UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Complexity in Complexity: Understanding Visual Complexity Through Structure, Color, and Surprise

Permalink
https://escholarship.org/uc/item/2g6408z2

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Sarıtaş, Karahan
Dayan, Peter
Shen, Tingke
et al.

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Complexity in Complexity: Understanding Visual Complexity Through Structure, Color, and Surprise

Karahan Sarıtaş[1], **Peter Dayan**[1,2], **Tingke Shen**[2*], **Surabhi S Nath**[1,2,3*]
[1]University of Tübingen, Tübingen, Germany
[2]Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[3]Max Planck School of Cognition, Leipzig, Germany

Abstract

Understanding human perception of visual complexity is crucial in visual cognition. Recently (Shen, et al. 2024) proposed an interpretable segmentation-based model that accurately predicted complexity across various datasets, supporting the idea that complexity can be explained simply. In this work, we investigate the failure of their model to capture structural, color and surprisal contributions to complexity. To this end, we propose Multi-Scale Sobel Gradient which measures spatial intensity variations, Multi-Scale Unique Color which quantifies colorfulness across multiple scales, and surprise scores generated using a Large Language Model. We test our features on existing benchmarks and a novel dataset containing surprising images from Visual Genome. Our experiments demonstrate that modeling complexity accurately is not as simple as previously thought, requiring additional perceptual and semantic factors to address dataset biases. Thus our results offer deeper insights into how humans assess visual complexity.

Keywords: visual complexity; natural images; image segmentation; image gradients; colorfulness; large language models; surprise

Introduction

Visual complexity is a fundamental attribute of images that reflects the level of detail, intricacy, and variation in visual elements within a scene (Snodgrass & Vanderwart, 1980a). This perceptual characteristic encompasses multiple dimensions, including the number of lines, density of elements (McDougall, Curry, & de Bruijn, 1999), quantity of objects (Olivia, Mack, Shrestha, & Peeper, 2004), clutter (Kyle-Davidson, Zhou, Walther, Bors, & Evans, 2023), symmetry (Kyle-Davidson, Solis, Robinson, Tan, & Evans, 2025), and variety of colors (Corchs, Ciocca, Bricolo, & Gasparini, 2016). Understanding visual complexity is crucial across domains, from user interface design (Miniukovich, Sulpizio, & De Angeli, 2018; Akça & Ömer Özgür Tanriöver, 2021) to cognitive psychology (Forsythe, 2009; Madan, Bayer, Gamer, Lonsdorf, & Sommer, 2018). While deep learning models predict complexity well, their black-box nature (Li et al., 2022) limits interpretability, which is essential for applications like education (Stoesz, Niknam, & Sutton, 2020; Ghai & Tandon, 2022) and information visualization (Y. Zhu, Suo, & Owen, 2007b, 2007a). Developing interpretable, perception-aligned features remains a key research goal.

A recent study by (Shen, Nath, Brielmann, & Dayan, 2024) bridged the gap between handcrafted features and deep learning approaches by proposing a simple two-feature model based on outputs from deep segmentation models (the numbers of segments and semantic classes in the image). Their simple, yet interpretable, model achieved superior performance compared to baseline on many naturalistic and art datasets, supporting the idea that the numbers of segments and classes are generic features that explain complexity well. However, their model fails to take into account the structure and arrangement of components in the image.

Here, we build upon the model from (Shen et al., 2024), by studying two of its major failure modes. First, we find that datasets with high structural and color regularity require two low-level features to explain complexity: Multi-Scale Sobel Gradient (MSG) and Multi-Scale Unique Color (MUC). MSG captures continuous spatial intensity variations across multiple scales, offering a richer representation of image structure. MUC quantifies color diversity also at multiple scales, and at different color resolutions. The latter feature builds on the concept of "colorfulness" from (Teresa, Ciocca, & Gagliardi, 2014), and provides a more robust and performant measure of chromatic complexity.

We discovered a second failure mode of (Shen et al., 2024) which concerns the rather underexplored contribution of whole-image, holistic or emergent information to visual complexity. Previous research by (Forsythe, Mulhern, & Sawey, 2008) demonstrated that participants in a visual complexity experiment exhibited a familiarity bias, perceiving familiar shapes as less complex compared to unfamiliar or novel ones. Similarly, (Snodgrass & Vanderwart, 1980b) found a negative correlation between visual complexity and familiarity in their experiment. (Forsythe, 2009) suggests that complexity judgments are context-dependent and that familiarity should be considered when developing a model of visual complexity. Motivated by their findings, we show that holistic, whole-image surprise judgments of scenes contribute significantly to the visual complexity of naturalistic images. We define surprising images as those containing unusual, or contextually novel elements, i.e., the opposite of familiarity. (Meyer, Reisenzein, & Schützwohl, 1997; Ortony & Partridge, 1987). To this end, we introduce a novel dataset, SVG, containing both surprising and randomly sampled images from the well-studied Visual Genome dataset (Krishna et al., 2016) and show that surprise scores generated by a large language model (LLM) explains significant variance in the

*Joint supervision

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

perceived complexity of images in this dataset.

