, 28(10):3351–3364, 2021. doi: 10/gn984r 1, 15

[3] Y. A. Andrienko, N. Brilliantov, and J. Kurths. Complexity of two-dimensional patterns. *Eur. Phys. J. B*, 15:539–546, 2000. doi: 10/c6bzm7 3, 6

[4] A. Ashokkumar and J. W. Pennebaker. Social media conversations reveal large psychological shifts caused by covid-19's onset across us cities. *Science advances*, 7(39):eabg7843, 2021. doi: 10/gm6vps 15

[5] F. Attneave. Physical determinants of the judged complexity of shapes. *J Exp Psychol*, 53(4):221, 1957. doi: 10/fhsj6w 2, 16

[6] F. B. Baker. *The basics of item response theory*. ERIC, 2001. 6

[7] M. Balzer and O. Deussen. Voronoi treemaps. In *IEEE INFOVIS*, pp. 49–56, 2005. doi: 10/b6xbtv 14

[8] G. Barua, T. Pingel, and T. Lim. Urban thermal map design considerations: color, shading, and resolution. *Cat & Geo Info Sci*, 51(4):549–564, 2024. doi: 10/pd8k 3

[9] R. Borgo, A. Abdul-Rahman, F. Mohamed, P. W. Grant, I. Reppa, L. Floridi, and M. Chen. An Empirical Study on Using Visual Metaphors in Visualization. *IEEE TVCG*, 18(12):2759–2768, 2012. doi: 10/npvk 1, 2, 3

[10] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What Makes a Visualization Memorable? *IEEE TVCG*, 19(12):2306–2315, 2013. doi: 10/f5h3pd 3, 4, 8, 14, 18

[11] R. P. Botchen, D. Weiskopf, and T. Ertl. Texture-based visualization of uncertainty in flow fields. In *IEEE Vis.*, pp. 647–654, 2005. doi: 10/c52bgr 14

[12] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10/c5bcq8 4, 15

[13] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE Trans Pattern Mach Intell*, 41(3):740–757, 2018. doi: 10/gfxgwp 4, 5

[14] A.-F. Cabouat, T. He, P. Isenberg, and T. Isenberg. PREVis: Perceived Readability Evaluation for Visualizations. *IEEE TVCG*, 31(1):1083–1093, 2025. doi: 10/njnz 2

[15] J. Canny. A Computational Approach to Edge Detection. *IEEE Trans Pattern Mach Intell*, pp. 679–698, 1986. doi: 10/fn3fdk 3

[16] X. J. Chai, N. Ofen, L. F. Jacobs, and J. D. Gabrieli. Scene complexity: influence on perception, memory, and development in the medial temporal lobe. *Front Hum Neurosci*, 4:1021, 2010. doi: 10/drkjb2 2, 3, 4, 16

[17] J. Chen, H. Cai, A. P. Auchus, and D. H. Laidlaw. Effects of Stereo and Screen Size on the Legibility of Three-Dimensional Streamtube Visualization. *IEEE TVCG*, 18(12):2130–2139, 2012. doi: 10/f4fqn5 2

[18] J. Chen, M. Ling, R. Li, P. Isenberg, T. Isenberg, M. Sedlmair, T. Möller, R. S. Laramee, H.-W. Shen, K. Wünsche, et al. VIS30K: A Collection of Figures and Tables From IEEE Visualization Conference Publications. *IEEE TVCG*, 27(9):3826–3833, 2021. doi: 10/gmsvxd 4, 5, 14

[19] J. Chen, M. Ling, R. Li, P. Isenberg, T. Isenberg, M. Sedlmair, T. Möller, R. Laramee, H.-W. Shen, K. Wünsche, and Q. Wang. IEEE VIS figures and tables image dataset. *IEEE Dataport*, 2020. doi: 10/kdqd 14, 18

[20] J. Chen, P. S. Pyla, and D. A. Bowman. Testbed evaluation of navigation and text display techniques in an information-rich virtual environment. In *IEEE VR*, pp. 181–289, 2004. doi: 10/dch9x2 9

[21] J. Chen, G. Zhang, W. Chiou, D. H. Laidlaw, and A. P. Auchus. Measuring the Effects of Scalar and Spherical Colormaps on Ensembles of DMRI Tubes. *IEEE TVCG*, 26(9):2818–2833, 2019. doi: 10/mcrw 3, 5

[22] M. Chen and H. Jäenicke. An Information-theoretic Framework for Visualization. *IEEE TVCG*, 16(6):1206–1215, 2010. doi: 10/d5rn2m 2, 8

[23] S. Chen, N. Andrienko, G. Andrienko, J. Li, and X. Yuan. Co-Bridges: Pair-wise Visual Connection and Comparison for Multi-item Data Streams. *IEEE TVCG*, 27(2):1612–1622, 2020. doi: 10/mzvm 14

[24] M.-T. Chi, S.-S. Lin, S.-Y. Chen, C.-H. Lin, and T.-Y. Lee. Morphable Word Clouds for Time-Varying Text Data Visualization. *IEEE TVCG*, 21(12):1415–1426, 2015. doi: 10/f7w6sq 14

[25] Y. Collet and M. Kucherawy. Zstandard Compression and the application/zstd Media Type. In *RFC8878*, 2018. doi: 10/nvxc 3, 5

[26] S. Corchs, F. Gasparini, and R. Schettini. No reference image quality classification for JPEG-distorted images. *Dgt Sigl Proc*, 30, 2014. doi: 10/f54jpv 16

[27] S. E. Corchs, G. Ciocca, E. Bricolo, and F. Gasparini. Predicting Complexity Perception of Real World Images. *PLoS one*, 11(6), 2016. doi: 10/gbpnrm 2, 4, 16

[28] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. TextFlow: Towards better understanding of evolving topics in text. *IEEE TVCG*, 17(12):2412–2421, 2011. doi: 10/fnpjpg 14

[29] F. R. De Siqueira, W. R. Schwartz, and H. Pedrini. Multi-scale gray level co-occurrence matrices for texture description. *Neurocomp*, 120:336–345, 2013. doi: 10/f5cfhx 3, 6

[30] D. DeCarlo, A. Finkelstein, S. Rusinkiewicz, and A. Santella. Suggestive Contours for Conveying Shape. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 401–408. ACM, 2023. doi: 10/bpk44q 15

[31] K. Dinkla, M. A. Westenberg, and J. J. van Wijk. Compressed Adjacency Matrices: Untangling Gene Regulatory Networks. *IEEE TVCG*, 18(12), 2012. doi: 10/f4fwnz 14

[32] D. C. Donderi. An Information Theory Analysis of Visual Complexity and Dissimilarity. *Perception*, 35(6):823–835, 2006. doi: 10/b6tfk4 4

[33] D. C. Donderi. Visual complexity: A review. *Psychological bulletin*, 132(1):73, 2006. doi: 10/frf92v 2

[34] Y. Du, C. Li, R. Guo, X. Yin, W. Liu, J. Zhou, Y. Bai, Z. Yu, Y. Yang, Q. Dang, et al. PP-OCR: A Practical Ultra Lightweight OCR System. *ArXiv*, 2020. doi: 10/nvxk 5

[35] B. Duffy, A. Dasgupta, R. Kosara, S. Walton, and M. Chen. Measuring Visual Complexity of Cluster-Based Visualizations. *ArXiv*, 2013. doi: 10/nvxd 1, 2

[36] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing Objects by their Attributes. In *Proc. CVPR*, pp. 1778–1785, 2009. doi: 10/cwdbn7 2

[37] M. Feixas, E. Del Acebo, P. Bekaert, and M. Sbert. An Information Theory Framework for the Analysis of Scene Complexity. In *Comput Graph Forum*, vol. 18(3), pp. 95–106, 1999. doi: 10/cxk5k3 1, 9

[38] C. Fernandez-Lozano, A. Carballal, P. Machado, A. Santos, and J. Romero. Visual complexity modelling based on image features fusion of multiple kernels. *PeerJ*, 7, 2019. doi: 10/mzzt 2, 16

[39] T. Fujiwara, N. Sakamoto, J. Nonaka, K. Yamamoto, K.-L. Ma, et al. A Visual Analytics Framework for Reviewing Multivariate Time-Series Data with Dimensionality Reduction. *IEEE TVCG*, 27(2):1601–1611, 2020. doi: 10/ghgt65 14

[40] R. M. Gray. *Entropy and Information Theory*. Sci. Bus. Media, 2011. doi: 10/gnms9f 3

[41] M. R. Greene. The information content of scene categories. In *Psychol. Learn. Motiv.*, vol. 70. Elsevier, 2019. doi: 10/nvxg 2, 4, 7

[42] T. Gschwandtnei, M. Bögl, P. Federico, and S. Miksch. Visual Encodings of Temporal Uncertainty: A Comparative User Study. *IEEE TVCG*, 22(1):539–548, 2015. doi: 10/ghppzj 14

[43] S. Harper, E. Michailidou, and R. Stevens. Toward a Definition of Visual Complexity as an Implicit Measure of Cognitive Load. *ACM Trans Appl Percept*, 6(2):1–18, 2009. doi: 10/bnfh3w 16

[44] C. Harris, M. Stephens, et al. A Combined Corner and Edge Detector. In *Alvey Vis. Conf.*, 1988. doi: 10/gfvtg5 6

[45] J. Hartmann, A. Sutcliffe, and A. D. Angeli. Towards a Theory of User Judgment of Aesthetics and User Interface Quality. *ACM Trans Comput Hum Interact*, 15(4):1–30, 2008. doi: 10/b57cf9 1

[46] D. Hasler and S. E. Suesstrunk. Measuring colorfulness in natural images. In *Hum Vis Elec Imaging VIII*, vol. 5007, pp. 87–95. SPIE, 2003. doi: 10/bj98jc 3, 4

[47] T. He, P. Isenberg, R. Dachselt, and T. Isenberg. BeauVis: A Validated Scale for Measuring the Aesthetic Pleasure of Visual Representations. *IEEE TVCG*, 29(1):363–373, 2022. doi: 10/kt3n 1, 2, 4, 14

