The Minimum Semantic Content (MSC) Dataset: A Large, Balanced Resource for Computational Aesthetics Research

Olivier Penacchio[a,b,c,*], Arslan Javed[a,b], Bogdan Raducanu[a,b], Xavier Otazu[a,b] and C.Alejandro Parraga[a,b]

[a]*Computer Science Dept., Engineering School, Universitat Autònoma de Barcelona (UAB), Campus UAB, Bellaterra, 08193, Barcelona, Spain.*
[b]*Computer Vision Centre, Campus UAB, Bellaterra, 08193, Barcelona, Spain.*
[c]*School of Psychology and Neuroscience, University of St Andrews, St Andrews, Fife KY16 9JP, United Kingdom.*
**Corresponding author email: penacchio@cvc.uab.cat*

Abstract

Image databases are central to empirical aesthetics, enabling tests of how image statistics relate to observers' appreciation. However, many existing databases have two key limitations: (1) they conflate low-level visual features with high-level semantic content, making it difficult to separate visual from cognitive influences on aesthetic judgments; and (2) they are imbalanced, overrepresenting highly appreciated images. To address these issues, we present the Minimum Semantic Content (MSC) database, a large, systematically curated resource for computational aesthetics. It comprises 10,426 natural scenes with reduced, homogenized semantic content, minimizing cognitive and emotional confounds. Each received 100 individual aesthetic ratings from naïve observers, drawn from a pool of approximately 10,000 participants, via crowdsourcing. The database includes both "beautified" and "uglified" versions, generated with a manipulation technique that promotes uniform coverage across the aesthetic spectrum. This broader distribution mitigates bias and overfitting in models. Validation also shows improved robustness in computational models overall. This database enables researchers to study how perceptual features shape aesthetic judgments, using stimuli with very limited semantic and contextual confounds.

Keywords: computational aesthetics, crowdsourcing, image preference, image uglification, low-level image features, balanced sampling

Background and summary

A central aim in empirical aesthetics is to uncover the perceptual and cognitive mechanisms that shape aesthetic judgments[1, 2]. To do this, many studies adopted a "bottom-up" approach, which posits that aesthetic responses emerge from a hierarchy of visual features—ranging from low-level sensory inputs to high-level semantic interpretations. Low-level features, such as contrast, lightness, saturation, and color, are processed early in the visual stream[3]. At the other end of the hierarchy, high-level features involve object recognition and semantic associations, linking visual input to meaning[4, 5]. Empirical studies have shown that features like color, contrast, saturation, aspect ratio, and contour modestly predict aesthetic preferences[6-8]. Additional predictors based on low-level features include image complexity, anisotropy, fractal self-similarity[9, 10] and image quality[11]. Symmetry has also been identified as a key factor, particularly in abstract patterns and facial aesthetics[12-14], as has the preference for smooth over angular contours[15-18]. Despite these insights, the predictive power of individual low-level features remains limited, except in cases where semantic content is minimized or uniform (such as in abstract artworks)[19] highlighting the need for more integrative models that capture interactions across multiple levels of visual processing.

One of the reasons for the lack of predictive power of low-level features is arguably the strong confound present at the level of the image databases used to evaluate them. Most databases are constructed by selecting examples from a limited number of image-sharing platforms (e.g. DPChallenge.com[20], Flickr.com[21]) where submissions are categorized under specific labels such as "titles," "themes," or "challenges". These labels provide a semantic context that can strongly influence how viewers judge an image's aesthetic value. For example, users on DPChallenge.com might rate an image more favorably if it closely matches the challenge theme—such as preferring a photo of a messy-haired pet for "Bad Hair Day," or nostalgic toys for "Through the Eyes of a Child." In some cases, winning entries are chosen for their humorous or topical interpretations.

This phenomenon illustrates what is referred to in computational learning as the *semantic gap*—the disconnect between the visual features extracted from an image and the meaning attributed to the image by human observers within a specific context[22]. While image-sharing platforms often follow similar structural formats, the semantic framing of images—shaped by labels, themes, or challenges—varies widely. This variability complicates efforts to predict aesthetic judgments based solely on visual content. For example, the top-rated image in the AVA database[23] is a stylized American flag, showing that symbolic and cultural factors can outweigh visual characteristics in determining aesthetic preference. This also exemplifies a form of cultural bias: although the image may be positively appraised within certain sociocultural contexts, it is plausible that the same symbol would elicit negative associations—and thus substantially lower aesthetic ratings—in other cultural or geopolitical settings. Additionally, AVA ratings are tightly clustered, with 95% of images scoring between 4 and 6 on a 1–10 scale. This compression reflects a narrow dynamic range in aesthetic judgements, restricting the ability to uncover meaningful relationships between low-level image features and viewer ratings across the full spectrum of possible ratings. Furthermore, the competitive nature of the original platforms likely discourages the posting of low-aesthetic-value—or 'ugly'—images, further biasing the database. As a result, statistical inference is constrained to a narrow subrange of aesthetic values, limiting the generalizability of findings[24].

The Minimum Semantic Content (MSC) Image Database[25] was developed to address these limitations. By systematically selecting and curating natural scenes with minimal semantic content, and by employing a novel image manipulation tool to generate both beautified and uglified versions of many scenes, the MSC database achieves a much more balanced distribution of aesthetic ratings across the full spectrum, from “very ugly” to “very beautiful” [26, 27]. Each image was evaluated by 100 naïve observers drawn from a much larger pool of approximately 10,000 participants worldwide, using a standardized rating protocol through a crowdsourcing platform. Importantly, the specific observers varied across images, ensuring both a consistent number of ratings per image and a large, diverse evaluator base, therefore avoiding the biases associated with a small, fixed group of raters.

The MSC dataset enables researchers to study how perceptual features shape aesthetic judgments using stimuli with limited semantic and contextual confounds—specifically, images that minimize recognizable objects, scenes, or symbols that could evoke high-level cognitive, emotional, or cultural interpretations. The MSC dataset therefore provides a unique resource for the empirical study of visual aesthetics, enabling the development and validation of computational models that predict aesthetic preference based on image features alone[7, 28, 29]. Limiting semantic context operationally entails excluding images containing people, animals, human-made objects, or elements carrying symbolic or functional meaning, such as written language of any kind (words, letters, numerals, signage, etc.). The resulting images are primarily natural textures, vegetation, rock formations, sky patterns, water surfaces, and other visually rich but semantically sparse content. Although such images cannot be entirely devoid of semantics, the remaining semantic content is *both reduced and homogeneous* across the database, as all stimuli depict natural scenes and textures. This homogeneity ensures that observers’ ratings primarily reflect variations in low-level visual properties rather than differences in semantic meaning or category. Moreover, a large body of vision science research has shown that the core coding mechanisms of the human visual system are adapted to the statistical regularities of natural scenes and textures[30-34]. Therefore, variations in low-level visual features within this class of stimuli are perceptually meaningful and engage early stages of visual processing, making natural scenes and textures a particularly well-suited stimulus class for investigating the perceptual foundations of visual aesthetics.