Methods

Modeling and Evaluation

In this work, we evaluate the effectiveness of sets of features to explain visual complexity on datasets using linear regression models. We follow the procedure in (Shen et al., 2024), and fit M repetitions of 3-fold cross-validated linear regression. M is determined by the size of the dataset to address statistical variability in smaller datasets. We measure performance using the mean Spearman correlation coefficient across all test splits. We compare the performance of our models to the same baselines in (Shen et al., 2024): handcrafted features (Corchs et al., 2016; Kyle-Davidson et al., 2023) and deep learning approaches (Saraee, Jalal, & Betke, 2020; Feng et al., 2023) for comparison. The code for our analysis, experimental setup, and dataset will be made publicly available upon acceptance.

Datasets

We evaluate our method on four publicly available datasets with human-rated complexity scores: RSIVL (49 images) (Corchs et al., 2016), VISC (800 images) (Kyle-Davidson et al., 2023), Savoias (1400 images across 7 categories) (Saraee et al., 2020), and IC9600 (9600 images across 8 categories) (Feng et al., 2022). We select image subsets to demonstrate the effectiveness of our new features to explain complexity - VISC and IC9600 architecture subset for MSG, Savoias Art and Suprematism subsets for MUC, and Savoias Interior Design and IC9600 Abstract for the combination of MSG and MUC. Finally, we use our novel dataset SVG to show that surprisal contributes significantly to complexity.

SVG: A Dataset of Surprising Images

To the best of our knowledge, there is currently no existing image set with surprising images that can be used to study visual complexity. To fill this gap, we introduce a new dataset containing 100 highly surprising images and 100 (on average, less surprising) images, the latter of which were sampled uniformly at random from the Visual Genome, subject to stratification according to bins of complexity based on IC9600 scores (Feng et al., 2023). To confirm that our subjectively selected images are more surprising than randomly sampled ones, we compare the distributions of LLM-generated surprise scores between the two subsets. A histogram of these scores shows clear separation: sampled images cluster at lower values, while handpicked images have higher surprise scores (mean: 39.45 for ordinary, 72.35 for surprising). A Kolmogorov-Smirnov test confirms a significant difference ($D = 0.67, p < 0.001$), supporting our hypothesis.

We collect visual complexity ratings from humans using an online experiment on Prolific (Palan & Schitter, 2018). Participants are shown pairs of images and asked to select the one that is more visually complex, similarly to the procedure in (Saraee, Jalal, & Betke, 2018). We followed the sampling strategy in (Saraee et al., 2018), and collect $6000 = 200 \times 30$ comparisons by sampling images with probabilities inversely proportional to the total number of times they have already been selected. We ensure that each pairwise comparison is assessed by three different participants. Hence there are in total $18,000$ evaluations. Each participant evaluates 200 pairwise comparisons in one session. We convert pairwise comparisons into scalar complexity ratings using the Bradley-Terry algorithm (Bradley & Terry, 1952).

Multi-Scale Sobel Gradient

We propose the Multi-Scale Sobel Gradient (MSG) algorithm (Algorithm 1) to capture structure in images. MSG analyzes spatial intensity variations across multiple resolutions in an RGB image. Note that the Sobel operator is typically applied to grayscale images (Kanopoulos, Vasanthavada, & Baker, 1988), but we found in ablations that the grayscale version of the algorithm (which first converts the colored images to grayscale) did not perform as well. Hence we use the color version of the algorithm for the rest of this work.