[48] C. Heaps and S. Handel. Similarity and features of natural textures. *J Exp Psychol*, 25(2):299, 1999. doi: 10/cvbnjw 1
[49] M. A. Hearst. Show It or Tell It? Text, Visualization, and Their Combination. *ACM CACM*, pp. 68–75, 2023. doi: 10/mzzv 1, 9
[50] J. Heer and M. Stone. Color Naming Models for Color Selection, Image Editing and Palette Design. In *ACM SIGCHI*, pp. 1007–1016, 2012. doi: 10/gfz5q4 3, 5
[51] H. Heft and J. L. Nasar. Evaluating Environmental Scenes Using Dynamic Versus Static Displays. *Environ Behav*, 32(3):301–322, 2000. doi: 10/bzvg7h 2, 16
[52] R. Herbrich, T. Minka, and T. Graepel. TrueSkill™: A Bayesian Skill Rating System. *Adv in Neural Info Proc Sys*, p. 569–576, 8 pages, 2006. doi: 10/pd8n 4, 5
[53] F. Heylighen. The Growth of Structural and Functional Complexity during Evolution. *Evol. Complex.*, 8:17–44, 1999. 2, 3
[54] M. Hlawatsch, P. Leube, W. Nowak, and D. Weiskopf. Flow Radar Glyphs—Static Visualization of Unsteady Flow with Uncertainty. *IEEE TVCG*, 17(12):1949–1958, 2011. doi: 10/cvc442 14
[55] G. Hu, Z. Pan, M. Zhang, D. Chen, W. Yang, and J. Chen. An interactive method for generating harmonious color schemes. *Color Research & Application*, 39(1):70–78, 2014. doi: 10/f5jsw3 4
[56] D. E. Huber and C. G. Healey. Visualizing Data with Motion. In *IEEE Vis.*, pp. 527–534, 2005. doi: 10/fjpgkx 14
[57] J. Hullman, E. Adar, and P. Shah. Benefitting InfoVis with Visual Difficulties. *IEEE TVCG*, 17(12):2213–2222, 2011. doi: 10/df5c8r 9
[58] S. Huron, Y. Jansen, and S. Carpendale. Constructing Visual Representations: Investigating the Use of Tangible Tokens. *IEEE TVCG*, 20(12):2102–2111, 2014. doi: 10/f6qjtr 14
[59] A. M. Iliyasu, A. K. Al-Asmari, M. A. AbdelWahab, A. S. Salama, M. A. Al-Qodah, A. R. Khan, P. Q. Le, and F. Yan. Mining visual complexity of images based on an enhanced feature space representation. In *IEEE Int Symp Intell Signal Process*, pp. 65–70, 2013. doi: 10/mzz4 2, 16
[60] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. Vispubdata.org: A Metadata Collection About IEEE Visualization (VIS) Publications. *IEEE TVCG*, 23(9):2199–2206, 2016. doi: 10/ggwwrv 14
[61] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller. Visualization as Seen through its Research Paper Keywords. *IEEE TVCG*, 23(1):771–780, 2016. doi: 10/f92gps 14
[62] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A Systematic Review on the Practice of Evaluating Visualization. *IEEE TVCG*, 19(12), 2013. doi: 10/f5h29z 14
[63] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Proc. CVPR*, pp. 145–152, 2011. doi: 10/cznr72 4
[64] J. Jakob, M. Gross, and T. Günther. A Fluid Flow Data Set for Machine Learning and its Application to Neural Flow Map Interpolation. *IEEE TVCG*, 27(2):1279–1289, 2020. doi: 10/pfhz 14
[65] H. Jänicke and M. Chen. A Salience-based Quality Metric for Visualization. In *Comput Graph Forum*, vol. 29(3), pp. 1183–1192, 2010. doi: 10/dpqxdr 2, 8
[66] H. Jänicke and G. Scheuermann. Visual Analysis of Flow Features Using Information Theory. *IEEE CG&A*, 30(1):40–49, 2009. doi: 10/bpkjdh 3
[67] H. Jänicke, A. Wiebel, G. Scheuermann, and W. Kollmann. Multifield Visualization Using Local Statistical Complexity. *IEEE TVCG*, 13(6):1384–1391, 2007. doi: 10/bzbh7k 2, 3, 8
[68] C. P. Kappe, L. Schuetz, S. Gunther, L. Hufnagel, S. Lemke, and H. Leitte. Reconstruction and Visualization of Coordinated 3D Cell Migration Based on Optical Flow. *IEEE TVCG*, 22(1):995–1004, 2015. doi: 10/pfjr 14
[69] B. Kerr. THREAD ARCS: An email thread visualization. In *IEEE InfoVis*, pp. 211–218, 2003. doi: 10/bsb89f 14
[70] G. Kindlmann and C.-F. Westin. Diffusion Tensor Visualization with Glyph Packing. *IEEE TVCG*, 12(5):1329–1336, 2006. doi: 10/cqqrk5 14
[71] P. Kondratieva, J. Kruger, and R. Westermann. The Application of GPU Particle Tracing to Diffusion Tensor Field Visualization. In *IEEE Vis.*, pp. 73–78, 2005. doi: 10/cb8g2k 14
[72] M. A. Kramer, M. N. Hebart, C. I. Baker, and W. A. Bainbridge. The features underlying the memorability of objects. *Science advances*, 9(17), 2023. doi: 10/gr72gs 1, 9, 15
[73] A. Krishnan, L. J. Williams, A. R. McIntosh, and H. Abdi. Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review. *Neuroimage*, 56(2):455–475, 2011. doi: 10/bdcbp4 15
[74] C. Kyle-Davidson, O. Solis, S. Robinson, R. T. W. Tan, and K. K. Evans. Scene complexity and the detail trace of human long-term visual memory. *Vision Research*, 227:108525, 2025. doi: 10/pfj3 2, 15
[75] C. Kyle-Davidson, E. Y. Zhou, D. B. Walther, A. G. Bors, and K. K. Evans. Characterising and dissecting human perception of scene complexity. *Cognition*, 231:105319, 2023. doi: 10/mzz5 15
[76] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE TVCG*, 18, 2012. doi: 10/drrh6j 14
[77] P. Q. Le, A. M. Iliyasu, J. Garcia, F. Dong, and K. Hirota. Representing Visual Complexity of Images Using a 3D Feature Space Based on Structure, Noise, and Diversity. *J Adv Comp Intell*, 16(5), 2012. doi: 10/mzz6 2, 16
[78] F. Lekschas, B. Bach, P. Kerpedjiev, N. Gehlenborg, and H. Pfister. HiPiler: Visual Exploration of Large Genome Interaction Matrices with Interactive Small Multiples. *IEEE TVCG*, 24(1):522–531, 2017. doi: 10/gcp7n9 14
[79] R. Li and J. Chen. Toward A Deep Understanding of What Makes a Scientific Visualization Memorable. In *IEEE SciVis*, pp. 26–31, 2018. doi: 10/gr633t 3, 4
[80] Y. Lin, W. Zeng, Y. Ye, and H. Qu. Saliency-aware color harmony models for outdoor signboard. *Comp Graph*, 105:25–35, 2022. doi: 10/nvxm 3
[81] D. Liu, P. Xu, and L. Ren. TPFlow: Progressive partition and multidimensional pattern extraction for large-scale spatio-temporal data analysis. *IEEE TVCG*, 25(1):1–11, 2018. doi: 10/gjh2k9 14
[82] Z. Liu, Y. Wang, M. Dontcheva, M. Hoffman, S. Walker, and A. Wilson. Patterns and Sequences: Interactive Exploration of Clickstreams to Understand Common Visitor Paths. *IEEE TVCG*, 23:321–330, 2017. doi: 10/f92f77 14
[83] G. L. Lohse, K. Biolsi, N. Walker, and H. H. Rueter. A Classification of Visual Representations. *ACM CACM*, 37(12):36–50, 1994. doi: 10/b92kbj 14
[84] G. Lommerse, F. Nossin, L. Voinea, and A. Telea. The Visual Code Navigator: An Interactive Toolset for Source Code Investigation. In *IEEE INFOVIS*, pp. 24–31, 2005. doi: 10/d8gnmz 14
[85] K. Lu, M. Feng, X. Chen, M. Sedlmair, O. Deussen, D. Lischinski, Z. Cheng, and Y. Wang. Palettailor: Discriminable Colorization for Categorical Data. *IEEE TVCG*, 27(2):475–484, 2020. doi: 10/gr632z 14
[86] M. L. Mack and A. Oliva. The perceptual dimensions of visual simplicity. *J. Vis*, 4(719), 2004. doi: 10/frr5nb 1, 2, 3, 6
[87] I. Mario, M. Chacon, D. Alma, and S. Corral. Image Complexity Measure: a Human Criterion Free Approach. In *Annu. Meet. North Am. Fuzzy Inf. Process. Soc.*, pp. 241–246, 2005. doi: 10/bdqq23 2, 16
[88] V. Matvienko and J. Krüger. Explicit Frequency Control for High-Quality Texture-Based Flow Visualization. In *IEEE SciVis*, pp. 41–48, 2015. doi: 10/k8sg 14
[89] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø. A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and intelligent laboratory systems*, 118:62–69, 2012. doi: 10/f4hx5x 6
[90] A. Mikhailiuk, C. Wilmot, M. Perez-Ortiz, D. Yue, and R. K. Mantiuk. Active Sampling for Pairwise Comparisons via Approximate Message Passing and Information Gain Maximization. In *IEEE Int Conf Pattern Recognit*, pp. 2559–2566, 2021. doi: 10/mzz8 4, 5, 15, 17
[91] S. Miksch, H. Leitte, and M. Chen. Knowledge-Assisted Visualization and Guidance. *Found. Data Vis.*, pp. 61–85, 2020. doi: 10/nvxn 3
[92] A. Miniukovich and A. De Angeli. Quantification of Interface Visual Complexity. In *Proc. Inter. Working Conf. on Adv. Visual Interfaces*, pp. 153–160, 2014. doi: 10/gm7pmd 2, 16
[93] A. Miniukovich, S. Sulpizio, and A. De Angeli. Visual Complexity of Graphical User Interfaces. In *Proc Int Conf Adv Vis Interfaces*, pp. 1–9, 2018. doi: 10/gh67h3 16
[94] T. Minka, R. Cleven, and Y. Zaykov. TrueSkill 2: An improved Bayesian skill rating system. Technical report, Microsoft Research, 2018. 15
[95] V. Molchanov and L. Linsen. Shape-preserving Star Coordinates. *IEEE TVCG*, 25(1):449–458, 2018. doi: 10/nvzg 14
[96] T. Munzner. *Visualization Analysis and Design*. CRC press, 2014. doi: 10/gd3xgq 8
[97] F. Nagle and N. Lavie. Predicting human complexity perception of real-world scenes. *Royal Soc Open Sci*, 7(5):191487, 2020. doi: 10/gq6f3m 4, 15
[98] P. H. Nguyen, K. Xu, A. Bardill, B. Salman, K. Herd, and B. L. W. Wong. SenseMap: Supporting browser-based online sensemaking through analytic provenance. *IEEE VAST*, pp. 91–100, 2016. doi: 10/ggd957 14
[99] A. Oliva, M. L. Mack, M. Shrestha, and A. Peeper. Identifying the

Perceptual Dimensions of Visual Complexity of Scenes. In *Proc Annu Meet Cogn Sci Soc*, vol. 26(26), 2004. 1, 2, 3, 4, 7, 14, 15, 16

[100] A. Oliva and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *Int J Comput Vis*, 42:145–175, 2001. doi: 10/dm4d3b 2

[101] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Prog. Brain Res.*, 155:23–36, 2006. doi: 10/b2jzn8 5

[102] M. A. Otaduy, N. Jain, A. Sud, and M. C. Lin. Haptic Display of Interaction between Textured Models. In *ACM SIGGRAPH Courses*, pp. 133–es, 2005. doi: 10/b95rph 14

[103] S. Palan and C. Schitter. Prolific.ac—A subject pool for online experiments. *J. Behav. Exp. Finance*, 17:22–27, 2018. doi: 10/gftct9 4

[104] S. Pandey and A. Ottley. Mini-VLAT: A Short and Effective Measure of Visualization Literacy. In *Comput Graph Forum*, vol. 42(3), pp. 1–11. Wiley Online Library, 2023. doi: 10/gtjwqb 9

[105] R. Y. Pang, K. Feng, S. Feng, C. Li, W. Shi, Y. Tsvetkov, J. Heer, and K. Reinecke. Interactive reasoning: Visualizing and controlling chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2506.23678*, 2025. doi: 10/pzxr 9

[106] H. Pfister, B. Lorensen, C. Bajaj, G. Kindlmann, W. Schroeder, L. S. Avila, K. Raghu, R. Machiraju, and J. Lee. The Transfer Function Bake-Off. *IEEE CG&A*, 21(3):16–22, 2001. doi: 10/bwv59n 3

[107] H. C. Purchase, E. Freeman, and J. Hamer. An Exploration of Visual Complexity. In *Proc Int Conf Diagram Represent Infer*, pp. 200–213, 2012. doi: 10/mz2b 1, 2, 4, 6, 7, 8, 14, 16, 27

[108] D. Rafiei. Effectively Visualizing Large Networks Through Sampling. In *IEEE Vis.*, pp. 375–382, 2005. doi: 10/dj6nf6 14

[109] M. D. Rahman, G. J. Quadri, B. Doppalapudi, D. A. Szafir, and P. Rosen. A Qualitative Analysis of Common Practices in Annotations: A Taxonomy and Design Space. *IEEE TVCG*, 31(1):360–370, 2025. doi: 10/nvr9 2, 4, 5

[110] G. Ramanarayanan, K. Bala, J. A. Ferwerda, and B. Walter. Dimensionality of visual complexity in computer graphics scenes. In *Hum Vis Elec Imaging XIII*, vol. 6806, pp. 142–151, 2008. doi: 10/fg9chj 2, 16

[111] A. R. Rao and G. L. Lohse. Towards a Texture Naming System: Identifying Relevant Dimensions of Texture. *Vision Research*, 36(11):1649–1669, 1996. doi: 10/fn7t6w 14

[112] K. Reinecke, T. Yeh, L. Miratrix, R. Mardiko, Y. Zhao, J. Liu, and K. Z. Gajos. Predicting Users' First Impressions of Website Aesthetics With a Quantification of Perceived Visual Complexity and Colorfulness. In *ACM SIGCHI*, pp. 2049–2058, 2013. doi: 10/gfs76j 4

[113] R. A. Rensink. Visualization as a stimulus domain for vision science. *J Vision*, 21(8):3–3, 2021. doi: 10/mg29 9

[114] I. Reppa, D. Playfoot, and S. J. McDougall. Visual Aesthetic Appeal Speeds Processing of Complex but not Simple Icons. In *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 52(18), pp. 1155–1159, 2008. doi: 10/fx6f5r 1

[115] J. Rigau, M. Feixas, and M. Sbert. Conceptualizing Birkhoff's Aesthetic Measure Using Shannon Entropy and Kolmogorov Complexity. In *Comput. Aesthet. Graph. Vis. Imaging*, pp. 105–112, 2007. doi: 10/m9kx 3

[116] J. Rigau, M. Feixas, and M. Sbert. Informational Aesthetics Measures. *IEEE CG&A*, 28(2):24–34, 2008. doi: 10/dz9m2k 4

[117] R. Rosenholtz. Visual Attention in Crisis. *Behavioral and Brain Sciences*, pp. 1–32, 2024. doi: 10/m9kw 3, 4, 9

[118] R. Rosenholtz, Y. Li, J. Mansfield, and Z. Jin. Feature Congestion: A Measure of Display Clutter. In *ACM SIGCHI*, pp. 761–770, 2005. doi: 10/bhtrg7 1, 2, 3, 4, 5, 6, 7

[119] R. Rosenholtz, Y. Li, and L. Nakano. Measuring visual clutter. *J Vision*, 7(2):1–22, 2007. doi: 10/bqtpr4 1, 2, 3, 4, 5, 6, 8, 14, 16, 26

[120] S. Rufiange and M. J. McGuffin. DiffAni: Visualizing Dynamic Graphs with a Hybrid of Difference Maps and Animation. *IEEE TVCG*, 19(12):2556–2565, 2013. doi: 10/f5h3wc 14

[121] G. Ryan, A. Mosca, R. Chang, and E. Wu. At a Glance: Pixel Approximate Entropy as a Measure of Line Chart Complexity. *IEEE TVCG*, 25(1):872–881, 2018. doi: 10/gf5khg 16

[122] A. Sanderson, G. Chen, X. Tricoche, D. Pugmire, S. Kruger, and J. Breslau. Analysis of Recurrent Patterns in Toroidal Magnetic Fields. *IEEE TVCG*, 16(6):1431–1440, 2010. doi: 10/fk85hx 14

[123] E. Saraee, M. Jalal, and M. Betke. Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding*, 195:102949, 2020. doi: 10/gpzpdb 1, 2, 4, 15, 16

[124] K. Sarıtaş, P. Dayan, K. Shen, and S. S. Nath. Complexity in Complexity: Understanding Visual Complexity Through Structure, Color, and Surprise. In *2^nd on Representational Alignment at ICLR*, 2025. doi: 10/pzj3 15

[125] V. Setlur and M. C. Stone. A Linguistic Approach to Categorical Color Assignment for Data Visualization. *IEEE TVCG*, 22(1):698–707, 2015. doi: 10/gr6324 14

[126] C. E. Shannon. A Mathematical Theory of Communication. *ACM Mob Comput Commun Rev*, 5(1):3–55, 2001. doi: 10/b39t 2, 5

[127] S. Silva, B. S. Santos, and J. Madeira. Using color in visualization: A survey. *Comput Graph*, 35(2):320–333, 2011. doi: 10/b59k9m 3

[128] J. G. Snodgrass and M. Vanderwart. A Standardized Set of 260 Pictures: Norms for Name Agreement, Image Agreement, Familiarity, and Visual Complexity. *J Exp Psychol*, 6(2):174, 1980. doi: 10/ftkftb 1, 2, 4, 16

[129] S. Sonkusare, M. Breakspear, and C. Guo. Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in cognitive sciences*, 23(8):699–714, 2019. doi: 10/gf4r7b 1

[130] T. C. Sprenger, R. Brunella, and M. H. Gross. *H-BLOB: A Hierarchical Visual Clustering Method Using Implicit Surfaces*. IEEE Vis., 2000. doi: 10/cvr4gs 14