Potential MSC users include researchers in psychology, neuroscience, computer vision, machine learning, and related fields who require high-quality, well-annotated image data for investigating the mechanisms underlying aesthetic judgements, training and benchmarking algorithms, or exploring the perceptual basis of beauty and ugliness in visual stimuli.

By making the MSC Image Database openly available, we aim to support a wide range of scientific inquiries and facilitate progress toward a deeper understanding of the visual determinants of aesthetic preference[27].

Methods

Ethics statement

Participants in the rating experiment (*naïve observers*) were recruited via a third-party online gaming platform. Participation was voluntary and incentivized through in-game tokens that allowed users to continue playing the game. Prior to participation, individuals were informed that they would be asked to rate images for research purposes and that their responses would be used in anonymized form for scientific analysis and data sharing. Informed consent was obtained through the platform before participants could take part in the task.

No personally identifying information (such as names, contact details, IP addresses, or demographic data) was collected or accessible to the authors. All rating responses were provided to the authors in fully anonymized form by the crowdsourcing platform. A separate group of *informed observers*, consisting of students and staff from our research institute, performed the image manipulations ("uglification" and "beautification"). These participants provided informed consent prior to participation. Data handling and storage complied with applicable ethical and data-protection standards. The study protocol was reviewed and approved by the Comitè d'Ètica en la Recerca (CERec) of the Universitat Autònoma de Barcelona (approval number 5453).

Database creation overview

We took the following steps to generate our image database. First, we collected a large number of images, which were obtained from four sources: (i) the public-domain image repository pdphoto.org[35] (https://pdphoto.org), (ii) the public-domain repository Photos-Public-Domain[36] (https://www.photos-public-domain.com), (iii) Flickr[21] (https://www.flickr.com), restricted to images explicitly released under public-domain or Creative Commons licenses permitting reuse, modification, and redistribution (C0, CC-BY, CC-BY-SA), and (iv) the authors' own photographic collections. Images with restrictive licenses (e.g. "all rights reserved" or licenses prohibiting modification or redistribution) were excluded. Redistribution of all images as part of the MSC database[25] complies with the original licensing terms specified by the respective data providers. Second, we removed semantic and emotional content by excluding images containing rich semantic information—such as people, animals, or human-made objects—from the initial pool. We also eliminated obvious examples of "postcard landscapes" or "holiday brochure" scenes that could elicit strong emotional responses.

Figure 1: *Representative examples of the "Original" database.* The MSC database was created from a set of images that predominantly depict natural scenes and materials with limited and homogeneous semantic content. A quantitative analysis based on automatically generated image captions indicates that the most frequent elements are trees (≈60% of images), sky (≈43%), foliage (≈37%), and shadows (≈36%), followed by branches and forest scenes (≈29%), rock or rocky surfaces (≈26%), and ground or terrain-related elements (≈24%). Together, these statistics reflect that the database is dominated by vegetation, geological materials, and large-scale outdoor scenes, providing visually rich yet semantically homogeneous stimuli.

The resulting 5,684 images were then cropped to remove frames and colored borders and resized to facilitate processing and distribution, with final dimensions ranging from 293 × 334 to 800 × 1200 pixels. We refer to this

collection as the *original* database (see Figure 1). Third, we created a subset of modified images by asking a small group of observers to manipulate the spatiochromatic characteristics of selected originals to produce either deliberately unattractive or attractive scenes, hereafter referred to as *uglified* and *beautified*, respectively, following the instructions provided (see details in the next section). Fourth, we augmented the database by adding these modified images to the original collection and by generating additional images through a procedure that recorded the uglification manipulations and applied them randomly to other originals. These automatically modified images were designated *auto-uglified*, following the same naming convention.

Finally, we obtained aesthetic ratings for the entire database from many non-expert observers using a crowdsourcing paradigm implemented by a third-party provider (Knowxel Crowdmobile S.L., Cerdanyola, Barcelona). The platform included algorithms to detect and exclude participants who responded randomly or without sufficient engagement.

The rating experiment consisted of two stages. An introductory training stage presented observers with 10 example images (manipulated to be either very beautiful or very ugly) in random order, along with the instruction: "Rate the beauty of this image from 1 (very ugly) to 5 (very beautiful)." This stage familiarized observers with the full dynamic range of the database; these initial ratings were discarded. The subsequent experimental stage required observers to rate at least 14 randomly selected images from the database using the same instructions, with the option to correct accidental clicks. In total, each of the 10,426 images received 100 independent ratings. The overall participant pool exceeded 10,000 individuals, although we could not control repeated participation or collect detailed demographic information beyond the indirect indicators noted above. The final database consists of 10,426 copyright-free images, including original photographs (54.5%) as well as beautified (8.8%), uglified (8.4%), and automatically manipulated (auto-uglified, 28.3%) versions. Details of the image manipulations are provided below.

The "Uglifier", an ad-hoc image manipulation software

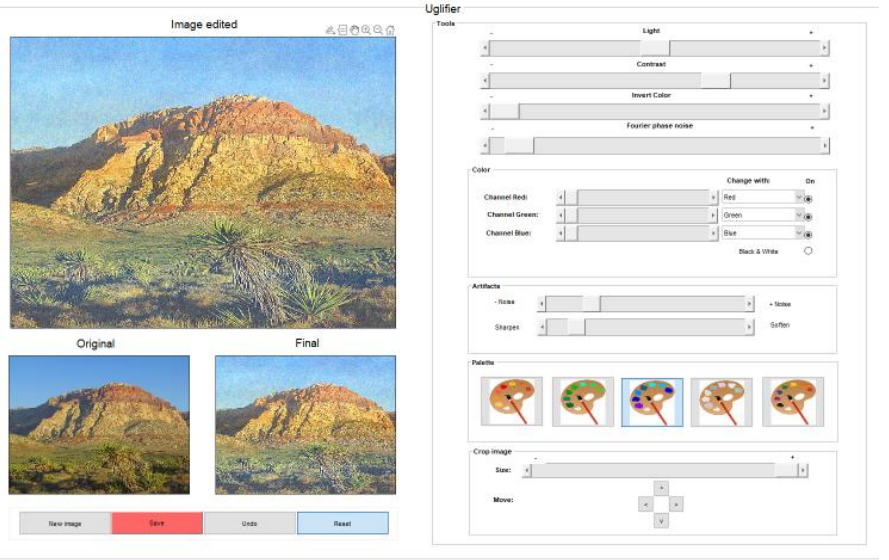


Figure 2: Screenshot of the "Uglifier" interface used to manipulate images to enhance the MSC database[25]. The tool allows users to systematically modify low-level visual features such as brightness, contrast, color channels, spatial structure (via Fourier phase noise), and noise artifacts. A range of color palettes and cropping options are also available. The example shows the original image (bottom left), the manipulated version (bottom right), and a large preview of the edited result (top). These transformations are designed to alter the aesthetic appearance of the image. The software also allowed users to undo the previous manipulations, reset the transformation and save the output always keeping the original image visible.