Algorithm 1 Multi-Scale Sobel Gradient

Require: Input image $img \in \mathbb{R}^{H \times W \times 3}$ in RGB format
Ensure: Scalar MSG score
Normalize image: $img \leftarrow img/255.0$
Define scales $S = \{1, 2, 4, 8\}$
and weights $W = \{0.4, 0.3, 0.2, 0.1\}$
Initialize $MSG \leftarrow 0$
for each scale $s \in S$ and weight $w \in W$ **do**
  Resize image to $\frac{H}{s} \times \frac{W}{s}$:
  $scaled \leftarrow Resize(img, (\frac{H}{s}, \frac{W}{s}))$
  for each channel $c \in \{R, G, B\}$ **do**
    $G_x = Sobel(scaled[:,:,c], dx = 1, dy = 0)$
    $G_y = Sobel(scaled[:,:,c], dx = 0, dy = 1)$
    $mag = \sqrt{G_x^2 + G_y^2}$
    $grad_c \leftarrow Mean(mag)$
  end for
  $s_grad \leftarrow Mean(grad_R, grad_G, grad_B)$
  $MSG \leftarrow MSG + w \cdot s_grad$
end for
return MSG

Multi-Scale Unique Color

We introduce MUC which intuitively counts the number of unique colors present in an image. We start from the "colorfulness" feature introduced by (Teresa et al., 2014) (which was first applied to image indexing and content querying) and derive MUC by making colorfulness multi-scale. Note, (Teresa et al., 2014) and subsequent works did not provide pseudocode for the colorfulness algorithm - we are the first to do so. MUC has two hyperparameters which control the coarse-graining of spatial resolution and color resolution (bit precision). In the rest of this paper we report results using the best bit precision for each dataset in terms of correlation to

complexity (this was typically 7–8 bits). We use a fixed set of spatial scale weights for both MSG and MUC on all datasets.

Algorithm 2 Multi-Scale Unique Color

Require: Image I (RGB), Number of bits to preserve per channel b
Ensure: Multi-Scale unique color score
Define scales $S = \{1, 2, 4, 8\}$
and weights $W = \{0.4, 0.3, 0.2, 0.1\}$
Initialize $MUC \leftarrow 0$
for each scale $s \in S$ and weight $w \in W$ **do**
  Resize image to $\frac{\text{width}}{s} \times \frac{\text{height}}{s}$:
  $I_s \leftarrow \text{resize}(I_{rgb}, (\frac{\text{width}}{s}, \frac{\text{height}}{s}))$
  Calculate bit shift: $shift \leftarrow 8 - b$
  Quantize image by bit shifting:
  $I_q \leftarrow (I_s \gg shift) \ll shift$
  Flatten image: $I_{flat} \leftarrow \text{reshape}(I_q, -1, 3)$
  Create color index:
  $idx \leftarrow I_{flat}[:,0] \cdot 2^{16} + I_{flat}[:,1] \cdot 2^8 + I_{flat}[:,2]$
  Count unique colors: $n_{\text{unique}} \leftarrow |\text{unique}(idx)|$
  Add weighted contribution:
  $MUC \leftarrow MUC + (w \cdot n_{\text{unique}})$
end for
return MUC

Generating Surprise Scores using LLMs

In order to have an automated pipeline for predicting complexity, we prompt an LLM using zero-shot Chain-of-Thought (Wei et al., 2023) to get surprise scores for each image in our SVG dataset. Our prompt is shown in Algorithm 3. We selected GEMINI-1.5-FLASH for its ability to provide rapid responses without compromising quality. The model generated surprise scores with step-by-step reasoning, producing structured output in the desired format. We also collect human surprise ratings on the SVG image set to validate our LLM-generated scores using a similar Prolific experiment to that described above.

Algorithm 3 Zero-shot-CoT

Q: Step by step, explain why this image is surprising or not. Consider factors like rare events, or unexpected content. Be precise in your reasoning. Then, on a precise scale from 0 to 100, rate the surprisal of this image.
Provide your reasoning and numeric rating as follows:
Reasoning: [your explanation]
Rating: ≪number≫

Results

Surprise to the Rescue: Capturing the Complexity Beyond Structural Features

The features we introduced: MSG, MUC, and Surprise, explain additional variance in perceived complexity beyond that captured by segmentation and class counts alone. We present these results in Table 1, showing that our models not only surpass all handcrafted features but also achieves performance comparable to the best supervised model. In the following subsections, we present each feature in turn and describe the datasets on which they improve complexity prediction. We then describe our penultimate model, which combines our three novel features to achieve state-of-the-art or near state-of-the-art performance on all datasets.

Explaining the Failure Modes with a Data-Centric Approach: Colors and Structure

On the VISC dataset, the MSG feature improves the mean Spearman's rank correlation coefficient from 0.56 to 0.68, surpassing all previous handcrafted features. Similarly, it achieves a notable increase of 0.10 in mean Spearman's rank correlation coefficient on the IC9. Architecture dataset. These two datasets comprise of mostly real-world natural images of buildings, and indoor and outdoor scenes which are highly textured or regular - for example a grass field, the windows on a highrise, or rows of seats in an airport. A Sobel operator applied to a spatially coarse-grained image can be interpreted as an (a) symmetry detector. Hence MSG detects (a)symmetry and quantifies the regularity of local patches (e.g. windows on a highrise) at multiple spatial scales, making it a well-suited feature for predicting complexity on VISC and IC9. Architecture.