[131] M. Stone and L. Bartram. Alpha, contrast and the perception of visual metadata. In *Color and Imaging conference*, vol. 16, pp. 355–359. Society of Imaging Science and Technology, 2008. doi: 10/pzvz 9

[132] Z. Sun and C. Firestone. Curious Objects: How Visual Complexity Guides Attention and Engagement. *Cog Sci*, 45(4):e12933, 2021. doi: 10/gpjgm5 2, 16

[133] R. Tibshirani. Regression Shrinkage and Selection Via the Lasso. *J of the Royal Stat Soc Ser B: Stat Metho*, 58(1):267–288, 1996. doi: 10/gfn45m 6

[134] A. N. Tuch, J. A. Bargas-Avila, K. Opwis, and F. H. Wilhelm. Visual complexity of websites: Effects on users' experience, physiology, performance, and memory. *Int J Hum Comput Stud*, 67(9):703–715, 2009. doi: 10/fk2j8c 1

[135] A. M. Turing. *Computing machinery and intelligence*. Springer, 2009. 8

[136] C. Tuttle, L. G. Nonato, and C. Silva. PedVis: A Structured, Space-Efficient Technique for Pedigree Visualization. *IEEE TVCG*, 16(6):1063–1072, 2010. doi: 10/fk22z7 14

[137] I. Viola and T. Isenberg. Pondering the Concept of Abstraction in (Illustrative) Visualization. *IEEE TVCG*, 24(9):2573–2588, 2017. doi: 10/gd3k7m 15

[138] S. Wallace, Z. Bylinskii, J. Dobres, B. Kerr, S. Berlow, R. Treitman, N. Kumawat, K. Arpin, D. B. Miller, J. Huang, et al. Towards Individuated Reading Experiences: Different Fonts Increase Reading Speed for Different Individuals. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(4):1–56, 2022. doi: 10/gqg7jp 4

[139] G. Wang, J. Guo, M. Tang, J. F. de Queiroz Neto, C. Yau, A. Daghistani, M. Karimzadeh, W. G. Aref, and D. S. Ebert. STULL: Unbiased Online Sampling for Visual Exploration of Large Spatiotemporal Data. In *IEEE Conf Vis Anal Sci Technol*, pp. 72–83, 2020. doi: 10/k8ss 14

[140] C. Ware. Color sequences for univariate maps: Theory, experiments and principles. *IEEE CG&A*, 8(5):41–49, 1988. doi: 10/cz578k 3

[141] C. Ware, M. Stone, and D. A. Szafir. Rainbow Colormaps Are Not All Bad. *IEEE CG&A*, 43(3):88–93, 2023. doi: 10/pd8q 3

[142] M. Wattenberg. A Note on Space-Filling Visualizations and Space-Filling Curves. In *IEEE INFOVIS*, pp. 181–186, 2005. doi: 10/dsngsw 14

[143] Y. Wei, H. Mei, Y. Zhao, S. Zhou, B. Lin, H. Jiang, and W. Chen. Evaluating Perceptual Bias During Geometric Scaling of Scatterplots. *IEEE TVCG*, 26(1):321–331, 2019. doi: 10/mzvh 14

[144] C. Weigle and R. M. Taylor. Visualizing intersecting surfaces with nested-surface techniques. In *IEEE Vis.*, pp. 503–510, 2005. doi: 10/dwgstw 14

[145] M. Weiß, K. Angerbauer, A. Voit, M. Schwarzl, M. Sedlmair, and S. Mayer. Revisited: Comparison of Empirical Methods to Evaluate Visualizations Supporting Crafting and Assembly Purposes. *IEEE TVCG*, 27(2):1204–1213, 2020. doi: 10/ghgt5z 14

[146] D. Whitney and A. Yamanashi Leib. Ensemble Perception. *Annu Rev Psychol*, 69:105–129, 2018. doi: 10/gfs6hb 5

[147] W. Willett, J. Heer, and M. Agrawala. Scented Widgets: Improving Navigation Cues with Embedded Visualizations. *IEEE TVCG*, 13(6):1129–1136, 2007. doi: 10/dpbgbm 9

[148] A. Wilson and R. Brannon. Exploring 2D Tensor Fields Using Stress Nets. In *IEEE Vis.*, pp. 11–18, 2005. doi: 10/fp72z7 14

[149] F. Windhager, A. Abdul-Rahman, M.-J. Bludau, N. Hengesbach,

H. Lamqaddam, I. Meirelles, B. Speckmann, and M. Correll. Complexity as Design Material Position Paper. In *IEEE BELIV*, pp. 71–80, 2024. doi: 10/nvwm 1, 2

[150] J. M. Wolfe. Guided Search 6.0: An updated model of visual search. *Psychon. Bull. Rev.*, 28(4):1060–1092, 2021. doi: 10/gh2s45 8

[151] L. H. Wurm, G. E. Legge, L. M. Isenberg, and A. Luebker. Color Improves Object Recognition in Normal and Low Vision. *J. Exp. Psychol. Hum. Percept. Perform.*, 19(4):899, 1993. doi: 10/b9w9k9 5

[152] L. Xu, T.-Y. Lee, and H.-W. Shen. An Information-Theoretic Framework for Flow Visualization. *IEEE TVCG*, 16(6):1216–1224, 2010. doi: 10/crhtcg 14

[153] J. Yang, A. Patro, S. Huang, N. Mehta, M. O. Ward, and E. A. Rundensteiner. Value and Relation Display for Interactive Exploration of High Dimensional Datasets. In *IEEE InfoVis*, pp. 73–80, 2004. doi: 10/cs7zp8 14

[154] J. Yang-Peláez and W. C. Flowers. Information Content Measures of Visual Displays. In *IEEE InfoVis*, pp. 99–103, 2000. doi: 10/bq34sd 2, 8

[155] V. Yoghourdjian, D. Archambault, S. Diehl, T. Dwyer, K. Klein, H. C. Purchase, and H.-Y. Wu. Exploring the limits of complexity: A survey of empirical studies on graph visualisation. *Vis Inf*, 2(4):264–282, 2018. doi: 10/ggjnn5 1, 2

[156] M. Zhang, L. Chen, Q. Li, X. Yuan, and J. Yong. Uncertainty-Oriented Ensemble Data Visualization and Exploration using Variable Spatial Spreading. *IEEE TVCG*, 27(2):1808–1818, 2020. doi: 10/ghgt7p 14

[157] H. Zhao, G. W. Bryant, W. Griffin, J. E. Terrill, and J. Chen. Evaluating Glyph Design for Showing Large-Magnitude-Range Quantum Spins. *IEEE TVCG*, 30(4):1868–1884, 2022. doi: 10/m9kv 5

[158] Y. Zhao, H. Jiang, Q. Chen, Y. Qin, H. Xie, Y. Wu, S. Liu, Z. Zhou, J. Xia, and F. Zhou. Preserving Minority Structures in Graph Sampling. *IEEE TVCG*, 27:1698–1708, 2020. doi: 10/gt6nb8 14

[159] L. Zhou and C. D. Hansen. A Survey of Colormaps in Visualization. *IEEE TVCG*, 22(8):2051–2069, 2015. doi: 10/f8wtp5 3

What Makes a Visualization Image Complex?

Additional material

I Reproducibility: VisComplexity2K: Perceived Visual Complexity Scores, and Objective Image-Quality Metric Databases

Our perceived visual complexity (VC) dataset, VisComplexity2K, consists of 1,800 visualization images, each annotated with VC scores collected from human participants via the online crowd-sourcing platform Prolific. The dataset also includes corresponding objective image-quality metrics and associated metadata. VisComplexity2K is publicly available through a Google spreadsheet `go.osu.edu/viscomplexitydata`. All data analysis code and images can be accessed via Colab on `go.osu.edu/vcgoogledrive`. `osf.io/bdet6` also has all the analysis results and source code. The key columns of VisComplexity2K include:

A The **VIS30K paper DOI** or **the magazine or web links (column D)** as a unique identifier to cross-link to other databases such as Vis30K [18], VisPubData [60], KeyVis [61], and the Practice of Evaluating Visualization [62, 76], as well as MASSVIS [10].

B The **image name (column C)**, either the MASSVIS or the VIS30K source image names, as a unique identifier to cross-link to other image datasets, e. g., MASSVIS [10] and VIS30K [19].

C The **thumbnail of each image (column U)**, either the original image or a sub-panel from the source image, provides a gateway for image annotation and analysis, e. g., through Google CoLab run-time execution `go.osu.edu/vcmodel`.

D An indicator **(column J)** specifying the **HP-Trees and Networks** visualization images, which are the direct node-link diagrams used in Purchase et al. [107] and the **RR-Color and Textured** images, representing the texture and color stimuli used in Rosenholtz et al. [119] to support the analysis in the main text.

E The **image link (column K)** that points to a web storage address where a full-resolution version is accessible through `go.osu.edu/vcgoogledrive`. Please note that all VIS30K image files are copyrighted, and for most the copyright is owned by IEEE. Some have creative commons licenses or are in the public domain. Yet other images are subject to different, specific copyrights as indicated in the figure caption in the paper.

F The **perceived visual complexity (column M)** for the images collected through the crowdsourcing. The interface to collect this data is shown in Apdx. Figure 10.

G The **O.MeC score (column BN)** and The **O.TiR score (column BJ)**, corresponding to the number of color and text-ink-ratio, respectively. The Python code for computing these is available in the sub-folders *Metric.MeaningfulColor(O.MeC)* and *Metric.TexttoInkRatio(O.TiR)* on `go.osu.edu/vcgoogledrive`.

H The other **10 image-metrics**, as shown in **column M to BE**. The Python code for computing these is available in the sub-folder *Metrics.10* on `go.osu.edu/vcgoogledrive`.

II Figure Image Courtesy by References

Since IEEE VIS 2025 has a 2-page citation limits and we used many images from publications, we provide this figure-by-figure citation list to credit the image source. Images in Figure 1 are taken from [10, 23, 28, 39, 42, 54, 56, 64, 69, 71, 78, 84, 102, 120, 122, 136, 148, 153]. Images in Figure 4 are taken from [10, 11]. Images in Figure 5 are taken from [10, 108, 145]. Images in Figure 8 are taken from [81, 108]. Images in Figure 9 are taken from [64, 88, 142, 156].

Images in Apdx. Figure 10 are taken from [31, 130]. Images in Apdx. Figure 21 are taken from [10, 24, 31, 70, 85, 95, 125, 139, 152]. Images in Apdx. Figure 13 are taken from [7, 23, 28, 42, 56, 58, 68, 69, 82, 84, 95, 98, 122, 143, 144, 148, 158].

III Addition Experimental Choices and Results

III.1 Experimental Data Samples

The main criteria when we chose the data was diversity of stimuli. It is a large, heterogeneous dataset spanning multiple visualization (node-link diagrams, heatmaps, charts), drawn from real-world sources, in order to increase ecological validity and generalizability of findings.

III.2 A Workshop and Pilot Studies

There are various approaches to collecting subjective VC experiences, whether dealing with a small set of visualization images or a large set common in vision science. To determine the most suitable method for gathering VC scores, we piloted several approaches including both low-fidelity manual sorting and high-fidelity active sampling-based experiments. This was necessary because a large-scale data gathering of visual experience data for thousands of visualization images had not been established previously in the visualization literature.

The first approach involved asking raters to score a small set of images (about 20-100). For example, He et al. [47] used 15 images to collect absolute beauvis scores. We tested this in our exploratory workshop but found it difficult to align scores across participants when attempting to rank a larger set of 200 images.

The second approach involved a hierarchical division of images into groups. For example, in the study by Oliva et al. [99], participants divided 100 images into four hierarchies and stated their criteria to categorize natural scenes as ‘openness’ etc. This is also the standard de facto approach to measure other perceptual dimensions from images (e. g., seminal work of texture and visualization image classifications [83, 111]). We piloted this approach in our workshop and found it effective for gathering participants’ comments to guide subsequent steps. However, similar to the first approach, it presented scalability challenges, limiting its applicability for larger datasets. During the workshop experiment, 46 participants, organized into 20 groups of 2-3 students, were instructed to perform the hierarchical sorting tasks based on perceived visual complexity of the images, defined as “*the amount of detail or intricacy in images*” [107]. These participants were graduate students having domain expertise in machine learning and computer vision, before they took the visualization class; and three participants had extensive industrial experience. Participants were asked to document the factors that influenced their sorting decisions.

Participants on day 1 sorted 20 images, with all groups completing the task in an average of 20 minutes. The scalability challenge arrived on day 2 for sorting 50 images, with the same instruction. Participants described the sorting task as "overwhelming." No group was able to finish within 1.5 hours, and only 50% of the groups managed to complete the task within 2 hours. First, participants noted that increasing image set made any linear sorting impractical; and two groups utilized a 2D Cartesian coordinate system to organize VC clusters. Second, their comments suggested the benefits of text display: without the knowledge of the data source and differences between MASSVIS and VIS30K, 11 groups pointed out that some (MASSVIS) images ‘*looked nice*’ and these (VIS30K) *show sophisticated scientific phenomena*. ‘*Text (annotated) greatly improved clarity; thus I rated the image with lower complexity*’. Participants also commented that ‘*After reading the texts, I know where to look.*’ Others mentioned object set-size and said ‘*there were not many items in the image; thus these (MASSVIS) images were simpler.*’ Third, half of the groups commented that “*these (VIS30K scientific visualization) images were difficult to understand without text or legend showing what an image represents.*” These subjective experiences led us to exclude 3D surface and volume rendering visualizations from our formal experiment, as they were challenging for participants and were likely to be rated more complex.

The third approach is active-sampling based, similar to how game competition assigns competing players [12] to score human players on a global scale. For example, both Bradley-Terry ranking algorithm [12] and Microsoft's TrueSkill 2™ [94, 97] enabled scoring thousands of images in computer vision [123]. We also tried this for our pilot study with 200 images and found that this approach was reliable. However, scaling up this approach to 1,800 images with one-time sampling would require a significantly larger number of trials, making it nearly infeasible in terms of cost and time.