To compensate for the bias towards beautiful images in the original database, we made a computational tool that allowed observers to manipulate the low-level features of images (Figure 2). This tool, called the "*Uglifier*", was designed to manipulate features described as important for aesthetic valuation in previous research with a user-friendly interface.

Through a series of trials with naïve observers, we identified effective and intuitive manipulations, avoiding those that were overly complex or time-intensive to prevent participant fatigue or predictable responses. These manipulations were chosen not only for their ability to adjust key visual features—such as color, contrast, and spatial frequencies—but also to foster a broader exploration of aesthetic preferences. These manipulations are consistent with prior findings showing that variations in spatial framing[37] and color statistics[38-40] can substantially influence aesthetic judgments as well as material perception.

Table 1: Description of the different image manipulation modules of the "Uglifier" algorithm available to users.

Module name	**Effect on the image**
Lightness slider	Overall lightness increase/decrease.
Contrast slider	Overall contrast increase/decrease.
Color inversion slider	Image color inversion. Switches between opponent colors (red/green, yellow/blue and light/dark).
Fourier phase noise slider	Introduces increasing amounts of randomization to the phase component of the Fourier representation of the image. Destroys the image content while keeping the original second-order Fourier statistics.
Color manipulation (contains three separate RGB channel sliders)	RGB relative contribution is increased/decreased. It can also exchange, disable, or equalize RGB channels (i.e., convert image to grayscale).
Artifacts (contains a random noise slider and a blur/sharpen slider)	Adds random noise to the image, blurs the image or performs edge enhancement using a circularly symmetric Gaussian filter.
Palette exchange	Exchanges the color palette with that of an external image which could be either reddish, bluish, greenish, or whitish (see Pitie et al.[41]).
Cropping and positioning slider	Allows to select a portion of the original image and discard the rest.

We asked 40 observers (university students, 30 male and 10 female) to manipulate a subset of 1,791 images randomly selected from the original database (approximately 45 images each). Observers were instructed either to "beautify" an image (i.e., to make it as beautiful as possible; 919 images) or to "uglify" it (i.e., to make it as ugly as possible; 872 images). They were free to apply any combination of the available manipulation modules and to iteratively adjust their parameters until they judged the result to fulfil the given instruction. Each manipulation session produced a single final image per original; no averaging, aggregation, or combination across observers was performed. Because these participants were explicitly aware of the purpose of the manipulation task, we refer to them as *informed observers*, in contrast to the *naïve observers* involved in the subsequent rating experiment (crowdsourcing).

Figure 3: *Examples of images created by informed observers instructed to make ugly images.* Each image pair contains the same image before (original) and after (uglified) being manipulated with the "Uglifier".

We recorded the manipulations performed when informed observers uglified images and generated another subset of 2,951 images by applying them randomly to the same number of original images (auto-uglified images). The idea here was to augment the data destined to counter the beauty bias since the manipulated images were expected to score low in the aesthetics axis. Table 1 shows the list and description of the image manipulation modules available to users of the "Uglifier" algorithm.

Figure 3 shows examples of images manipulated using the Uglifier. After adding the modified images to the original database, we ended up with a total of 10,426 images.

Obtaining aesthetic valuations from rating results (crowdsourcing)

In our rating paradigm, 100 non-expert observers valuated each MSC image, resulting in 10,426 rating histograms. For our analysis, we converted each rating distribution to a single aesthetic valuation number. After a visual inspection of the data, we decided to fit a truncated Gaussian to each histogram. This provided two statistical descriptors: the mean (μ) and the SD of each valuation distribution. For the fittings, we assumed that the first bin corresponded to valuations between 0.5 to 1.5, the second bin to valuations between 1.5 and 2.5, and so on. Our fitting code was customized from Ryabov's code[42] with the following constraints: (i) the value of μ was restricted between 0.5 and 5.5, which is the full range of the bars and (ii) the SD of the Gaussian distribution was not allowed to have values below 0.5 (half of a histogram bar). Using these constraints, each image of the database was assigned a μ value in the range [0.5, 5.5] and an SD in the range [0.5, 1.9]. However, to allow users to adopt other fitting procedures, we provide the original set of aesthetic judgements.

Data Records

Description of the database files, formats, and structure

The MSC database[25] is distributed as a single compressed file named MSC_database.zip. The archive contains a subfolder called "dataset" containing all 10,426 images in JPEG (.jpg) format. The filenames are structured to indicate the origin and type of each image:

- Unmodified images are labeled with their original image number (e.g., *9830.jpg*, *9845.jpg*).
- Modified images (either uglified or beautified) include the suffix *_uglifier* after the original number (e.g., *650_uglifier.jpg*, *651_uglifier.jpg*) to indicate that they were manually altered using the Uglifier software.
- Randomly modified images are prefixed with *random_* before the original number (e.g., *random_3785.jpg*, *random_3786.jpg*).

In addition to the images, the database includes a comma-separated values (.csv) file that provides detailed metadata and results for each image:

- Columns B–F: Raw ratings, with each column representing the number of votes received for each of the five aesthetic rating categories (1 = very ugly, 5 = very beautiful).
- Columns H–L: Further raw ratings, indicating the category with the maximum number of votes for each image. In the case of a tie, a value of -1 is recorded.
- Columns N–R: Results from truncated histogram fittings to the ratings. These columns include the final value of μ (mean), the final value of σ (standard deviation), the final *FVAL* and *ExitFlag* values from Matlab's[43] *fminsearch* function, and the number of iterations required.
- Columns T–Z: Image categorization, specifying whether an image is "*uglified*," "*beautified*," "*auto_uglified*," or "*unmodified*." This section also indicates if an image served as the starting point for a modification (e.g., "*uglified_original*" denotes an image used as the basis for an uglification process).

This structure ensures that users can easily identify the type and provenance of each image, as well as access comprehensive metadata and aesthetic ratings for further analysis.

Data availability

The MSC database is openly available via the Open Science Framework (OSF)[25], and can be accessed at https://doi.org/10.17605/OSF.IO/ZGSVJ. The OSF project includes two components: (1) "*The MSC Database*" and (2) "*Validation Software*". The first component contains the complete image database as well as a comprehensive .csv file with the raw rating results. This arrangement allows users to easily determine the type and provenance of each image and provides access to detailed metadata and aesthetic ratings, supporting a wide range of analyses and applications. All images included in the MSC database are either in the public domain or released under licenses permitting reuse and redistribution.