Table 1 shows that MUC improves model performance on Sav. Art dataset, increasing Spearman's correlation from 0.73 to 0.81, and in Sav. Suprematism, where it improves performance from 0.89 to 0.94. These datasets are comprised of 2D art. The efficacy of MUC in these domains likely stems from its ability to quantify color-based complexity, which may be more crucial for people's aesthetic judgments than judgments of naturalistic images.

Two datasets, Sav. Interior Design and IC9. Abstract experience very large performance gains (from 0.61 to 0.87 and from 0.66 to 0.83 respectively) when both MSG and MUC features are included in the model. Interior Design is a dataset of synthetic renderings of rooms for IKEA advertisements. The dataset likely benefits from MSG for the same reason as VISC - the images contain regular, highly symmetric elements like books in a bookshelf or panels of a kitchen cabinet. Furthermore, unlike naturalistic images Sav. Interior Design images were designed with color themes in mind for aesthetic reasons, similar to art images which is likely why the Sav. Interior Design dataset benefits from MUC like the Sav. Art dataset. IC9. Abstract on the other hand, consists of real-life photographs. However, the images are taken in an artistic way with aesthetic value in mind, likely increasing the importance of colorfulness in predicting complexity.

Figure 1 presents two examples from the Sav. Interior Design dataset where the baseline model (Shen et al., 2024) assigns the same complexity due to similar SAM segmentations and FC-CLIP classes, despite differing ground truth complexities. However, the two images differ greatly in their MSG

Table 1: Model performance on 7 datasets and comparison with previous models. The models from previous work are classified as being based on either handcrafted features or Convolutional Neural Networks (CNNs). * for supervised methods indicate their own *test* set. Bold indicates the best model.

Model/Dataset	*VISC*	*IC9. Arch.*	*Sav. Art*	*Sav. Suprematism*	*Sav. Int.*	*IC9. Abstract*	*SVG*
Handcrafted features							
Corchs 1 (10 features)	0.62	0.66	0.68	0.80	0.85	0.74	0.73
Kyle-Davidson 1 (2 features)	0.60	0.54	0.55	0.79	0.74	0.69	0.70
CNNs							
Saraee (transfer)	0.58	0.59	0.55	0.72	0.75	0.67	0.72
Feng (supervised)	**0.72**	**0.92***	**0.81**	0.84	**0.89**	**0.94***	0.83
Previous simple model							
$\sqrt{num_seg} + \sqrt{num_class}$	0.56	0.66	0.73	0.89	0.61	0.66	0.78
Visual features							
$MSG + MUC$	0.60	0.65	0.64	0.91	0.84	0.76	0.72
Baseline + visual features							
$\sqrt{num_seg} + \sqrt{num_class} + MSG$	0.68 $\uparrow$ 0.13	0.76 $\uparrow$ 0.10	0.75	0.90	0.79	0.79	0.78
$\sqrt{num_seg} + \sqrt{num_class} + MUC$	0.62	0.71	**0.81** $\uparrow$ 0.08	**0.94** $\uparrow$ 0.05	0.80	0.76	0.79
$\sqrt{num_seg} + \sqrt{num_class} + MSG + MUC$	0.68	0.77	0.81	0.94	0.87 $\uparrow$ 0.26	0.83 $\uparrow$ 0.17	0.79
Baseline + semantic feature							
$\sqrt{num_seg} + \sqrt{num_class} + Surprise$	0.60	0.67	0.74	0.89	0.60	0.67	0.83 $\uparrow$ 0.05
Baseline + all features							
$\sqrt{num_seg} + \sqrt{num_class} + MSG + MUC + Surprise$	0.71	0.78	**0.81**	**0.94**	0.87	0.84	**0.85**

values (20 vs. 92 on a normalized scale of 100), possibly because the lower image has patterns on the bookshelves and blanket which contribute substantially to MSG. The lower (higher) MSG values in the top (bottom) image decrease (increase) the complexity predictions in model, making the predictions much closer to the ground truth complexity ratings.

Table 1 shows that including surprise scores generated using GEMINI-1.5-FLASH improves Spearman's correlation on the SVG dataset from 0.78 to 0.83 over (Shen et al., 2024)'s baseline model. Our model containing surprise scores matches the performance of (Feng et al., 2023)'s supervised network.