We finally used a multi-stage active sampling algorithm [90], which reduces the number of comparisons by 10 fold while maintaining the same level of reliability. In each phase, the algorithm optimizes the information obtained to adapt the algorithms to assign an absolute score. Figure 11 shows the validation of this approach.

III.3 Partial Least Square (PLS) for Variable Selection: Additional Discussion

PLS is a relatively new method and is gaining interests in choosing variables in many scientific domains such as bio-informatics, chemistry and neuroscience [73]. It is becoming widely used mainly due to scientists' ability to collect large datasets with many variables and PLS's ability to handle a large number of variables. The process of choosing the most useful variable is called variable selection, where variables in our case are the objective image metrics (predictors) and the response is the collected VC scores via the crowdsourcing experiment. Unlike principal component regression, which maximizes variance in predictors, PLS explicitly maximizes the covariance between predictors and the target variable (VC score), allowing us to extract latent components that are optimally aligned with perceptual responses. This made PLS an appropriate tool to retain and analyze all metrics jointly while still yielding interpretable component structures.

Our calculation. We assessed the appropriate number of PLS components to retain by running models with 1 to 10 components and selecting the number where the regression R^2 began to plateau. Based on this procedure across different experiments, we selected 5 components for all final analyses to ensure a balance between model complexity and interpretability. For the bootstrapping procedure to assess the reliability of each metric's contribution, we resampled the data with replacement across multiple iterations to estimate the distribution of the PLS regression coefficients, from which we computed standard errors and approximate p-values. This enabled us to identify statistically significant and stable metric variables. Lastly, to aid interpretation, we complemented coefficient analysis with individual metric vs. VC plots (e. g., Figure 7 and extended in the supplemental) that visualize how VC scores vary across the value range of each feature.

III.4 Verbal Reasoning

While the main text focuses on quantitative analysis, we also conducted an additional analysis of the 698 self-reported verbal reasoning responses collected from the post-questionnaire. In these responses, participants explained their rationale for two randomly selected trials. The words and their frequencies offer insights into the cognitive processes involved, helping to construct vocabularies that represent semantic categories relevant to the perception of visual complexity.

This approach aligns with prior work in high-order perception research, such as studies on memorability [72] and attitude [4]. Table 6 shows the proportion of words grouped by semantic category, ordered by descending word frequency. Table 7 summarizes the resulting comments into the categories. The most frequently cited themes were clarity and readability (47.3%) and color and contrast (39.7%), followed by information density (34.1%), number of elements (33.0%), abstractness and familiarity (25.9%), visual clutter (16.8%), interconnectedness (7.4%), and beautifulness (2.6%).

The quantitative results also align with our metric-based findings, supporting the existence of a hierarchy of perceptual features that influence visual complexity. Interconnectedness (7.4%) reflects participants' attention to the relationships between visualization elements, with frequently used terms such as "connection" (9 mentions) and "interact" (6 mentions), primarily associated with node-link diagrams.

The high-level attributes of clarity and readability associated with verbal reasoning may explain that some high-level details could not be explained by the PLS models. For communication purposes, the familiarity, structure, surprise [124], distinctiveness [74, 75], and aesthetic appeal, are important. Furthermore, we also did not measure *correctness*, (i. e., the representations to use visualization that represent key knowledge in order to access meaning) and *semantic distance* (the goodness-of-fit between representation and its meaning). Our image-based metrics do not represent this set of high-level attributes, but our method can capture these to further examine complexity.

IV Additional Implications on Using Complexity to Support Design

Ideally, this project outcome would support new design. *Declutter* removes any additional ink not needed: grid lines, labels, colors, and 3D effects are taken as redundant visual elements [2]. The color removal in their decluttering experiment was appropriate for conditions where colors were not used to encode data. Subsequently, the authors added highlight colors and text annotations to convey key takeaways from the figures. In doing so, the purpose of the images shifted from classic rhetorical visualizations to communicative tools, integrating textual explanations with semiotic marks. The authors observed a weak improvement in perceived VC, a finding that is consistent with the insights from our inquiry. On the other hand, human perception of "redundancy" depends on the expertise of the viewer: People with less experience would benefit from more data-relevant elements.

Our regression model revealed that metrics associated with edge density and feature points, along with color count, play a dominant role in determining perceived complexity, which is in close alignment with the work of Oliva et al. [99]. For visualization applications, extensive studies of visual abstraction and suggestive contours, which are driven by perceptual principles [30, 137], could potentially help reduce VC.

Table 3: A comprehensive literature review of terms, factors, their definitions, and real-world applications reported to influence visual complexity (see the main text section 2). This review served as the foundation for summarizing subjective factors in Table 1 and guiding the selection of objective metric measurements in Table 2.

Category	Factor	Definition	Source	Goal
Information-theoretic	Entropy	The amount of information required to encode the distribution of pixel values.	[121, 123]	To measure the information content and "visual richness" of an image.
	Compression ratio	The size of the compressed image divided by the size of the original raw image.	[26, 123]	To measure image complexity.
	Noise	The unwanted variations and features present in the image.	[77]	To propose a 3D feature space to represent visual complexity of images.
	Quantity of information	The number of units of information on the screen.	[93]	To measure the facets of visual complexity for GUIs.
	Meaningfulness	The meaningfulness of the stimuli.	[99]	To judge visual complexity of images.
	Understandability	The difficulty of semantically understanding a scene. It relates to two aspects: quantity of elements and relationships between elements.	[27]	To evaluate real world image complexity.
	Prototypicality	Compliance with past experience or habit.	[92]	To quantify interface visual complexity.
	Cognitive load	The cognitive effort required for interaction with the interface.	[43]	Definition of visual complexity as an implicit measure of cognitive load.
	Perceivability of detail	Limitations of human perception requiring more effort to perceive details make interfaces seem more complex.	[93]	To measure the facets of visual complexity for GUIs.
Clutter	Clutter	The number and complexity of items or their representation or organization.	[119]	To formalize the concept of visual clutter.
		The amount of distractors and focal points competing for attention.	[99]	To evaluate clutter measurement methods of perceived scene complexity.
	Feature congestion	The energy or strength of edges, color/luminance variations, and orientations within a local neighborhood.	[123]	To calculate and measure visual clutter and complexity of local area.
	Visual clutter	Influences of search performance of target.	[92]	To quantify interface visual complexity.
Color	Colorfullness	The variety and diversity of colors present in the region.	[119]	To evaluate visual clutter.
	Color variability	The amount of color variation across a region quantified by summing the euclidean distances between the color of each pixel and the mean color of the region.	[119]	To measure attributes for visual clutter.
		The number of dominant colors, perceived color depth and idiosyncratic color preferences.	[92]	To quantify interfaces.
	Colors	Number of unique RGB colors.	[107]	To measurement attribute.
	Color regions	The numbers of regions of different color/texture that can be segmented.	[27]	To indicate varied colors/textures and visual complexity.
	Color harmony	The coordination of colors, luminance, and hues in the image.	[27]	To evaluate the color balance of an image and reflect visual complexity.
	Color coherence	The level of continuous and homogeneous intensity of color signal.	[59]	To measure the chromatic contributions to visual complexity and model system representing human visual system's assessment or evaluation of visual complexity.
	Luminance variations	The amount of local luminance variation across a region quantified by summing the luminance contrast between each pixel and its neighbors.	[119]	To evaluate visual clutter.
	Gradient strength	The average magnitude of graylevel gradients, indicating the overall level of contrast and transitions in the image.	[87]	To derive a human criterion-free metric for image complexity.
	Saturation complexity	The vividness of a color.	[38]	To analyze the effectiveness and informativeness of the HSV channels in predicting visual complexity.
	Hue complexity	The color type.	[38]	
	Value complexity	The brightness of a color.	[38]	
	Lighting complexity	The complexity of lighting including the contrast of light and shadow and richness of shadow effects.	[110]	To measure the perceived visual complexity in a computer graphics scene.
Shape and Structure	Symmetry	The degree to which one half of the image predicts the other half.	[99]	
		The similarity of an object reflection.	[92]	To quantify interface visual complexity.
	Organization	The spatial structure, regularity, and placement of items.	[99]	To reflect how the combination of elements affects visual complexity.
	Spatial Frequency	The rate of change or repetition of image patterns across space.	[27]	To indicate the number of cycles or repetitions per unit visual angle or image width.
	Variety of visual form	The number of visual features like colors, shapes, sizes, textures used to represent information.	[93]	GUIs.
	Spatial organization	Repetition and regular positioning of visual elements simplifies the perceived information.	[93]	GUIs.
	Ease of grouping	The similarity within a visual group, and the difference between one visual group and other groups.	[92]	To quantify interface visual complexity.
	Grid	Regular repetition of similar structural elements.	[92]	To quantify interface visual complexity.
	Edges	The contours and boundaries in an image that exhibit significant graylevel changes.	[87]	Photos & images.
	Edge density	The ratio between the number of edge pixels and the total number of pixels in an image.	[87]	To measure image complexity.
		The area of the image occupied by edges.	[107]	To measure visual complexity.
	Edge congestion	The density of edges.	[92]	To quantify interface visual complexity.
	Structure	The visual patterns and edges present in the image. Images with more edges and patterns are considered to have higher structure.	[77]	To propose A 3D feature space to represent visual complexity of images.
	Compactness	The ratio of squared perimeter to area.	[5]	To judge complexity of shapes.
	Turns	The number of sides or turns in the shape.	[5]	To judge complexity of shapes.
	Skeletons	The internal structure of the shape.	[132]	To measure shape complexity by explaining why a shape has the external features.
	Texture variation	A higher contrast, correlation, energy and lower homogeneity indicate more texture variation and complexity.	[27]	To predict visual complexity across different image categories.
	Material complexity	The richness, variety and intricacy of materials and textures.	[110]	To measure the perceived visual complexity in a computer graphics scene.
	Co-occurrence matrix	symbiotic relationship between different texture.	[59]	To evaluate visual complexity.
Object-based	Number of objects	The number of objects or components that make up a visual scene.	[99]	To evaluates clutter measurement methods of perceived scene complexity.
	Object categories	The number of unique object categories in a scene.	[16]	To account for the effect of perceptual grouping. This means that similar objects can be perceived as a group, reducing the perceived complexity.
	Element relationships	More complex interactivity and relationships between elements in a scene.	[51]	To evaluate scene complexity.
	Element heterogeneity	Diversity and heterogeneity of elements.	[128]	To evaluate visual complexity of image.
	Element diversity	To represent variety of visual elements present in the image including objects, textures, and corners, etc.	[77]	To propose a 3D feature space based on structure, noise, and diversity.

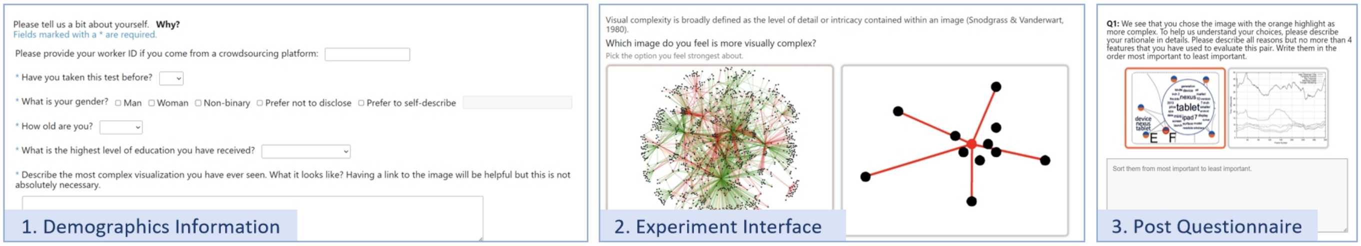
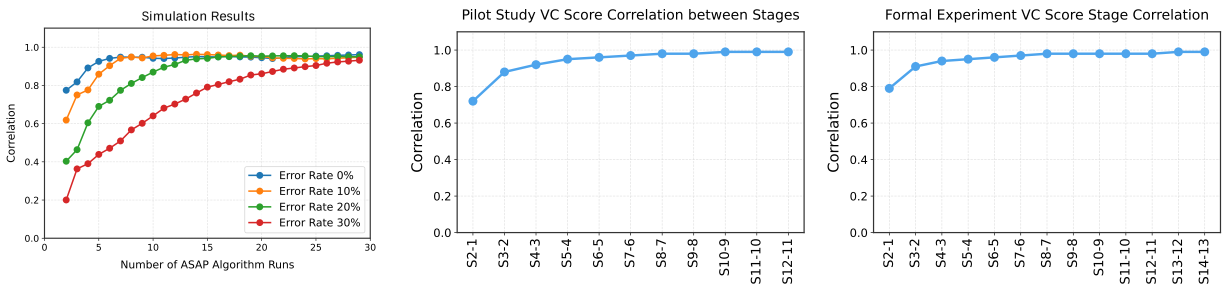


Fig. 10: **Interface for collecting perceived visual complexity.** Participants perform multiple paired comparisons between images, with new pairs assigned using an active sampling method (see the main text subsection 3.3). The collected data were used to assess observer consistency and to compute perceived visual complexity for our quantitative metric analysis.



(a) Simulation of noise input.

(b) Pilot study reliability.

(c) Formal study reliability.

Fig. 11: **Reliability of the Mikhailiuk et al.'s multi-stage active sampling algorithm [90]**. This figure validates the active sampling algorithm and estimates the number of stages required for our data collection (see the main text subsection 3.2). **(a). Robustness to human input noise.** We simulated user error by randomly flipping 0%, 10%, 20%, and 30% of comparison outcomes, leading the algorithm to recommend a different set of images in the next phase. Despite the noise, the algorithm remained robust—at noise levels $\leq$ 20%, the correlation between noisy and original image scores remained above 0.9 and stabilized within 13 stages. **(b). Pilot test of algorithmic reliability (N=90 images).** In a 12-stage pilot, 72 different participants (6 per stage) ranked 90 images. The plot shows high correlations between image scores across adjacent stages, indicating convergence. **(c). Reliability in the formal full study (N=1,800 images).** Our stopping rule required a correlation > 0.95 across three consecutive stages. After the 13^{th} stage, the VC score correlation with the previous stage reached 0.98. Our data collection stopped at the 14^{th} stage, and results from this stage were used for analysis. **Observations.** The results indicated that the scores stabilized by the 10th stage after approximately 890 comparisons. Comparisons generated by the active sampling algorithm were randomly distributed across 6 participants per stage.