Technical validation

Population-level consistency of image manipulations

Although the number of informed observers involved in the image manipulation stage was relatively modest (N= 40), the effectiveness and robustness of their manipulations can be evaluated empirically using the subsequent large-scale rating experiment (crowdsourcing). As shown in Figure 4, images manually *uglified* by informed observers are predominantly assigned lower aesthetic values by naïve observers, whereas *beautified* images are shifted toward higher values relative to the original images. Importantly, *auto-uglified images* (generated by randomly reapplying recorded manipulation profiles) exhibit a distribution that is visually comparable to that of manually uglified images, suggesting that the manipulation strategies extend beyond the original set of edited images. Despite variability in individual preferences, this population-level separation confirms that the manipulation procedures reliably shift aesthetic judgments in the intended direction when evaluated by a large and independent sample of naïve observers (see Figure 6 for a more detailed analysis).

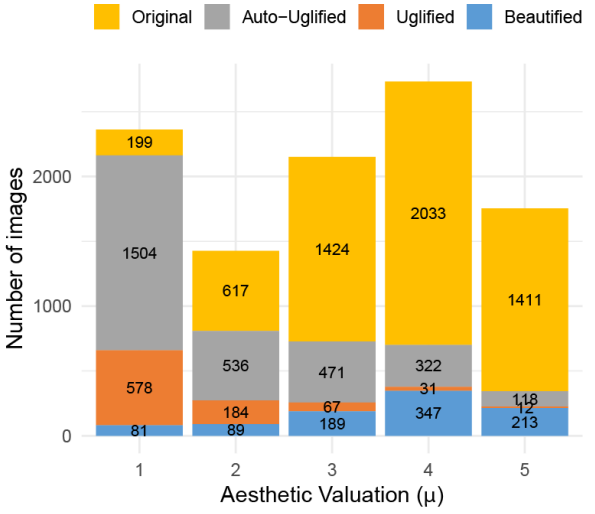


Figure 4: *Population-level validation of beautification and uglification procedures*. Distribution of mean aesthetic valuations (μ = 1–5) assigned by naïve observers to *beautified*, *uglified*, *auto-uglified*, and *unmodified* (original) images. The distribution of both uglified and auto-uglified images shows a bias towards low values, while the distribution of beautified images shows a bias towards high values, hinting at the effectiveness of the manipulation procedures at the population level (see Figure 6 and corresponding statistics).

Image low-level metrics for technical validation

To characterize the low-level visual properties of the MSC database and to validate the effects of achieving a more balanced distribution of aesthetic ratings through the beautification and uglification procedures, we used the aesthetics toolbox developed by Redies *et al*[44]. This open-access, user-friendly, Python-based toolkit computes a standardized set of quantitative image properties (QIPs) derived from luminance, color, spatial frequency, and structural characteristics, many of which have been shown in previous work[6-10, 12, 14, 37] to correlate with aesthetic

valuation across a wide range of visual stimuli. Importantly, these measures are largely independent of high-level semantic content and are therefore well suited for analyzing databases designed to minimize semantic confounds. To address the difficulty of comparing results across studies that use heterogeneous scripts and implementations, the toolbox integrates and harmonizes original code from multiple research groups into a single, unified platform. In total, it provides 43 QIPs grouped into conceptual categories, summarized in Table 2. We excluded image dimensions from the analysis due to their very limited dynamic range, which makes them unsuitable for the analyses presented below; no other metrics were pre-selected or omitted. The QIP toolbox was applied uniformly to all images in the MSC database, including original, beautified, and uglified versions, allowing us to assess whether the manipulation procedures produced systematic and measurable changes in low-level image statistics across images. By relying on a standardized and previously validated toolbox, we ensure that the reported image statistics are reproducible, comparable to earlier studies, and interpretable within the broader literature on visual aesthetics and natural image statistics.

Table 2: Summary of the most relevant metrics used in the aesthetics toolbox of Redies *et al* [44].

Category	Metric	Description
Image Dimensions	Image Size	Sum of image width and height
	Aspect Ratio	Ratio of width to height
Lightness, Contrast, and Complexity	RMS Contrast	Standard deviation in the L* channel (lightness)
	Lightness Entropy	Shannon entropy of the L* channel
	Complexity (HOG)	Mean gradient strength from HOG features
	Edge Density	Sum of edge responses from Gabor filters
Color Metrics	Channel Means & SDs	Mean and standard deviation of RGB, HSV, and L*a*b* channels
	Color Entropy	Entropy of hue values in HSV color space
Balance and Symmetry	Balance Score	Symmetry of luminance across multiple axes
	DCM (Deviation of Center of Mass)	Distance between perceptual and geometric center
	Mirror Symmetry	Average symmetry across vertical, horizontal, and diagonal axes
	CNN-based Symmetry	Symmetry from low-layer CNN features
Scale Invariance and Self-Similarity	Fourier Spectrum Slope	Slope of log-log Fourier amplitude spectrum
	Fourier Spectrum Sigma	Deviation from fitted power-law line in spectrum
	Fractal Dimension (2D)	Box-counting fractal dimension on binarized image
	Fractal Dimension (3D)	Box-counting on grayscale lightness values
	PHOG-based Self-Similarity	Similarity of HOG features at different pyramid levels
	CNN-based Self-Similarity	Similarity of CNN features across image regions
Feature Distribution and Entropy	Homogeneity	Entropy of black pixel distribution in subregions
	Anisotropy (HOG)	Directional imbalance in gradient orientations
	Edge-Orientation Entropy (EOE)	Entropy of edge orientations from Gabor filters
CNN-Based Variability	Sparseness and Variability	Variance in CNN feature activations

Bootstrapping and balanced sampling procedure

To assess whether differences between statistics (coefficient of variation, Jensen–Shannon divergence, Spearman rank correlation coefficient) were significant, we used a bootstrapping procedure with $\boldsymbol{N} = 10{,}000$ resamples with replacement[45]. To compute confidence intervals, we adjusted the confidence level according to the number of tests using a Bonferroni correction. Specifically, we used a confidence level of $1 - \alpha$ ($with\ \alpha = 0.05$) for single tests (coefficient of variation, Jensen–Shannon divergence[46]), and a level of $1 - \alpha/\boldsymbol{N_{metrics}}$ (with $\boldsymbol{N_{metrics}} = 45$) for multiple comparisons (Spearman rank correlation coefficient) involving the QIP toolbox metrics.

For the balanced sampling procedure, we first binned the aesthetic scores into 25 intervals and then identified the longest sequence of consecutive bins containing at least 50 data points each. Statistics (Spearman rank correlation coefficient) were computed using a bootstrap procedure in which each sample was generated by sampling with replacement within the selected bins. We generated $\boldsymbol{N} = 10{,}000$ bootstrap samples for each statistic.