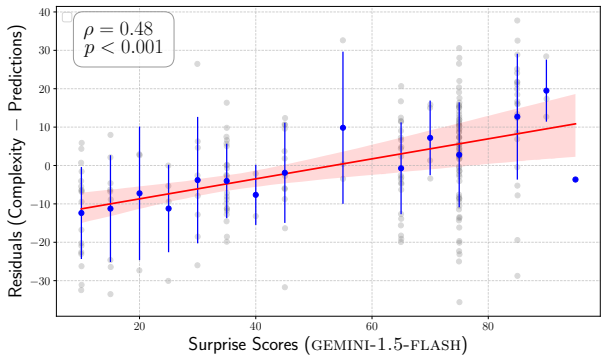

Figure 3: Correlation between residuals (actual complexity - baseline predictions) and surprise scores (SVG).

Furthermore, Figure 3 shows that the surprise scores have a correlation of 0.48 ($p < 0.001$) with complexity after regressing out SAM number of segments and FC-CLIP number of objects, showing that surprise captures variance in complexity not explained by the number of segments or objects. These quantitative results are corroborated by participants' anecdotes on how they rated complexity on the SVG dataset. Many reported that "unusual", "surprising", or "weird", objects or events in the image, as well as judgments that the images are "confusing" or "unreal" caused them to rate the visual complexity higher. Figure **??** presents a selection of responses from participants in our experiment, in which they describe the specific strategies they employed to assess visual complexity. Their responses provide a deeper understanding of the cognitive processes involved in evaluating image complexity and highlight how the *element of surprise* works as a latent dimension for visual complexity evaluation.

Figure 2 shows two examples from the SVG dataset where the model from (Shen et al., 2024) underpredicts visual complexity. Both images have high (85) surprise scores. Hence our model including surprise scores correctly increases its complexity predictions, making the predictions much closer to the ground truth complexity. Furthermore, one advantage of using LLMs in our automated pipeline for complexity prediction is LLMs offer (often) interpretable, explicit reasoning in natural language (Singh, Inala, Galley, Caruana, & Gao, 2024). Two examples of explanations generated from GEMINI-1.5-FLASH using CoT prompting (Z. Zhu et al., 2024; Wei Jie, Satapathy, Goh, & Cambria, 2024), can be seen in Figure 2. In the top image, GEMINI-1.5-FLASH successfully identifies the intuitively surprising fact that sheep are racing, an unusual activity for sheep. In the bottom image, GEMINI-1.5-FLASH identifies that it's unusual for the man to be doing an uncontrolled backflip, arguably consistent with human intuition for why the image is surprising.

To explore quantitatively how well our GEMINI-1.5-FLASH surprise scores aligned with human judgments of surprise, we computed Spearman's correlation between GEMINI-1.5-FLASH scores and the human surprise scores we collected. We find a high correlation of 0.75. This supports the fact GEMINI-1.5-FLASH explanations are good proxies for human judgments of image surprisal which contribute to hu-

(a) B: 50, P: 39, G: 40 (b) MSG: 20

(c) B: 50, P: 67, G: 68 (d) MSG: 92

Figure 1: Left column: original images from *Sav. Int.*. Right column: gradient visualizations. B: baseline prediction using number of segmentations and classes. G = ground truth complexity. P = predicted complexity using baseline and MSG. All values are scaled between 0 and 100. The first image has 177 segmentations and 35 classes, while the second has 185 segmentations and 38 classes. Due to these similarities, the baseline model predicts nearly identical complexity scores for both. However, MSG acts as a latent dimension, refining predictions to better align with ground truth.

Surprise score: 85
Reasoning: The image depicts sheep dressed in jockey silks and participating in a race. This is surprising because it is not a typical activity for sheep. The juxtaposition of animals in a human sport is unexpected and humorous.

Surprise score: 85
Reasoning: The image depicts a skateboarder performing an unexpected maneuver on a beach, resulting in an unusual pose mid-air. The rarity stems from the seemingly uncontrolled backflip and the unusual beach setting for such a stunt.

Figure 2: Racing sheep with toy riders or a flying skater from the **SVG** dataset illustrate the improvement in complexity predictions when incorporating surprise. The baseline model underestimates the visual complexity, with scores of B: 54 and B: 46, while ground truth values are G: 60 and G: 55. Incorporating surprise scores (85) reduces this gap, yielding adjusted predictions of P: 61 and P: 55, demonstrating the role of surprise in aligning predictions more closely with human perception. Explanations provided by GEMINI-1.5-FLASH enhance the interpretability of the assigned surprise scores.

man perceptions of visual complexity. In summary, GEMINI-1.5-FLASH surprise scores improve the predictive power of our complexity model on the SVG dataset and the natural language explanations generated alongside these scores facilitate a clearer understanding of why the image is complex.