Fig. 12: **A 300 Example image set with visual complexity scores. (a). From 1,800 images,** *Top:* The least visually complex visualization images from our experiment. *Middle:* The images around the median visual complexity scores. *Bottom:* The most visually complex images from our experiment.

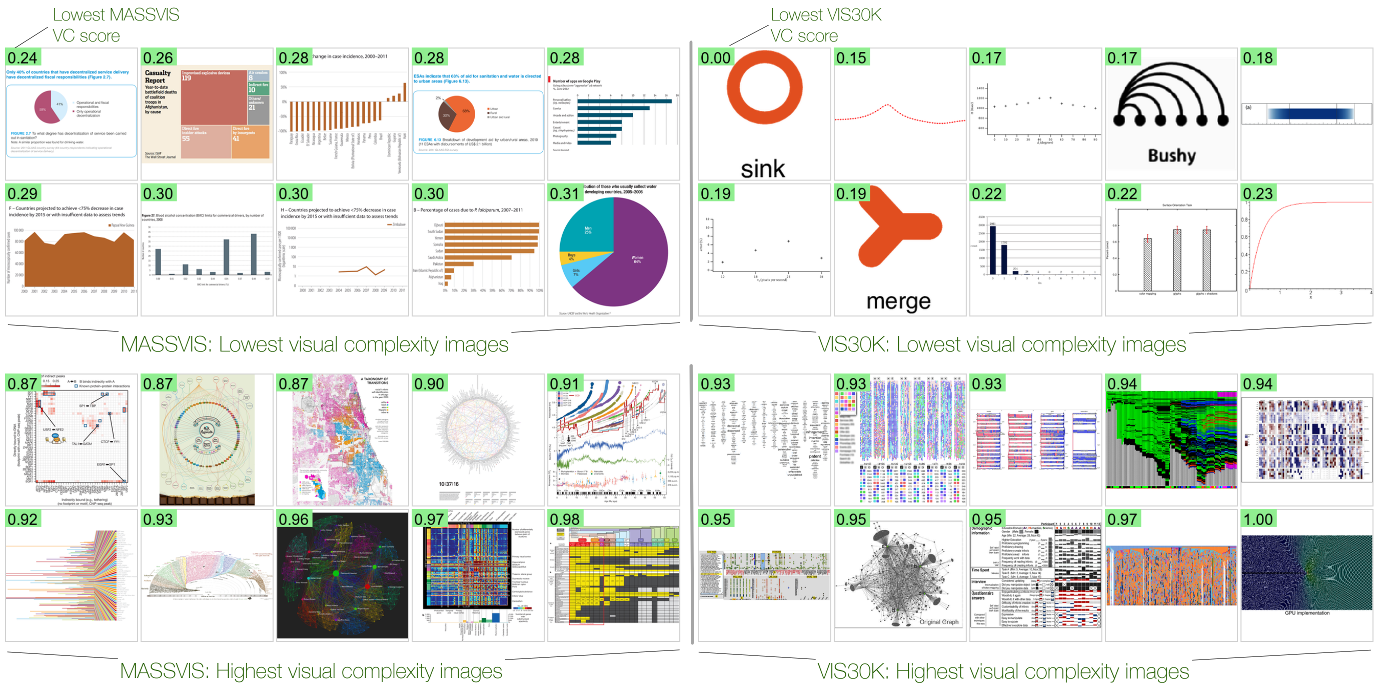


Fig. 13: The lowest and highest visually complex images in MASSVIS (left) [10] and VIS30K (right) [19] (see the main text subsubsection 4.3.2).

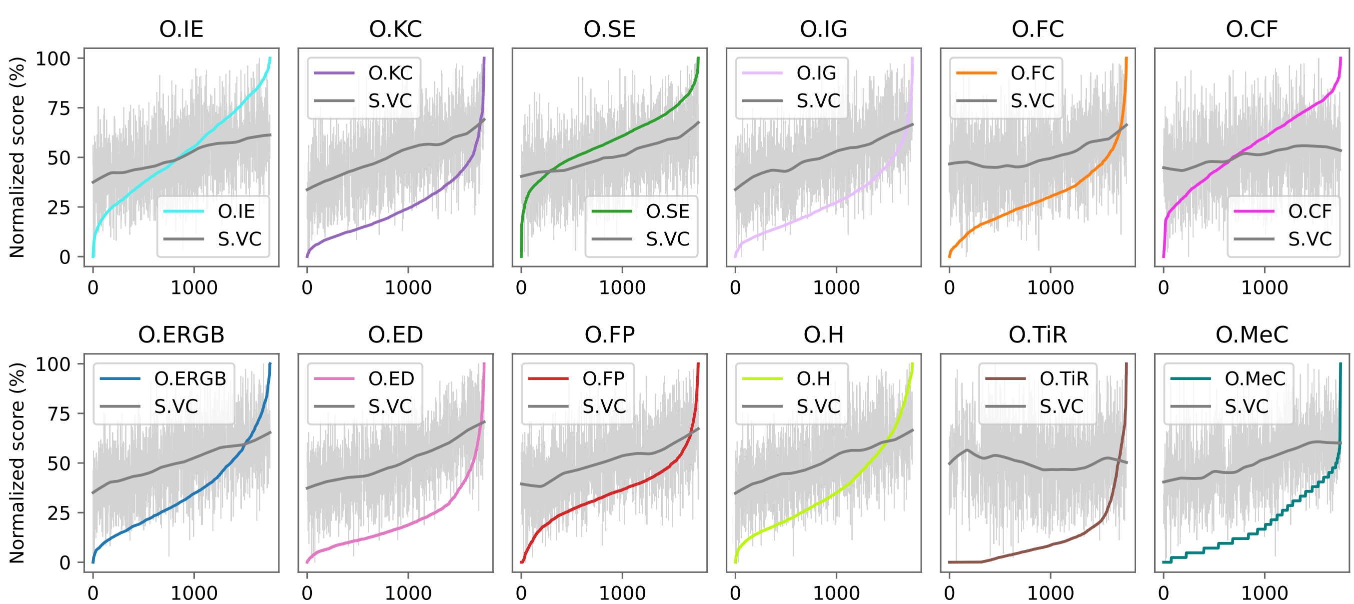


Fig. 14: **Metrics vs. Perceived VC.** The colored lines show the metric values in ascending order for the 1800 images, while the light-gray lines in the background are perceived VC scores, overlaid by a locally weighted smoothing line in gray.

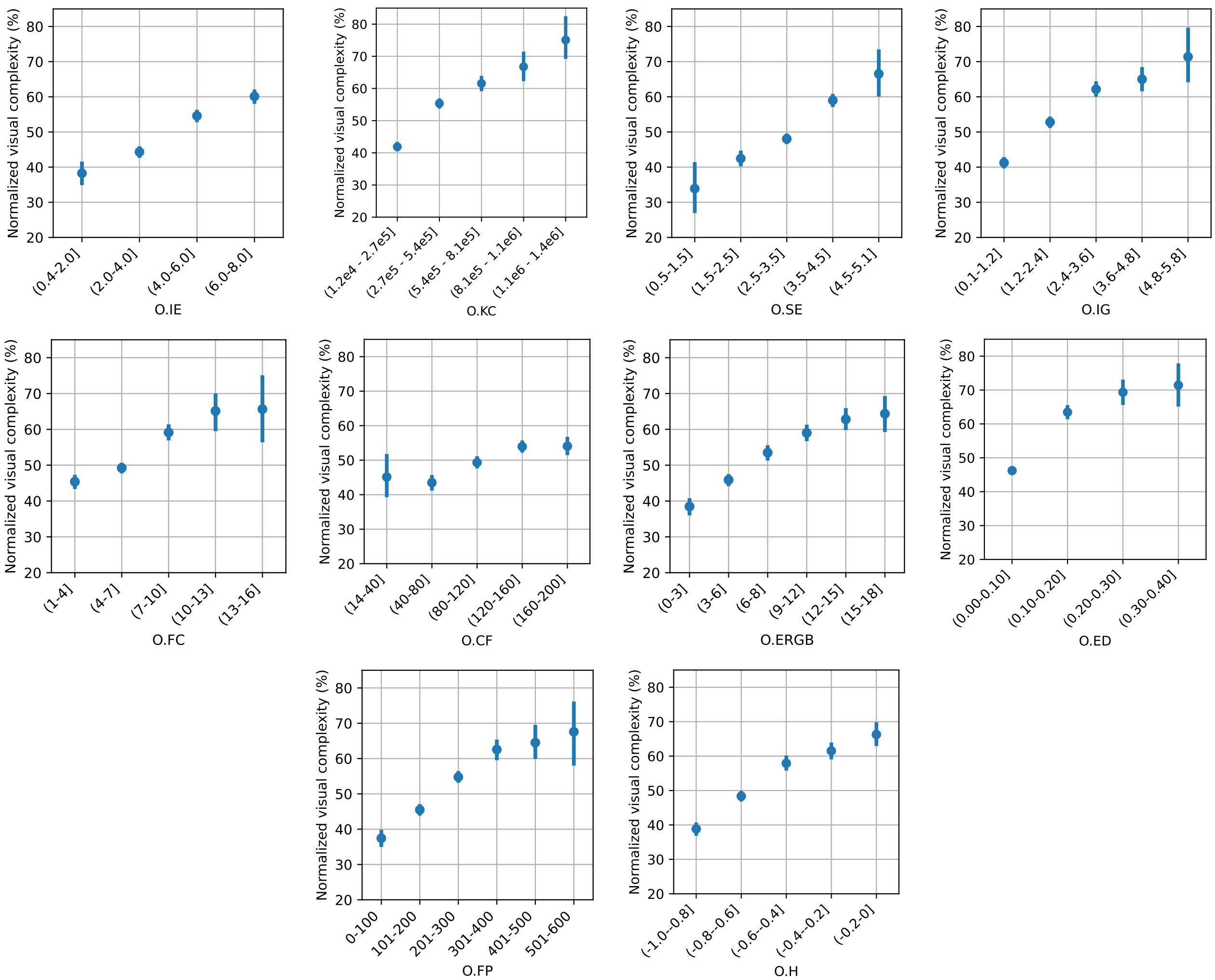


Fig. 15: **Mean Visual Complexity (VC) across for visualization of all 10 objective metrics.** The O.TiR and O.MeC metric correlation results are in the main text Figure 7. Error bars represent 95% confidence intervals.

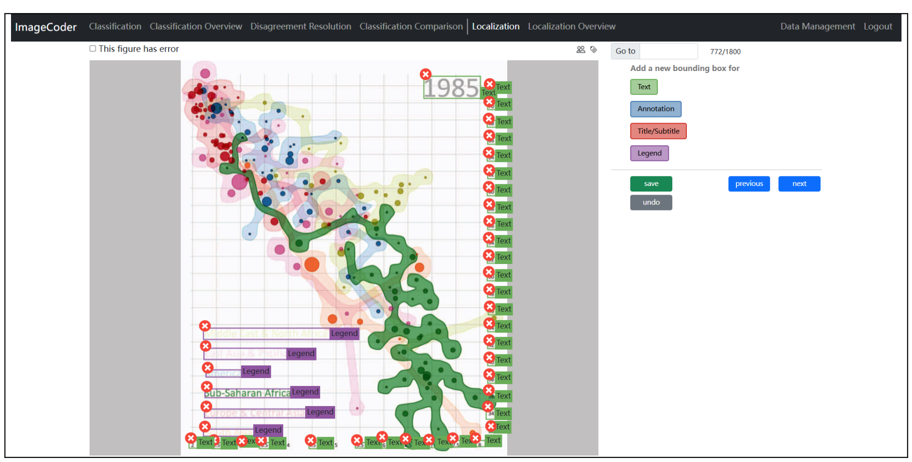
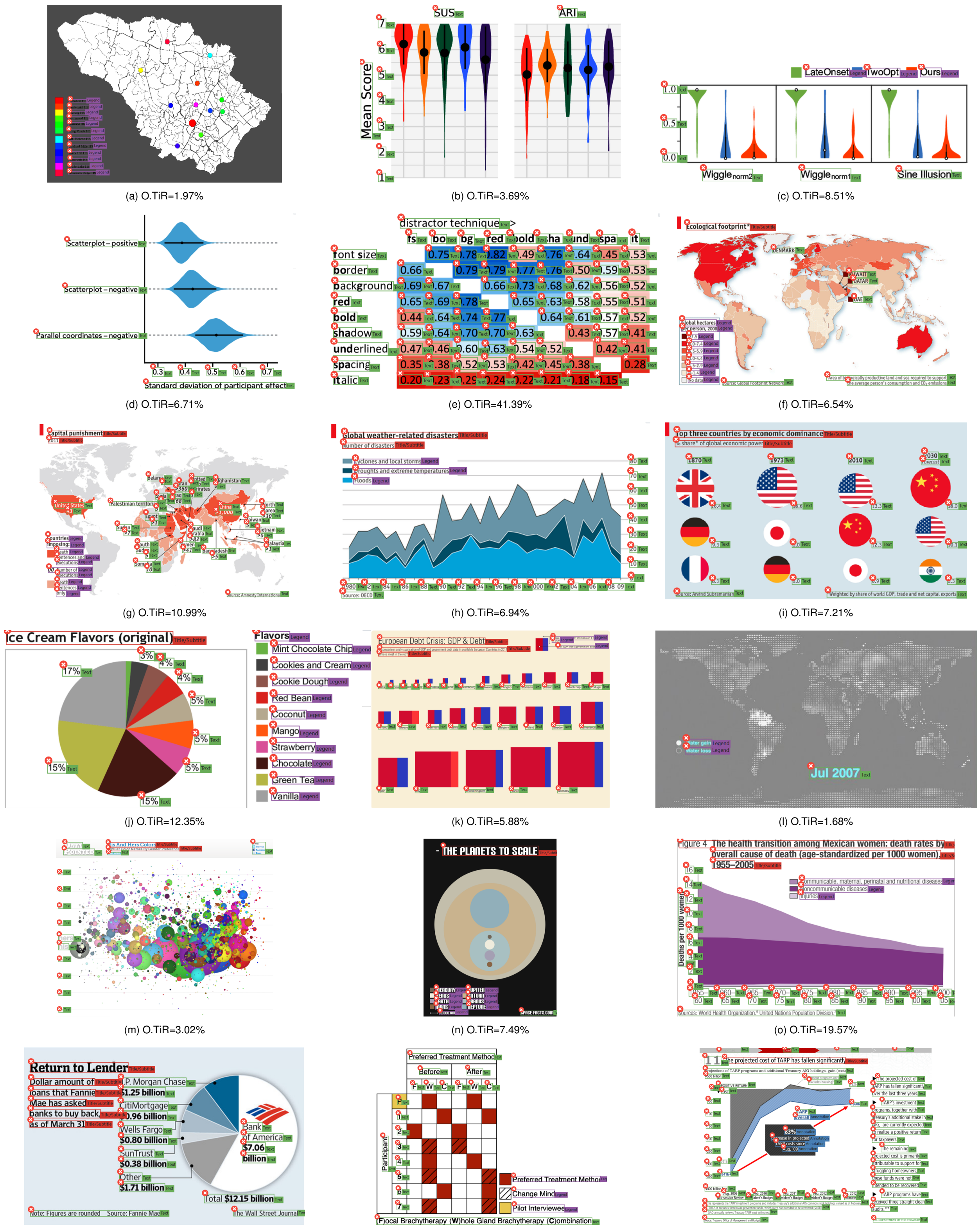


Fig. 16: **Interface for curating the text labels.** Here, our labels clearly separated ‘annotation’ from other texts such as ‘Title/subtitle’, and ‘legend’ (see main text subsection 4.1).