Limitations, caveats and special considerations

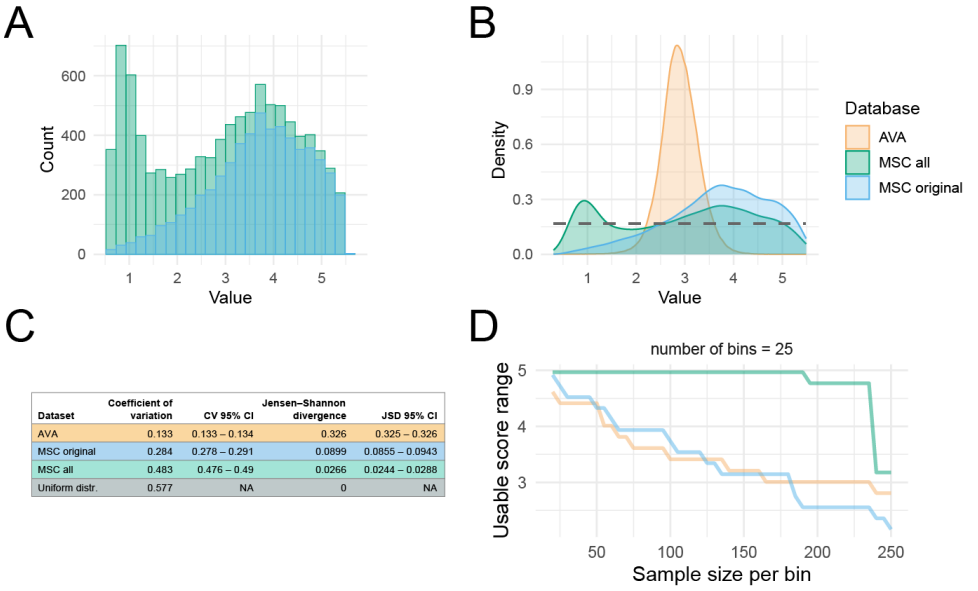
- **Image resolution.** The final sizes of the MSC images range from 293–800 pixels in height and 334–1200 pixels in width, which may be considered small by current standards. This choice was informed by the fact that the crowdsourcing company reported most observers viewed images on mobile phones. Based on the standard for 20/20 vision (Snellen), the human eye can resolve detail at about 60 pixels per degree (ppd). With a typical mobile viewing distance of 400 mm and screen dimensions of 150 x 70 mm, the maximum perceivable resolution is approximately 1274 x 600 pixels. This is consistent with the image sizes used in databases like AVA. Since we could not control the devices or internet speeds used by participants, we adopted these standard image sizes. While we recognize that image resolution can influence aesthetic judgments, the setup likely prevented participants from viewing images on larger screens where higher resolutions would be more relevant.
- **Residual Semantic Content.** Although the MSC database was designed to minimize semantic content, it is not entirely free of semantics. Images of natural objects can still be interpreted and attributed meaning by observers, depending on context. We acknowledge that it is nearly impossible to create an ecological database completely devoid of semantic content. To address this, we specifically avoided images depicting art or those likely to induce false recognition or subjective priming among naïve (crowdsourcing) participants.
- **Scope and Generalizability.** It is possible to argue that image aesthetics are inherently linked to semantics, and thus the MSC database represents only a subset of the full spectrum of images, with characteristics distinct from everyday visual experiences. Our intention was to create a controlled resource that isolates perceptual factors from semantic ones for a broad class of images. This reductionist approach is deliberate and parallels methods in neuroscience that separate complex processes to study individual components, such as the distinction between "what" (ventral) and "where" (dorsal) pathways[47], or specialized visual areas (e.g., V1, V2, V3)[48]. By minimizing and holding semantic factors constant, the MSC database provides a foundation for models that can later incorporate semantic complexity. While the database may not capture the full range of real-world aesthetics, this methodology enables a focused investigation of perceptual contributions to aesthetic decisions and supports future research that integrates semantic elements.
- **Potential Stereotypy in Uglifier Manipulations.** There is a possibility that the "Uglifier" tool could produce stereotyped results, such as consistently generating "ugly" images with unrealistic colors, low saturation, or similar features. We have addressed this concern in the last section (Figure 10,Figure 11 and Figure 12).
- **Limited participant demographic information.** Crowdsourcing participants were recruited through a popular online sports-related computer game, which likely resulted in a participant pool skewed toward younger users with a male preponderance. Their geographic distribution probably reflects the global popularity of the game, with users broadly distributed across Europe, Asia, and the Americas. Beyond indirect indicators provided by the crowdsourcing platform, we do not have access to detailed demographic information (e.g., age, gender, education, or cultural background). Participants were not screened for expertise in photography, art, or aesthetics; accordingly, the terms *naïve* and *non-expert* indicate that no expertise-related selection criteria were applied.

Database validation

To validate the full database (all)—which includes both original and systematically modified images derived from the unmodified set (original) through targeted “uglification” and “beautification” procedures—we addressed three key questions: (1) Does the database provide a more balanced coverage of the full spectrum of aesthetic ratings? (2) Are

the 'uglified' images indeed rated as less aesthetically pleasing? (3) Does this more balanced distribution alter statistical inferences typically drawn in computational aesthetics based on low-level image features? We report affirmative answers to each of these points below, systematically comparing our results with those obtained using the widely adopted AVA database[23]. We included a comparison with AVA/DPChallenge because these ratings are commonly interpreted in the literature as proxies for beauty judgments[7, 23, 49-51]. Our analysis therefore aligns with prevailing practice while highlighting the need for databases—such as MSC—in which the rated subjective dimension is explicitly specified.

More balanced representation of aesthetic valuations. Regression models are constrained to the range of data on which they are trained. When regressing aesthetic valuations against low-level image metrics—or indeed in any regression problem—if most observations cluster within a narrow segment of the possible range, the model cannot accurately estimate the true shape of the relationship or its confidence intervals beyond that segment[52, 53]; the model will also struggle to generalize to values outside the limited training range.



Dataset	Coefficient of variation	CV 95% CI	Jensen–Shannon divergence	JSD 95% CI
AVA	0.133	0.133 – 0.134	0.326	0.325 – 0.326
MSC original	0.284	0.278 – 0.291	0.0899	0.0855 – 0.0943
MSC all	0.483	0.476 – 0.49	0.0266	0.0244 – 0.0288
Uniform distr.	0.577	NA	0	NA

Figure 5: Comparison of the distributional balance of aesthetic scores across databases. A: Histogram of aesthetic scores for the original database (MSC original, light blue) and the enhanced database (MSC all, light green). B: Distributions (relative densities) of aesthetic scores for the three databases—MSC original, MSC all, colors as in A, and AVA, orange—compared to a uniform distribution (grey dotted line). C: Bootstrapped summary statistics quantifying deviation from uniformity for each database, using the coefficient of variation (left) and Jensen–Shannon divergence relative to the uniform distribution (right). D: Usable range of aesthetic scores for balanced regression analyses as a function of the required number of samples per bin (25 sampling bins). MSC all consistently supports a wider usable range than the other two databases. (Note that the range of aesthetic scores for AVA has been scaled to match that of the proposed database for comparison purposes.)