The last row of Table 1 shows the performance of our penultimate model including MSG, MUC, surprise as well as the two features from (Shen et al., 2024). Interestingly, we find combining MSG and MUC with surprise only improves performance on all datasets over using either of these features alone. Furthermore, our penultimate model is significantly more performant than every baseline except (Feng et al., 2023)'s supervised network which is competitive. Taken together, in contrast to what (Shen et al., 2024) previously suggested, the complexity of naturalistic images cannot be explained simply using two generic segment and object features. Some datasets require additional features to explain visual complexity fully.

In Figure 4, we present another pair of images where, this time, the addition of surprise scores causes their predicted complexity values to diverge in opposite directions.

Discussion

This work introduces three features to solve the inadequacy of the generic constructs proposed by (Shen et al., 2024) to explain complexity in certain datasets: multi-scale gradient analysis, multi-scale unique color, and a surprise score. We experimentally show that complexity can be multifaceted - requiring both generic features (segmentation and object counts) and dataset specific features to predict accurately. Our results demonstrate that pixel-level structural and chromatic features, MSG and MUC effectively capture key aspects of visual complexity, with MSG quantifying continuous spatial intensity variations across scales and MUC capturing color diversity, both complementing object-level information from deep segmentation models. Additionally, we introduce surprise, an image-level "cognitive" feature, as a previously underexplored dimension of visual complexity, representing the degree to which the image as a whole deviates from expected objects or events. We introduce a new dataset SVG containing images with unexpected or novel elements to show that surprise contributes significantly to perceived complexity. We constructed our SVG dataset by manually selecting surprising images and hence acknowledge the limitations in scalability and the potential subjectivity of this approach. We hope our preliminary yet confirmatory results will inspire future research to develop larger datasets using objectively defined measure to further validate the link between surprise and complexity.

MSG versus Patch Symmetry and Edge Density Patch symmetry (Kyle-Davidson et al., 2023; Olivia et al., 2004; Rosenholtz, Li, & Nakano, 2007) and edge density (Guo, Qian, Li, & Asano, 2018; Ciocca, Corchs, & Gasparini, 2015) are two commonly used low-level features for cap-

(a) *num_seg* = 191, *num_class* = 9, *MSG* = 59, *MUC* = 922, *Surprise* = 25

(b) *num_seg* = 172, *num_class* = 9, *MSG* = 56, *MUC* = 820, *Surprise* = 85

Surprise as the hidden dimension
We present a case where both images share similar visual features, including segmentations, classes, MSG, and MUC. The right image, with lower feature values, was initially predicted to have lower complexity. The left image received a complexity score of 60, while the right received 55. However, the element of surprise shifts the outcome, with improved predictions at 52 (72) compared to ground truth values of 25 (74). Despite their similarity, humans may perceive the second image as more complex due to its unusual composition, where multiple traffic signs are clustered together.

Figure 4: Comparison of two images having similar values of visual features but with differing complexity evaluations.

Figure 5: Representative samples from previous datasets show initial complexity predictions ($B = 52$), derived from $\sqrt{num_seg}$ and $\sqrt{num_class}$, refined using **MSG** (VISC, IC9 Architecture), **MUC** (Sav. Art, Sav. Suprematism), and **MSG + MUC** (Sav. Int. Design, IC9 Abstract). The enhanced predictions (P) align closer to ground truth scores (G).

turing structural information in an image. We run ablation studies to investigate whether MSG is more performant than these two features. We evaluate on all 16 datasets: VISC, RSIVL, SAVOIAS (5 categories - excluding Advertisement and Visualizations due to heavy presence of text), IC9600 (8 categories), and SVG using permutation tests. We compare MSG against patch symmetry because (Shen et al., 2024) reported that it captures structural information well on the VISC dataset. We compare MSG against Canny edge density because Canny is a standard metric for predicting visual complexity. It has been applied in (Guo et al., 2018; Corchs et al., 2016; Rosenholtz et al., 2007).

MSG performs significantly better than Canny edge density in five datasets. For example, on IC9. Abstract, MSG achieves significantly higher correlation with the complexity compared to edge density (0.517 vs 0.705). Edge density only outperformed MSG on two datasets. Compared to patch symmetry, MSG performs better on 7 datasets. Patch symmetry outperforms MSG on only one dataset. Our ablations show that MSG is currently the most performant structural feature for predicting complexity, outperforming both edge density and patch symmetry across diverse image datasets.