(a) O.TiR=1.97% (b) O.TiR=3.69% (c) O.TiR=8.51%

(d) O.TiR=6.71% (e) O.TiR=41.39% (f) O.TiR=6.54%

(g) O.TiR=10.99% (h) O.TiR=6.94% (i) O.TiR=7.21%

(j) O.TiR=12.35% (k) O.TiR=5.88% (l) O.TiR=1.68%

(m) O.TiR=3.02% (n) O.TiR=7.49% (o) O.TiR=19.57%

(p) O.TiR=19.52% (q) O.TiR=10.41% (r) O.TiR=23.07%

Fig. 17: **O.TiR examples.** Each subcaption shows its O.TiR value. Additional results for computing O.TiR for all images in VisComplexity2K are available in the corresponding Google Sheet `go.osu.edu/viscomplexitydata`, under the tab `O.TiRResults`. Also see main text subsection 4.1.

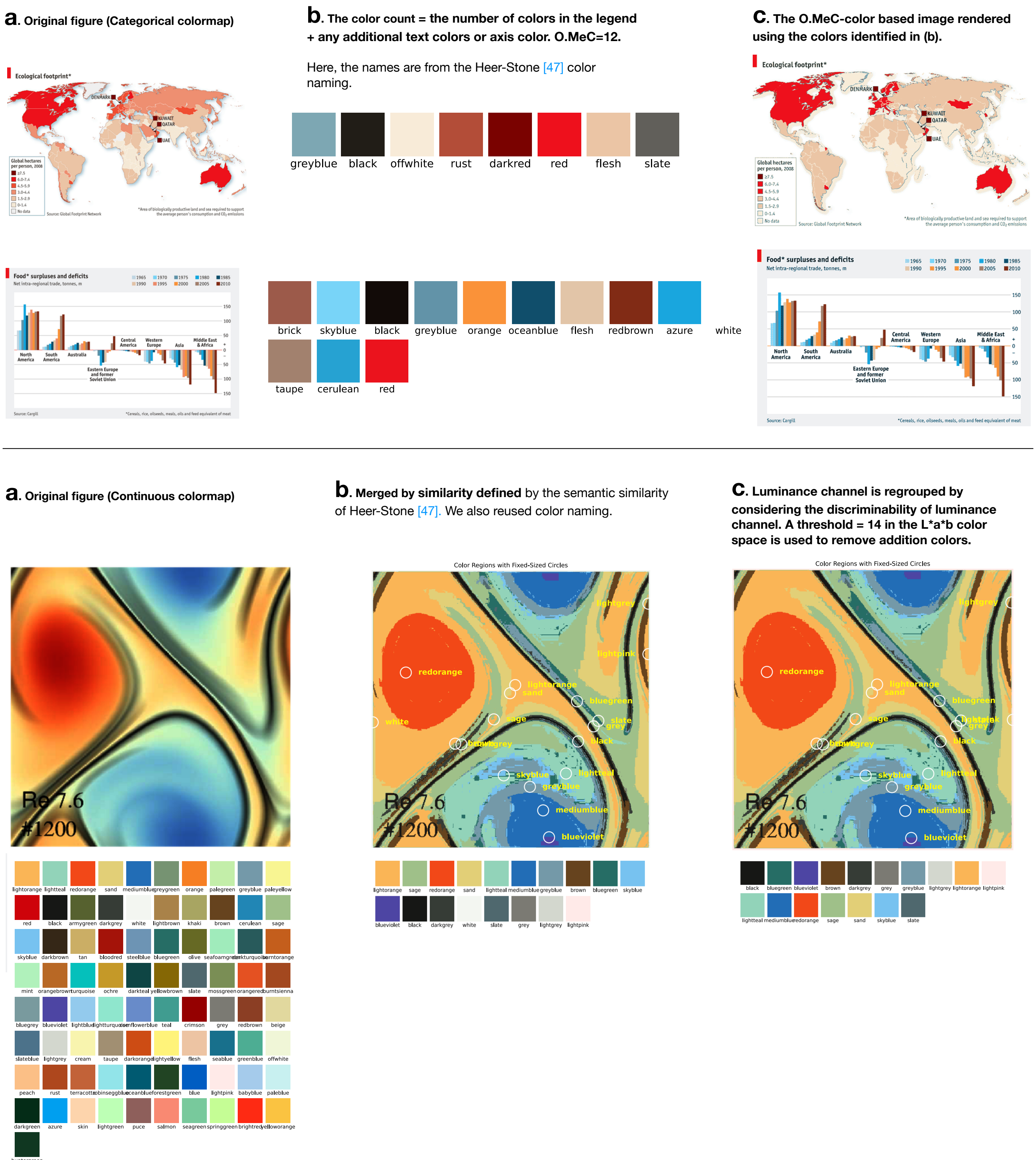
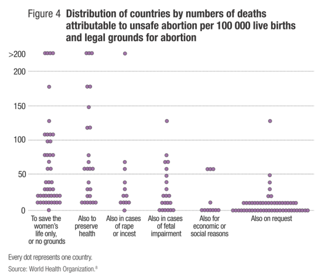
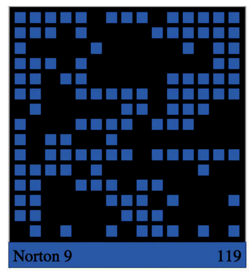


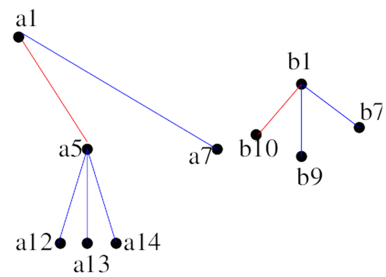
Fig. 18: **The O.MeC metric computational workflow.** *Top*: categorical color map processing. *Bottom*: continuous color map processing. Step-by-step results for computing O.MeC on both continuous and discrete color images in VisComplexity2K are available in the corresponding Google Sheet `go.osu.edu/viscomplexitydata`, under the tab `O.MeCResults`. Also see main text subsection 4.1.



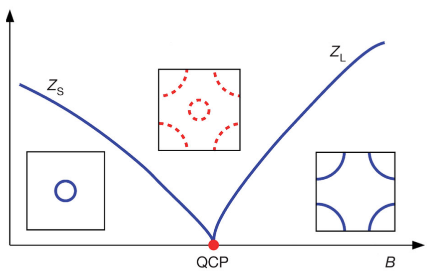
(a) O.MeC=2



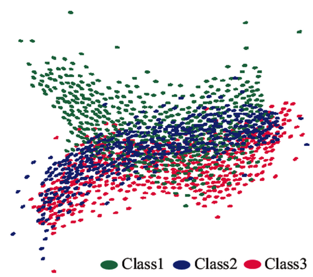
(b) O.MeC=2



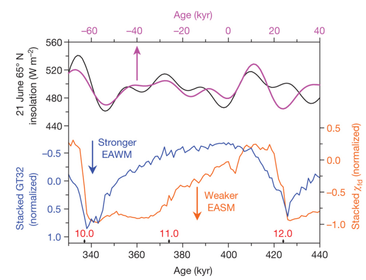
(c) O.MeC=3



(d) O.MeC=3

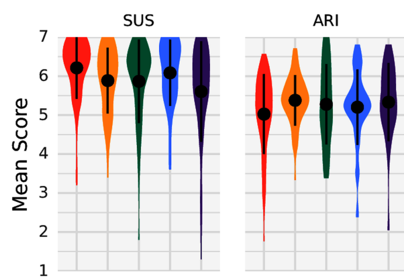


(e) O.MEC=4

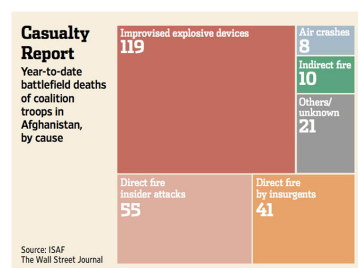


(f) O.MeC=5

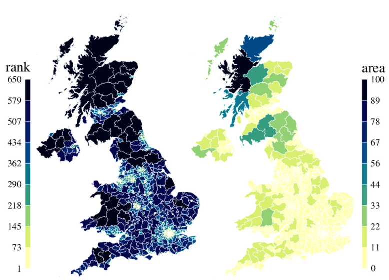
(g) O.MeC=6



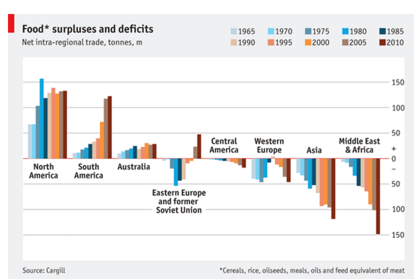
(h) O.MeC=7



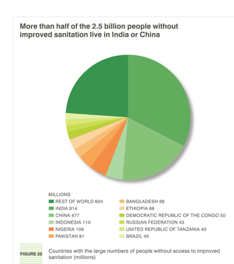
(i) O.MeC=9



(j) O.MeC=10

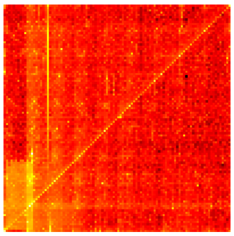


(k) O.MeC=13

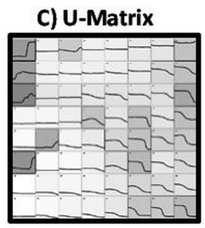


(l) O.MeC=14

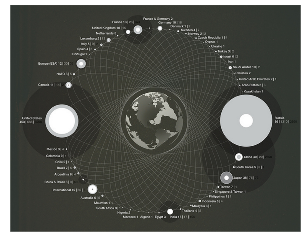
Fig. 19: More O.MeC examples for **discrete** representations. Subcaptions contain the O.MeC number (see main text subsection 4.1).



(a) O.MeC=4

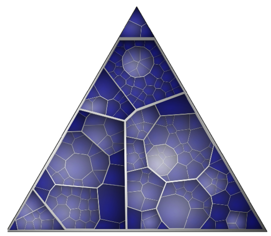


(b) O.MeC=5

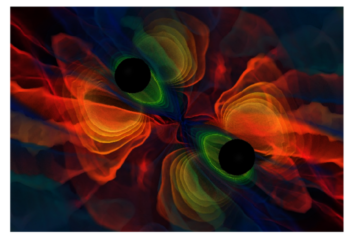


(c) O.MeC=7

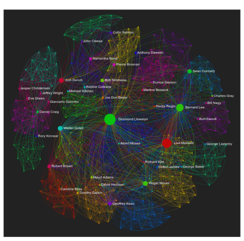
(d) O.MeC=8



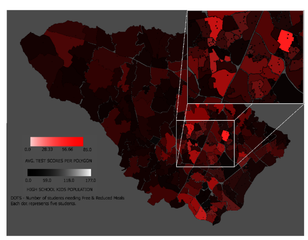
(e) O.MeC=11



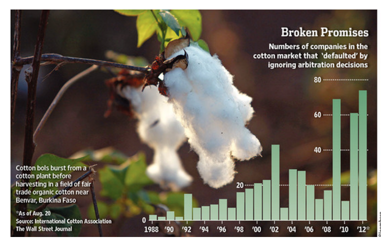
(f) O.MeC=14



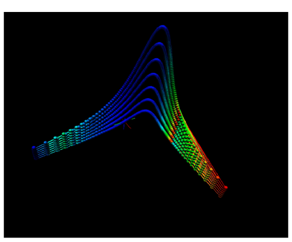
(g) O.MeC=15



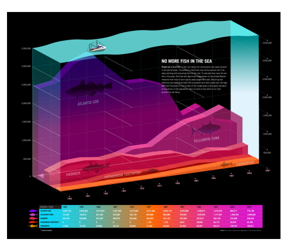
(h) O.MeC=15



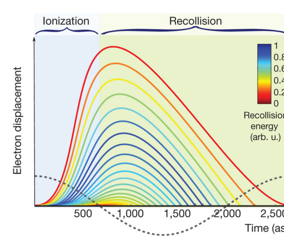
(i) O.MeC=16



(j) O.MeC=18



(k) O.MeC=22



(l) O.MeC=24

Fig. 20: More O.MeC examples for **continuous** representations. Subcaptions contain the O.MeC number (see main text subsection 4.1).