To evaluate the extent to which our database addresses any bias towards a subrange of the aesthetic valuations, we compared the distribution of aesthetic valuations in AVA, MSC original, and MSC all against a uniform distribution (Figure 5B), using two complementary metrics: the coefficient of variation—a normalized measure of dispersion relative to the mean—and the Jensen–Shannon divergence—a symmetric, bounded measure of similarity between probability distributions.

We found significant differences in divergence from the uniform distribution: AVA exhibited the greatest deviation, followed by MSC original, while MSC all was closest to uniform (Figure 5C). Importantly, the inclusion of 'uglified' images in MSC all brought the distribution significantly closer to uniformity compared to MSC original, a prerequisite to better supports regression analyses that rely on a broad and balanced sampling of the predictor variable.

To further illustrate this point, we quantified the range of aesthetic scores that could support evenly sampled regression analyses—an essential condition for reducing bias in estimating relationships between variables. We did

so by binning the score variable and identifying the widest consecutive range of bins that satisfied a minimum number of datapoints per bin, across various threshold values (see *Methods*). As shown in Figure 5D, MSC all supports a substantially broader range of aesthetic scores for a given sampling threshold, thereby enabling more robust and generalizable regression analyses. For instance, with 25 bins, MSC all allows more than 175 datapoints per bin while still covering the full aesthetic score range. In contrast, the same sampling density would restrict the usable score range to approximately 3 units for both AVA (despite AVA having a much larger number of datapoints overall) and MSC original.

Uglified images have indeed lower aesthetic valuations. To confirm the greater representation of images with lower aesthetic scores, we analyzed the effect of the uglification and auto-uglification processes. Figure 6 (panels A and C) shows the score of the (auto-)uglified images against the original images. Most images lie below the diagonal, indicating that the transformed images received lower aesthetic valuations. This was confirmed by two one-sided Mann–Whitney U tests, corrected for multiple comparisons using the Bonferroni method (adjusted $\alpha = .025$), indicating that scores in the original distribution were significantly higher than in the distribution of the uglified images ($W = 579{,}476$, $p < 2.2\text{e-}16$), and in the distribution of the auto-uglified images ($W = 1{,}905{,}702$, $p < 2.2\text{e-}16$) (Figure 6, panels B and D).

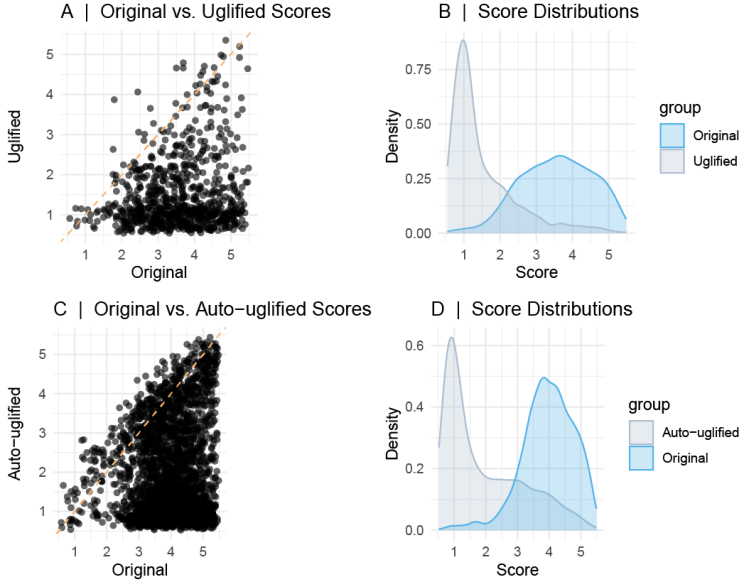


Figure 6: *Effect of the uglification process used to construct the database on aesthetic valuations*. A: Aesthetic scores of uglified images plotted against the scores of their original counterparts. B**:** Distribution of aesthetic scores for the uglified images (grey) compared to the distribution of the original images from which they were derived (blue). C and D are the counterparts of A and B, respectively for the process of auto-uglification. Darker regions in panels A and C indicate higher local point density.

We also verified that the uglifying process did not lead to stereotyped images. For this purpose, we compared the distributions of low-level metric values (see below) between uglified images and the original (hence, unmodified) images that received low valuations and found a very good match (see Figure 10,Figure 11 and Figure 12).

A more balanced representation modifies statistical inference. To further validate the proposed database, we demonstrated that its more balanced representation of aesthetic scores often leads to markedly different statistical inferences—directly impacting conclusions about the relationship between low-level image metrics and aesthetic valuation. This has significant implications for empirical aesthetics studies that do not explicitly control imbalances in aesthetic score distributions.

To explore this, we conducted two types of analyses: one based on standard correlation methods typically used in the literature and applied to the full database, and another using a balanced sampling approach across aesthetic scores—the dimension of interest. In the first analysis, we explored the correlations between aesthetic scores and a range of image metrics commonly used in empirical aesthetics across the three databases, without applying any

correction for imbalanced sampling—that is, by regressing aesthetic score directly onto low-level image metrics, as is typically done in the literature[7, 27, 51, 54-57]. *Figure* 7 displays the Spearman rank correlations between aesthetic score and each metric for the three databases.

Figure 7: Regression analysis across the three databases using all metrics from the QIP toolbox[44]. We computed Spearman rank correlations between aesthetic scores and 45 image metrics from the QIP toolbox for each database (AVA, orange; MSC original, light blue; MSC all, green) using a bootstrapping procedure. Each row corresponds to a different metric. The 95% confidence intervals for the correlation coefficients are shown in the color of the corresponding database. (For AVA, the intervals are extremely narrow due to its large sample size and are therefore not visible.) Colored markers on the right indicate statistically significant differences in correlation strength between databases: blue stars mark cases where the correlation for MSC original is significantly greater than for AVA; green stars indicate where MSC all shows a significantly greater correlation than AVA; and green circles show where MSC all has a significantly greater correlation than MSC original. These results support the conclusion that using a database with a more uniform distribution of aesthetic valuations enhances the ability to detect, and accurately quantify, relationships between low-level image metrics and aesthetic valuation.