Single versus Multi-Scale We perform ablation studies to investigate the necessity of multi-scale computations for MSG and MUC. We measure the correlation between complexity and our visual features using both single and multiple scales. We find that multi-scale MSG outperforms single-scale MSG on 6 datasets, while single-scale MSG outperforms multi-scale MSG on just two datasets out of the 16 datasets tested. The advantage of the multi-scale approach was particularly pronounced in Sav. Suprematism, where correlation increased substantially from 0.474 to 0.616. Similarly, multi-scale MUC outperformed single-scale MUC on 5 datasets, while single-scale MUC outperformed multi-scale MUC on only one dataset. The multi-scale versions of MSG and MUC maximize the correlation with the perceived visual complexity.

Future work While our work uses whole-image, holistic surprise scores, GEMINI-1.5-FLASH identifies specific object relationships and object-contexts that make each image more or less surprising. Therefore, it would be interesting for future works to derive surprise scores (or even other semantic features) from object-centric representations such as scene graphs. Our dataset, SVG enables this type of investigation by including human-rated complexity scores, and inheriting region descriptions, and scene graphs from the Visual Genome dataset. These rich annotations open a promising path for developing and validating features that reflect semantic rather than purely visual contributions to complexity.

References

Akça, E., & Ömer Özgür Tanriöver. (2021). A comprehensive appraisal of perceptual visual complexity analysis methods in gui design. *Displays*, *69*, 102031. doi: https://doi.org/10.1016/j.displa.2021.102031

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, *39*(3/4), 324–345. Retrieved from https://doi.org/10.2307/2334029 (Accessed 21 Jan. 2025) doi: 10.2307/2334029

Ciocca, G., Corchs, S., & Gasparini, F. (2015). Complexity perception of texture images. In V. Murino, E. Puppo, D. Sona, M. Cristani, & C. Sansone (Eds.), *New trends in image analysis and processing – iciap 2015 workshops* (pp. 119–126). Cham: Springer International Publishing.

Corchs, S. E., Ciocca, G., Bricolo, E., & Gasparini, F. (2016). Predicting complexity perception of real world images. *PLOS ONE*, *11*(6), e0157986. doi: 10.1371/journal.pone.0157986

Feng, T., Zhai, Y., Yang, J., Liang, J., Fan, D.-P., Zhang, J., ... Tao, D. (2023). Ic9600: A benchmark dataset for automatic image complexity assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*(01), 1–17. doi: 10.1109/TPAMI.2022.3232328

Forsythe, A. (2009). Visual complexity: Is that all there is? In D. Harris (Ed.), *Engineering psychology and cognitive ergonomics* (pp. 158–166). Berlin, Heidelberg: Springer Berlin Heidelberg.

Forsythe, A., Mulhern, G., & Sawey, M. (2008, 03). Confounds in pictorial sets: The role of complexity and familiarity in basic-level picture processing. *Behavior research methods*, *40*, 116-29. doi: 10.3758/BRM.40.1.116

Ghai, A., & Tandon, U. (2022, 01). Analyzing the impact of aesthetic visual design on usability of e-learning: An emerging economy perspective. *Higher Learning Research Communications*, *12*, 1-22. doi: 10.18870/hlrc.v12i2.1325

Guo, X., Qian, Y., Li, L., & Asano, A. (2018). Assessment model for perceived visual complexity of painting images. *Knowledge-Based Systems*, *159*, 110-119. doi: https://doi.org/10.1016/j.knosys.2018.06.006

Kanopoulos, N., Vasanthavada, N., & Baker, R. (1988). Design of an image edge detection filter using the sobel operator. *IEEE Journal of Solid-State Circuits*, *23*(2), 358-367. doi: 10.1109/4.996

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... Li, F.-F. (2016). *Visual genome: Connecting language and vision using crowdsourced dense image annotations.* Retrieved from https://arxiv.org/abs/1602.07332

Kyle-Davidson, C., Solis, O., Robinson, S., Tan, R. T. W., & Evans, K. K. (2025). Scene complexity and the detail trace of human long-term visual memory. *Vision Research*, *227*, 108525. doi: https://doi.org/10.1016/j.visres.2024.108525

Kyle-Davidson, C., Zhou, E. Y., Walther, D. B., Bors, A. G., & Evans, K. K. (2023). Characterising and dissecting human perception of scene complexity. *Cognition*, *231*, 105319. doi: https://doi.org/10.1016/j.cognition.2022.105319

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., ... Dou, D. (2022). *Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond.* Retrieved from https://arxiv.org/abs/2103.10689