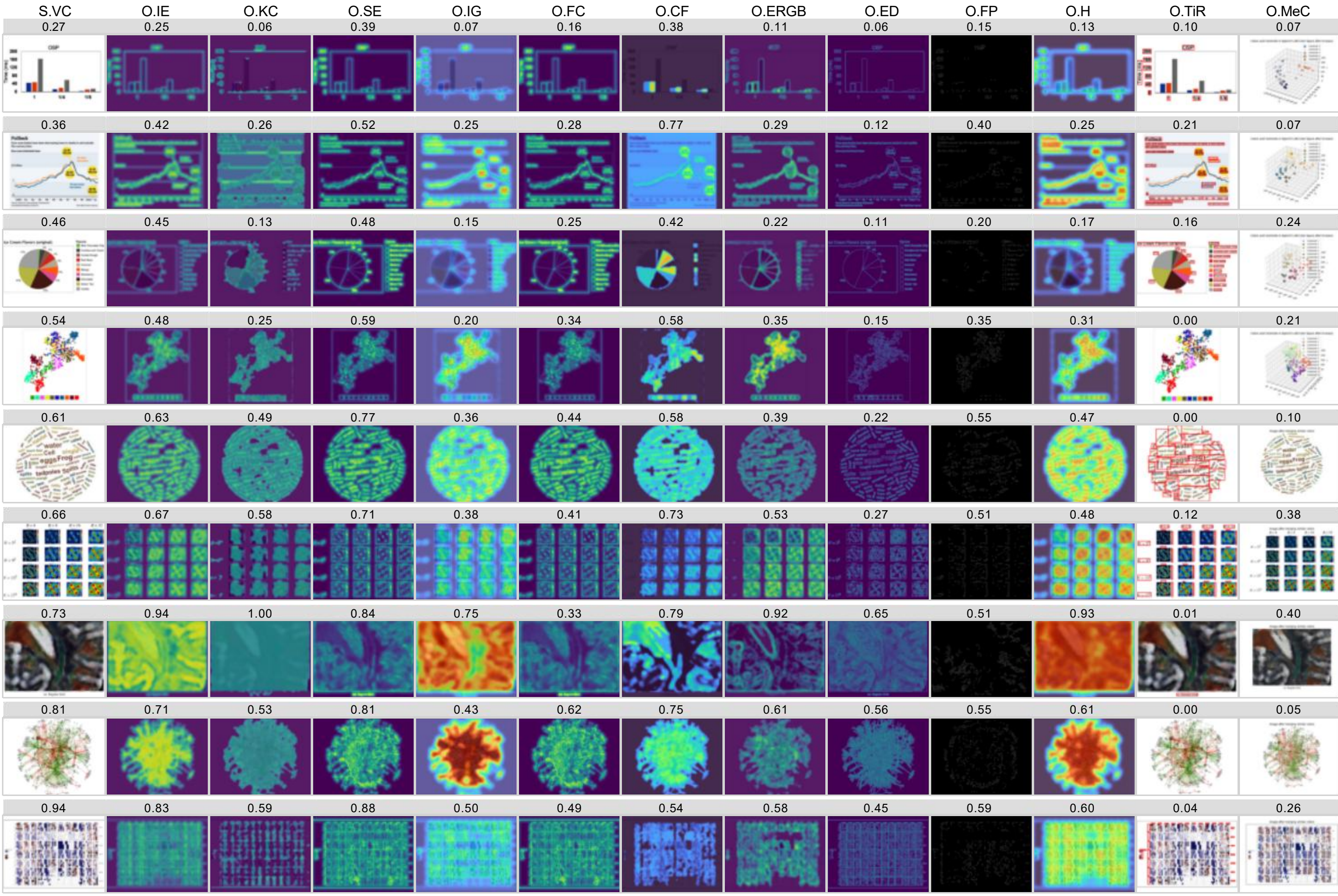


Fig. 21: **Additional examples of image VC and metric scores.** Each row shows the original image followed by visual representations of the 12 objective metrics, along with their corresponding scores. Values indicate normalized percentage scores (0-100%).

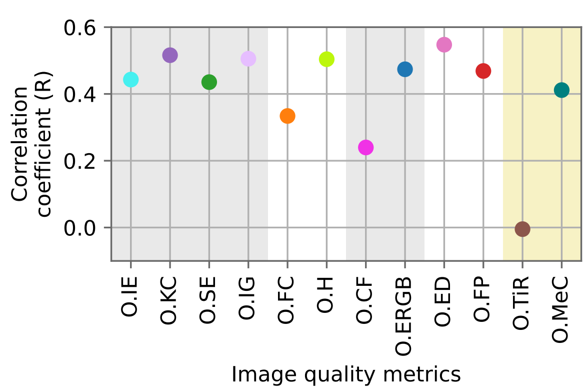


Fig. 22: **Single variable analysis: correlation between each metric and perceived VC.** Pearson correlation coefficients (r) were used to assess the linear relationships between each objective metric and perceived VC. The associated effect size is typically categorized as small ($r \geq 0.10$), medium ($r \geq 0.30$), and large ($r \geq 0.50$). **Observation.** All but TiR metrics were positively correlated, with most showing moderate to strong correlations. Between the two object-based metrics, O.MeC demonstrated a moderate correlation with perceived VC ($r = 0.41$), whereas O.TiR showed a negative but statistically insignificant correlation with perceived VC ($r = -0.01, p = 0.56$).

Table 4: **Metric correlation analyses.** P-values for between-metric Pearson correlations used in the main text subsection 4.2 and Apdx. Figure 23. O.TiR has large p-values with most other metrics, indicating non-significant correlations. O.MeC exhibits non-significant correlations with O.TiR and O.FC. In all other cases, $p \leq 0.05$ indicates significant correlations.

	O.IE	O.KC	O.SE	O.IG	O.FC	O.H	O.CF	O.ERGB	O.ED	O.FP	O.TiR	O.MeC
O.IE												
O.KC	<0.001											
O.SE	<0.001	<0.001										
O.IG	<0.001	<0.001	<0.001									
O.FC	<0.001	<0.001	<0.001	<0.001								
O.H	<0.001	<0.001	<0.001	<0.001	<0.001							
O.CF	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001						
O.ERGB	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001					
O.ED	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001				
O.FP	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001			
O.TiR	0.28	0.79	<0.001	0.225	<0.001	0.414	0.088	0.028	0.052	0.12		
O.MeC	<0.001	<0.001	<0.001	<0.001	0.04	<0.001	<0.001	<0.001	<0.001	<0.001	0.157	

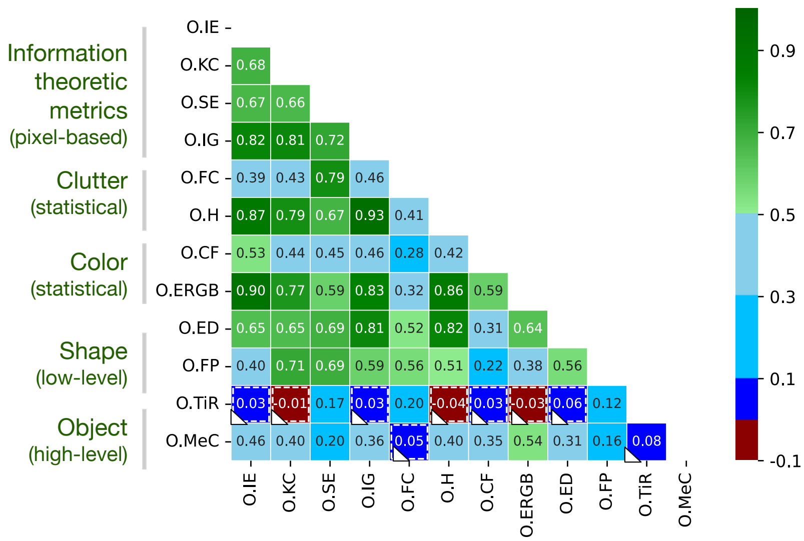
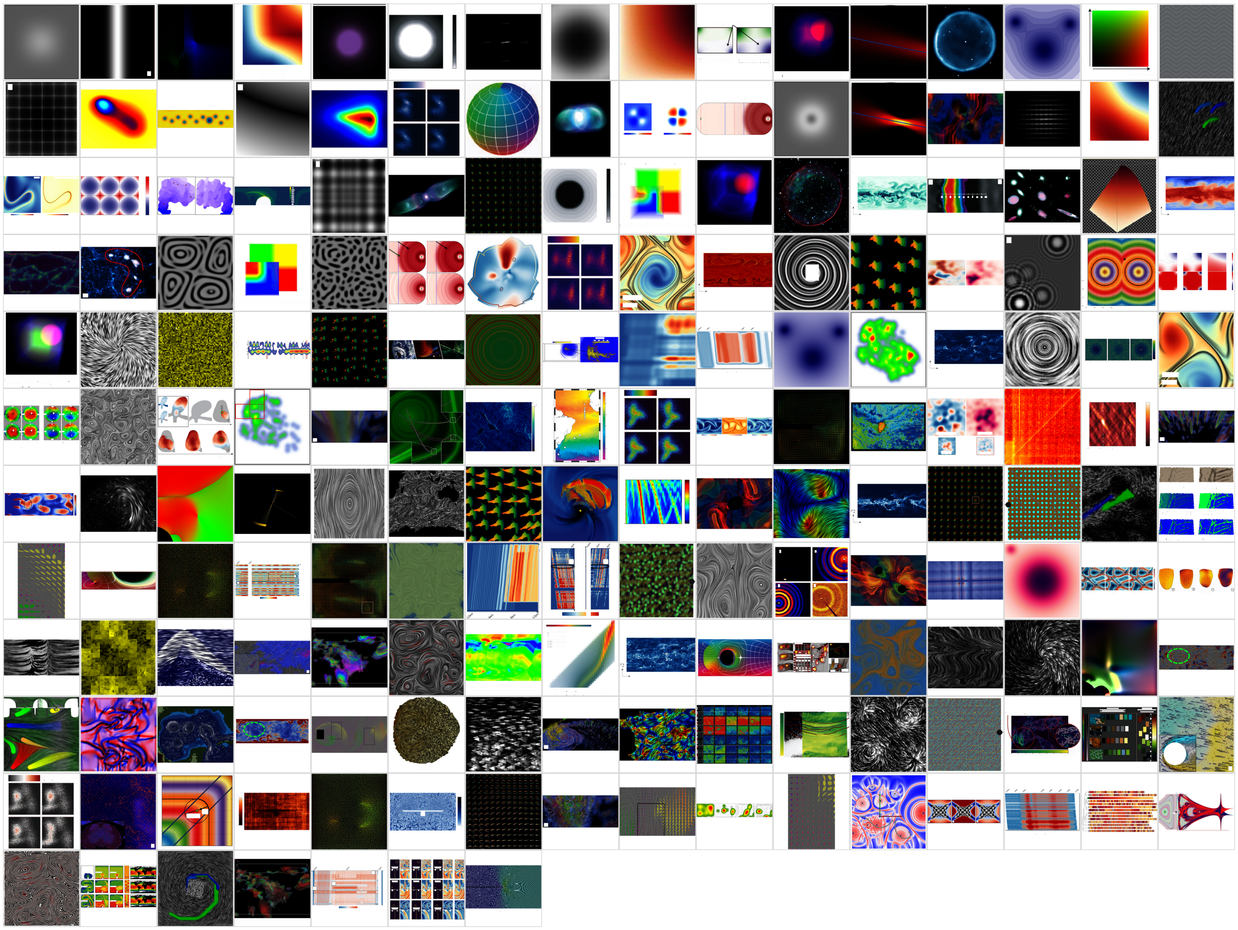


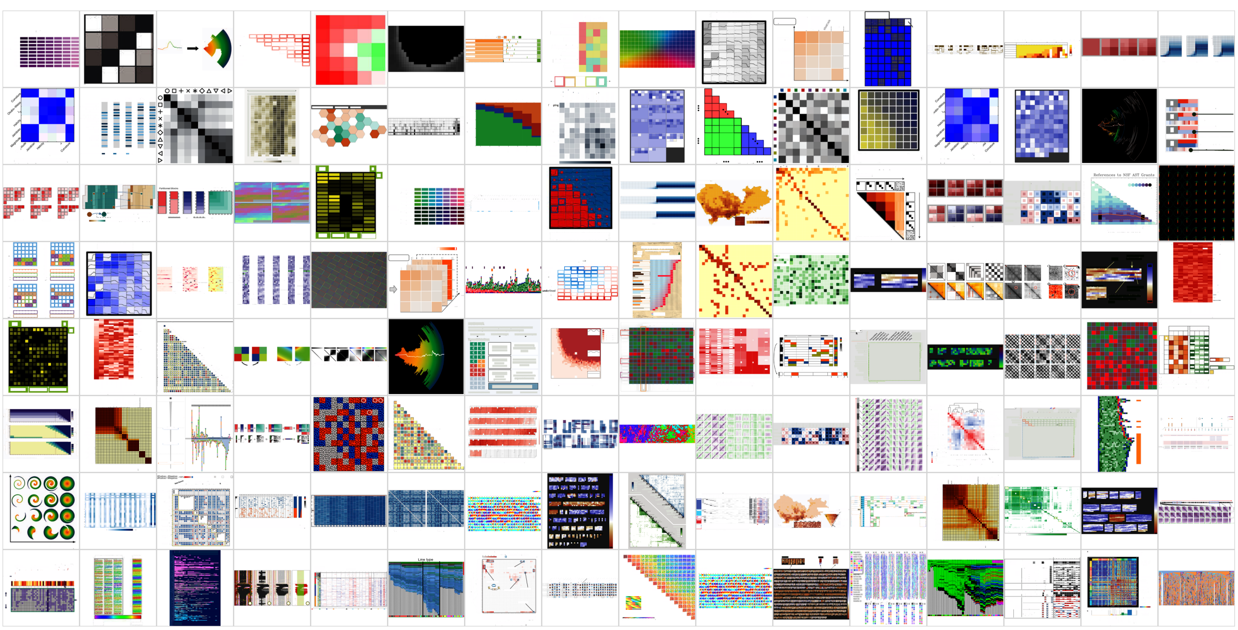
Fig. 23: **Metric correlation analyses.** Pearson correlation coefficients (r) used in the main text subsection 4.2 between the image quality metrics (Table 2), computed across all images. Cells with a white triangle indicate non-significant correlations at the 0.01 level; all other correlations are statistically significant with $p < 0.001$ (see Table 4). **Observations. (1) Strong within-factor correlations** were observed among the four mathematical *information-theoretic metrics* (O.IE, O.KC, O.SE, O.IG) and between the two shape-based metrics (O.ED, O.FP). In the case of information-theoretic metrics, complexity is mathematically defined by the length of the system description needed to encode an image. When the same input is processed through different computational models, the resulting compressions tend to be similar. Similarly, for shape-based metrics, the relationship is intuitive: more edges (O.ED) inherently lead to more feature points (O.FP). **(2) Metrics O.IG, O.FC and O.H are sensitive to texture and show strong mutual correlations.** (3)**Cross-factor correlations were also present**. Shape-based metrics (O.ED, O.FP), along with O.H and O.ERGB, exhibited strong correlations with information-theoretic metrics (O.IE, O.KC, O.SE, O.IG), with $r \in [0.58, 0.93]$ except for $r = 0.41$ between O.FP and O.IE. These findings suggest that different metric categories may capture overlapping aspects of visual complexity. (4) In contrast, **object-based metric O.TiR showed little to no correlation with other metrics**, suggesting a fundamental distinction between text-based and graphical features. O.MeC also had relatively weak correlations with other metrics, supporting its role in capturing semantic or structural complexity rather than low-level image statistics. Finally, there was no meaningful correlation between O.FC and O.MeC, indicating they capture distinct dimensions of visual information. The p-values for between-metric correlations are reported in Table 4.