The results clearly indicate a systematic underestimation of the relationship between image statistics and aesthetic judgments in databases with a less balanced distribution of aesthetic valuations. Specifically, correlations were significantly higher for MSC original than for AVA in 73.3% of the metrics (33 out of 45), for MSC all than for AVA in 80% of the metrics (36 out of 45), and for MSC all than for MSC original in 44.4% of the metrics (20 out of 45). Figure 8 illustrates the differences in correlations between aesthetic score and image metrics for AVA and MSC all, using three representative metrics. It highlights how a more even distribution of aesthetic scores (x-axis) allows clearer and more consistent trends to emerge—trends that remain obscured when scores are heavily concentrated around a narrow range.

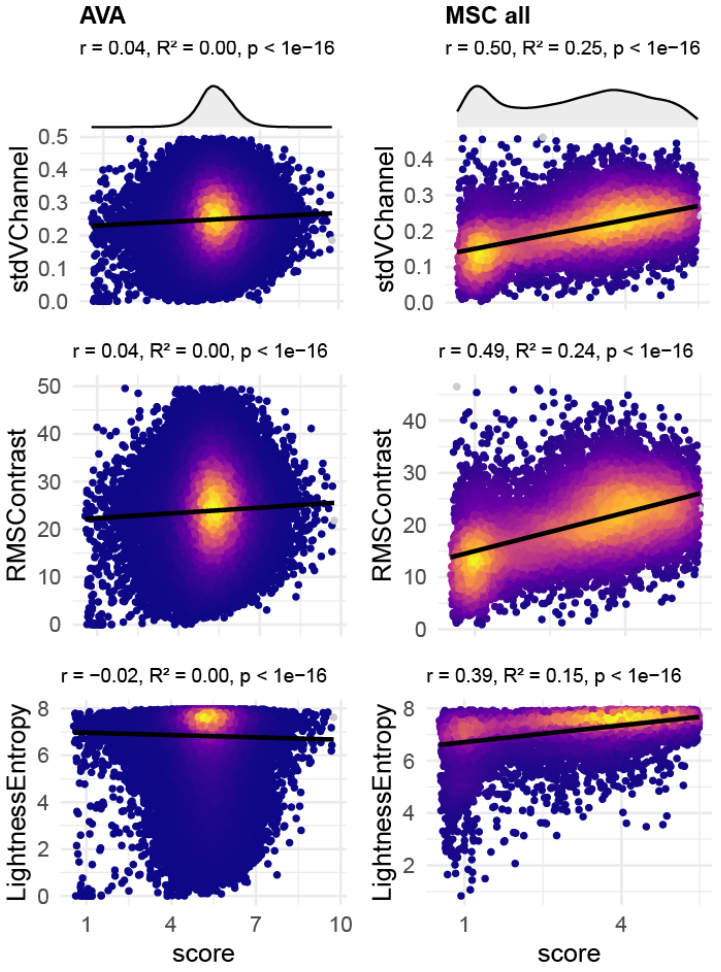


Figure 8: Comparison of estimated correlations between aesthetic scores and three image metrics from the QIP toolbox across two databases. Each row corresponds to one metric: standard deviation of the V channel (top), RMS contrast (middle), and lightness entropy (bottom). The left column shows results for the AVA database, and the right column for the proposed MSC all databases. Each panel displays the Spearman correlation coefficient, the R^2 value of a fitted linear regression (black line), and the Bonferroni-corrected p-value for the correlation. Point colors reflect local data density, with bright yellow indicating high density and dark blue indicating low density (AVA: N = 255,494; MSC all: N = 10,426). Grey histograms at the top of the top-row panels show the distribution of aesthetic scores for each database, allowing comparison with the uniform distribution (see also Figure 5B).

To further demonstrate the value of our enhanced database for studying the relationship between aesthetic judgments and low-level image metrics, we conducted a second analysis based on a balanced sampling of the dimension of interest, aesthetic scores (see Figure 9). Because MSC all includes previously underrepresented 'ugly' images, it yields a more uniform distribution of aesthetic scores—allowing for balanced sampling across the entire rating spectrum, a condition unmet by both the original database and the widely used AVA (see Figure 5C). This

broader coverage reveals markedly different—and in some cases reversed—relationships between image features and aesthetic value.