Madan, C. R., Bayer, J., Gamer, M., Lonsdorf, T. B., & Sommer, T. (2018). Visual complexity and affect: Ratings reflect more than meets the eye. *Frontiers in Psychology*, *8*, 2368. doi: 10.3389/fpsyg.2017.02368

McDougall, S. J., Curry, M. B., & de Bruijn, O. (1999). Measuring symbol and icon characteristics: Norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols. *Behavior Research Methods, Instruments, & Computers*, *31*(3), 487–519. doi: 10.3758/BF03200730

Meyer, W. U., Reisenzein, R., & Schützwohl, A. (1997). Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion*, *21*(3), 251–274. Retrieved from https://doi.org/10.1023/A:1024422330338 doi: 10.1023/A:1024422330338

Miniukovich, A., Sulpizio, S., & De Angeli, A. (2018). Visual complexity of graphical user interfaces. In *Proceedings of the 2018 international conference on advanced visual interfaces.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3206505.3206549 doi: 10.1145/3206505.3206549

Olivia, A., Mack, M. L., Shrestha, M., & Peeper, A. S. (2004). Identifying the perceptual dimensions of visual complexity of scenes.. Retrieved from https://api.semanticscholar.org/CorpusID:8049814

Ortony, A., & Partridge, D. (1987, 01). Surprisingness and expectation failure: What's the difference? In (p. 106-108).

Palan, S., & Schitter, C. (2018). Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22-27. doi: https://doi.org/10.1016/j.jbef.2017.12.004

Rosenholtz, R., Li, Y., & Nakano, L. (2007). Measuring visual clutter. *Journal of vision*, *7 2*, 17.1–22. Retrieved from https://api.semanticscholar.org/CorpusID:10847280

Saraee, E., Jalal, M., & Betke, M. (2018). *Savoias: A diverse, multi-category visual complexity dataset.* Retrieved from https://arxiv.org/abs/1810.01771

Saraee, E., Jalal, M., & Betke, M. (2020). Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding*, *195*, 102949. doi: https://doi.org/10.1016/j.cviu.2020.102949

Shen, T., Nath, S. S., Brielmann, A., & Dayan, P. (2024). *Simplicity in complexity : Explaining visual complex-*

ity using deep segmentation models. Retrieved from https://arxiv.org/abs/2403.03134

Singh, C., Inala, J. P., Galley, M., Caruana, R., & Gao, J. (2024). *Rethinking interpretability in the era of large language models.* Retrieved from https://arxiv.org/abs/2402.01761

Snodgrass, J. G., & Vanderwart, M. (1980b). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174–215. doi: 10.1037//0278-7393.6.2.174

Snodgrass, J. G., & Vanderwart, M. L. (1980a). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology. Human learning and memory*, *6 2*, 174–215. Retrieved from https://api.semanticscholar.org/CorpusID:16871046

Stoesz, B., Niknam, M., & Sutton, J. (2020, 12). Defining the visual complexity of learning management systems using image metrics and subjective ratings. *Canadian Journal of Learning and Technology / La revue canadienne de l'apprentissage et de la technologie*, *46*. doi: 10.21432/cjlt27899

Teresa, A., Ciocca, G., & Gagliardi, I. (2014, 06). Good 50×70 project: A portal for cultural and social campaigns. *Archiving Conference*, *11*. doi: 10.2352/issn.2168-3204.2014.11.1.art00048

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models.* Retrieved from https://arxiv.org/abs/2201.11903

Wei Jie, Y., Satapathy, R., Goh, R., & Cambria, E. (2024). How interpretable are reasoning explanations from prompting large language models? In *Findings of the association for computational linguistics: Naacl 2024* (p. 2148–2164). Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.138

Zhu, Y., Suo, X., & Owen, G. (2007b, 11). Complexity analysis for information visualization design and evaluation. In (Vol. 4841, p. 576-585). doi: 10.1007/978-3-540-76858-6_56

Zhu, Y., Suo, X., & Owen, G. S. (2007a). Complexity analysis for information visualization design and evaluation. In G. Bebis et al. (Eds.), *Advances in visual computing* (pp. 576–585). Berlin, Heidelberg: Springer Berlin Heidelberg.

Zhu, Z., Chen, H., Ye, X., Lyu, Q., Tan, C., Marasovic, A., & Wiegreffe, S. (2024, June). Explanation in the era of large language models. In R. Zhang, N. Schneider, & S. Chaturvedi (Eds.), *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 5: Tutorial abstracts)* (pp. 19–25). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.naacl-tutorials.3/ doi: 10.18653/v1/2024.naacl-tutorials.3