Table 5: **Effect sizes from Partial Least Squares (PLS) analysis, measured by f-squared (f^2).** These values indicate the contribution of each metric to the overall explained variance in the perceived VC across four experiments. Bold values represent small ($f^2 \geq 0.02$) or medium ($f^2 \geq 0.15$) effect sizes. A value of $f^2 \geq 0.35$ is considered a large effect size. **Observation.** O.ED emerged as the strongest metric, showing a significant effect in three out of four PLS models. It was followed by O.MeC, O.FP, and O.FC, each significant in two of the four models, and O.IE, which was significant in one model. Notably, O.ED mainly contributed to node-link diagrams; O.FC is the only factor that has a relatively large effect-size to represent the continuous color and texture pattern representations.

Stimulus	O.IE	O.KC	O.SE	O.IG	O.FC	O.H	O.CF	O.ERGB	O.ED	O.FP	O.TiR	O.MeC
Overall (Figure 6)	-0.000	0.000	0.002	0.001	0.003	0.001	0.001	0.003	**0.050**	**0.031**	0.007	**0.068**
HP-Node-link (Figure 8)	0.001	0.006	-0.003	0.001	0.007	-0.001	0.005	**0.026**	**0.288**	0.012	0.001	0.019
Dsct.-Grid & Matrix (Figure 9)	**0.026**	-0.000	0.014	0.003	**0.021**	0.001	-0.001	0.007	**0.059**	**0.058**	0.003	**0.148**
RR-Cont.-Grid & Matrix (ColorPatn) (Figure 9)	0.000	0.009	0.015	0.016	**0.152**	-0.001	0.007	0.011	0.014	0.001	0.001	0.001



(a) Pixel location is given (183 images).



(b) Pixel location is not given (128 images).

Fig. 24: *Top (a):* Images in the style of Rosenholtz et al. [119], where colors and textures dominate the visual representation. *Bottom (b):* Examples represent discrete grid and matrix images (see the main text subsection 5.2).

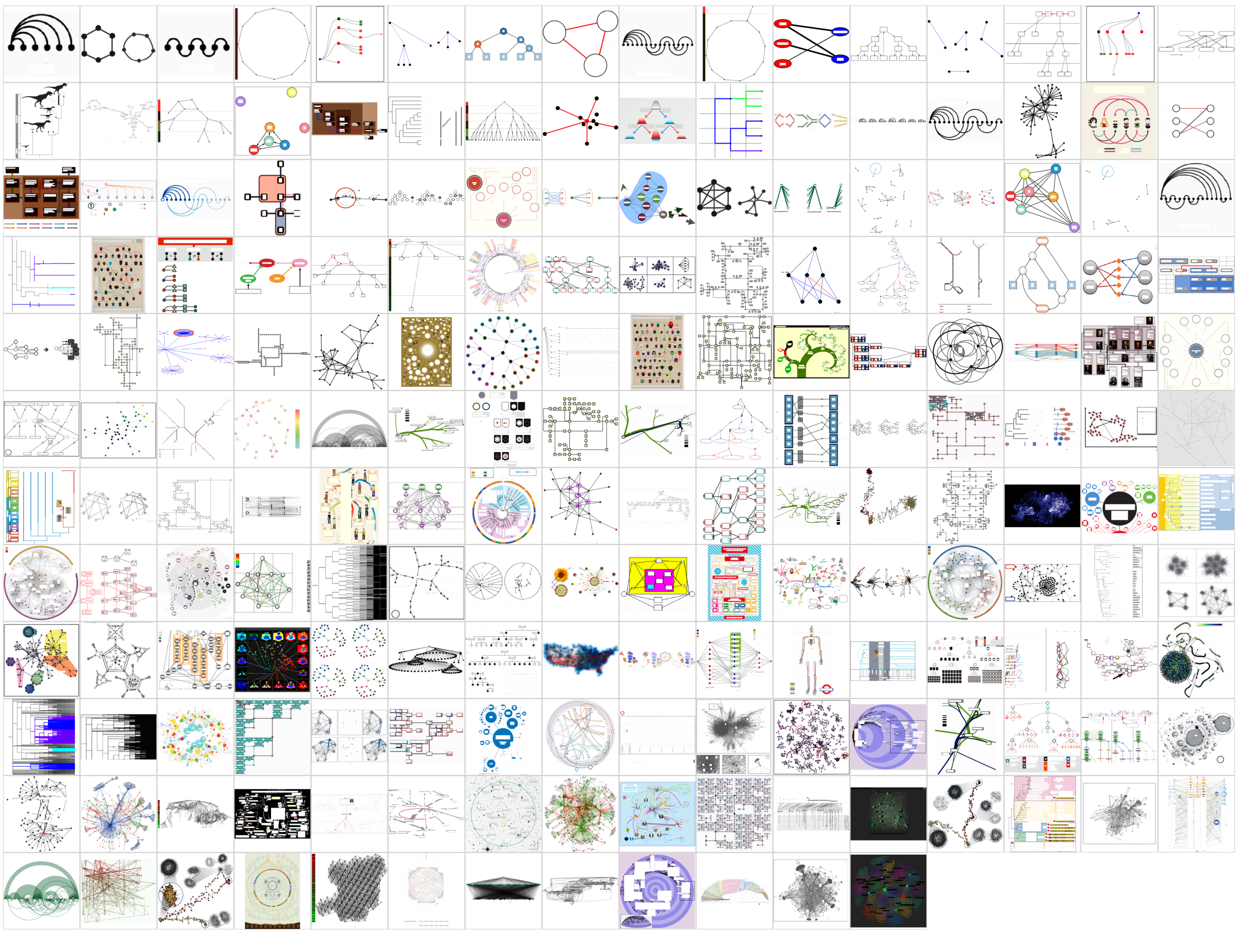
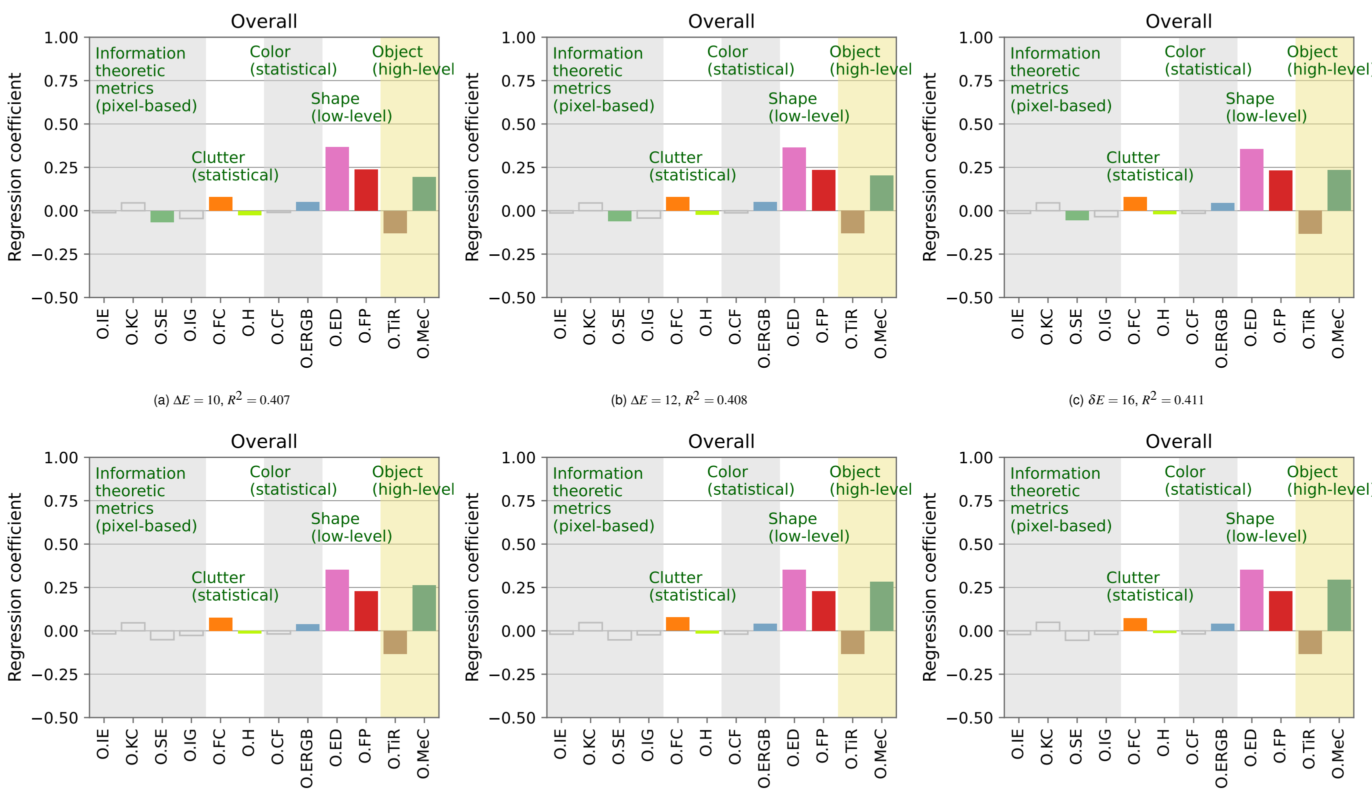


Fig. 25: **Direct node-link diagrams** in the style of Purchase [107] (see the main text subsection 5.1).



(a) $\Delta E = 10$, $R^2 = 0.407$

(b) $\Delta E = 12$, $R^2 = 0.408$

(c) $\delta E = 16$, $R^2 = 0.411$

(d) $\delta E = 18$, $R^2 = 0.415$

(e) $\delta E = 20$, $R^2 = 0.415$

(f) $\delta E = 22$, $R^2 = 0.416$

Fig. 26: **PLS modeling results of VC**, when O.MeC is computed using different threshold values ranging from ΔE = 10, 12, 14, 16, 18, 20, and 22. Each subplot corresponds to a specific threshold ΔE and the corresponding PLS regression coefficient. **Observations.** Despite variations in threshold values, the modeling results remain largely consistent, indicating minimal impact of the thresholding on the overall model structure.

Table 6: Top-10 keywords generated from the post-questionnaire comments. Keywords in italics are unique to their category. The numbers in brackets and the width of the bar denote the number of participant comments mentioning that keyword or topic (see Apdx. subsection III.4).

Clarity and Readability (317)	**Color Usage (271)**	**Number of Elements (226)**	**Information Density (221)**	**Abstractness and Familiarity (173)**	**Visual Clutter (113)**	**Interconnectedness (49)**	**Beautifulness (16)**
(55) clear	(249) color	(40) line	(95) information	(29) complex	(39) detail	(17) line	(3) appealing
(39) easy	(21) *contrast*	(32) number	(30) detail	(19) *familiar*	(23) element	(8) *connection*	(3) *like*
(39) hard	(18) variety	(31) element	(20) legend	(17) pattern	(21) clutter	(7) overlap	(3) pattern
(39) label	(16) different	(30) text	(19) complex	(17) *structure*	(15) *excessive*	(6) *intersect*	(3) *pleasing*
(29) difficult	(10) *shade*	(22) lot	(17) point	(13) *unfamiliar*	(15) overlap	(5) *interconnect*	(2) color
(28) font	(9) lot	(21) detail	(16) lot	(11) difficult	(8) layout	(5) point	(2) *impact*
(27) layout	(7) *dark*	(16) variety	(13) *concentration*	(11) *subject*	(7) complex	(4) *crossing*	
(27) text	(7) *range*	(14) many	(13) layer	(10) graph	(5) dot	(4) element	
(21) information	(7) *saturation*	(13) point	(12) amount	(9) *idea*	(5) information	(4) lot	
(21) small	(6) number	(12) amount	(11) graph	(8) clear	(5) tidy	(3) arrow	

Table 7: **Categorization of subject VC terms** from the user-studies (see Apdx. subsection III.4).

Terms	Reasoning
Clarity and Readability	Poor readability due to small font sizes, cluttered layouts, or lack of clear labels can increase perceived complexity. Clear and concise labeling, along with a logical flow of information, tends to reduce complexity.
Color Usage	The use of a wide range of colors, especially when they are vibrant or poorly contrasting, can add to the complexity. Images with more color variations and saturation are often seen as more visually complex.
Information Density	High information density, where a lot of data is presented in a small space, can make an image appear more complex. This includes the presence of detailed legends, annotations, and multiple layers of information.
Number of Elements	Images with more elements, such as shapes, lines, and text, are often perceived as more complex. The presence of multiple data points, overlapping elements, and intricate patterns can increase the perceived complexity. *Glyph*, *Node-link*, and *Point* were most concerning.
Abstractness and Familiarity	Images that lack a clear structure or recognizable patterns are often perceived as more complex. Familiarity with the content also plays a role; images depicting unfamiliar subjects or using unfamiliar formats are seen as more complex.
Visual Clutter	Overlapping elements, excessive details, and lack of negative space can contribute to a sense of clutter, making an image appear more complex.
Beautifulness	Visual appeal, good-looking, personal preference, emotional feeling.