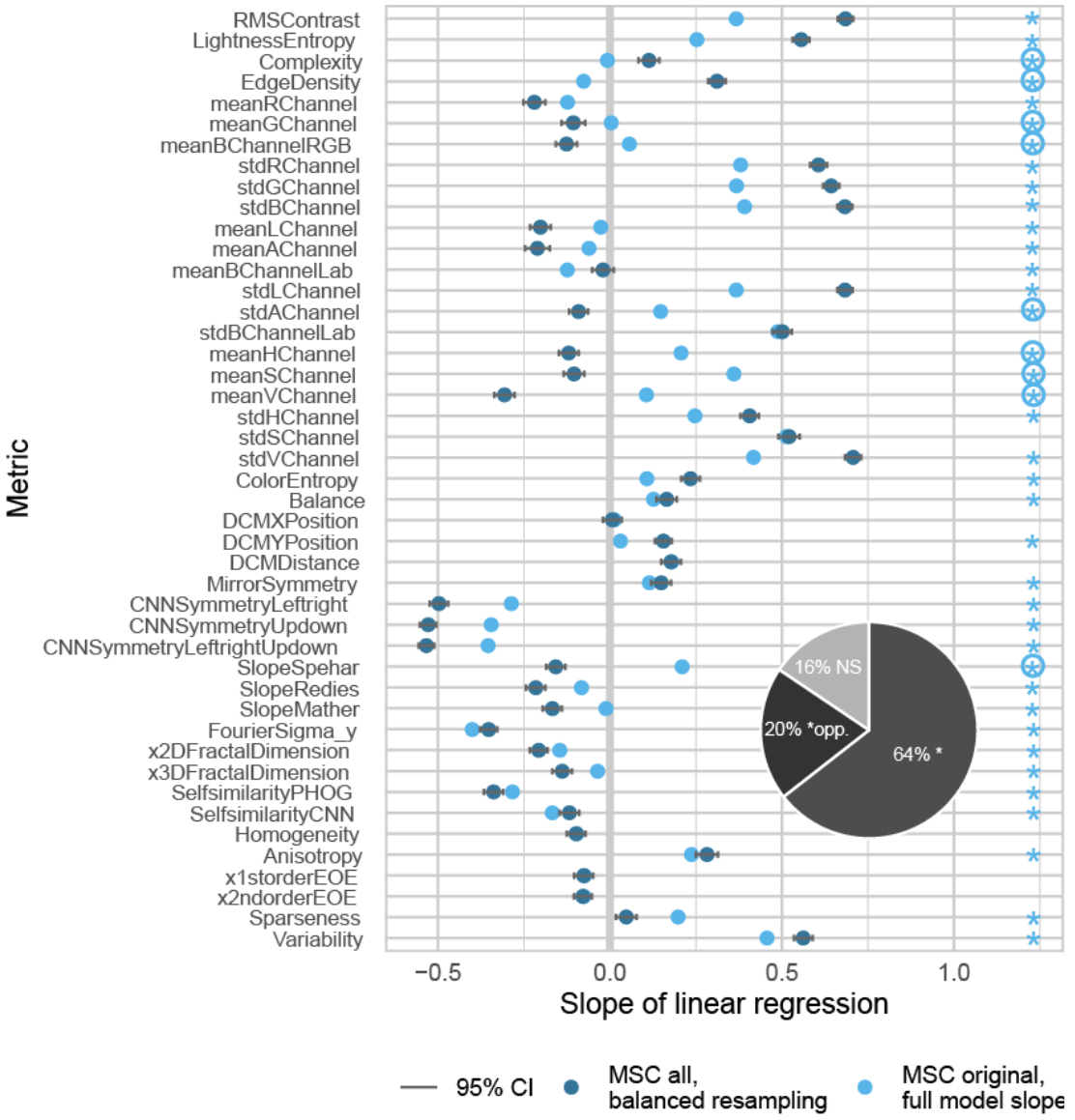


Figure 9: *Comparison of correlations between low-level image metrics and aesthetic valuations in the original database* (MSC original**)** *and the proposed extension with a more uniform distribution of aesthetic scores* (MSC all). As in Figure 7, each row corresponds to one metric from the QIP toolbox. Light blue points show the correlation coefficients for MSC original, while dark blue points represent the bootstrapped correlations for MSC all, with grey lines indicating 95% confidence intervals. A star on the right of a row marks a statistically significant difference in correlation between the two databases. A circle indicates that the correlation not only differs significantly but also has the opposite sign—implying that the two databases predict inverse relationships. This highlights the impact of using balanced sampling over a wide range of aesthetic scores in MSC all, compared to traditional analyses based on more skewed databases. The pie chart summarizes the results, showing the percentage of metrics with significant differences (mid and dark grey, four in five metrics), including the proportion with reversed correlations (dark grey, one in five metrics).

In this validation, we applied a bootstrap procedure to estimate Spearman correlations for each of the 45 image metrics in QIP toolbox and aesthetic valuations, comparing results from the proposed database (MSC all) and the original (MSC original). Each bootstrap sample was drawn using a balanced sampling strategy in MSC all (25 bins,

150 datapoints per bin). The differences were substantial: 84% of metrics (38 out of 45) showed significantly different correlation coefficients, and 20% (9 out of 45) even reversed in direction. In practical terms, this means that for four out of five metrics, the strength of association with aesthetic value changes significantly when using a more representative database—and for one in five, the interpretation of the relationship is fundamentally reversed.

Does the "Uglifier" produce stereotyped results?

To evaluate whether the image manipulation software used to enhance the original database ("Uglifier") generates stereotypically "ugly" results, we analyzed the low-level features of two groups: 1. the original (hence, unmodified) images rated with low value (aesthetic rating $\mu \leqslant 2.5$) by uninformed participants, and 2. the uglified images (Figure 10,Figure 11 and Figure 12 below). Visual inspection indicated that the distributions of the low-level features are similar across both categories. This was confirmed by two metrics. First, the overlap coefficient (OVL) between the two distributions, defined for two distributions f and g as $\mathrm{OVL}(f,g) = \int \min(f(x), g(x))dx$, was high, with mean = 0.78 and SD = 0.11; the maximum for this metric is 1 if the two distributions are identical. Second, the Jensen-Shannon divergence[46] was low, with mean = 0.088 and SD = 0.071, where identical distributions have a Jensen-Shannon divergence of 0. For instance, although informed participants using the Uglifier often exaggerated features such as color contrast to make images appear "uglier," similar exaggerations were also present in the color profiles of unmodified images that were classified as "ugly" by the crowd. Figure 10,Figure 11 and Figure 12 show the distributions of values for the 45 metrics in the QIP toolbox[44].

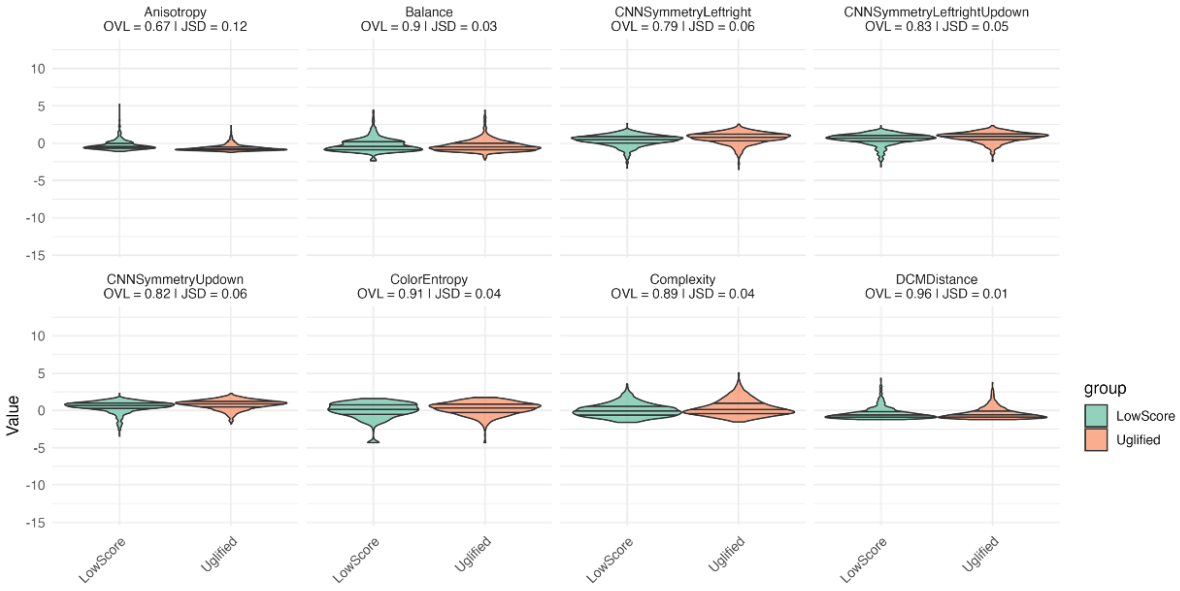


Figure 10: Comparison of the distributions of values for the low-level metrics in the QIP toolbox between original (unmodified) images with low aesthetic ratings and uglified images. The first 8 metrics are shown (see Figures S2-3 below for the other metrics). Each panel shows one metric, with the left distribution (green) corresponding to the subset of original images that received an average rating below 2.5 and the right distribution (orange) corresponding to the subset of uglified images. The overlap coefficient (OVL) and Jensen-Shannon divergence (JSD) are given at the top of each panel. Both metrics are bounded between 0 and 1. Identical distributions have an OVL of 1 and a JSD of 0.

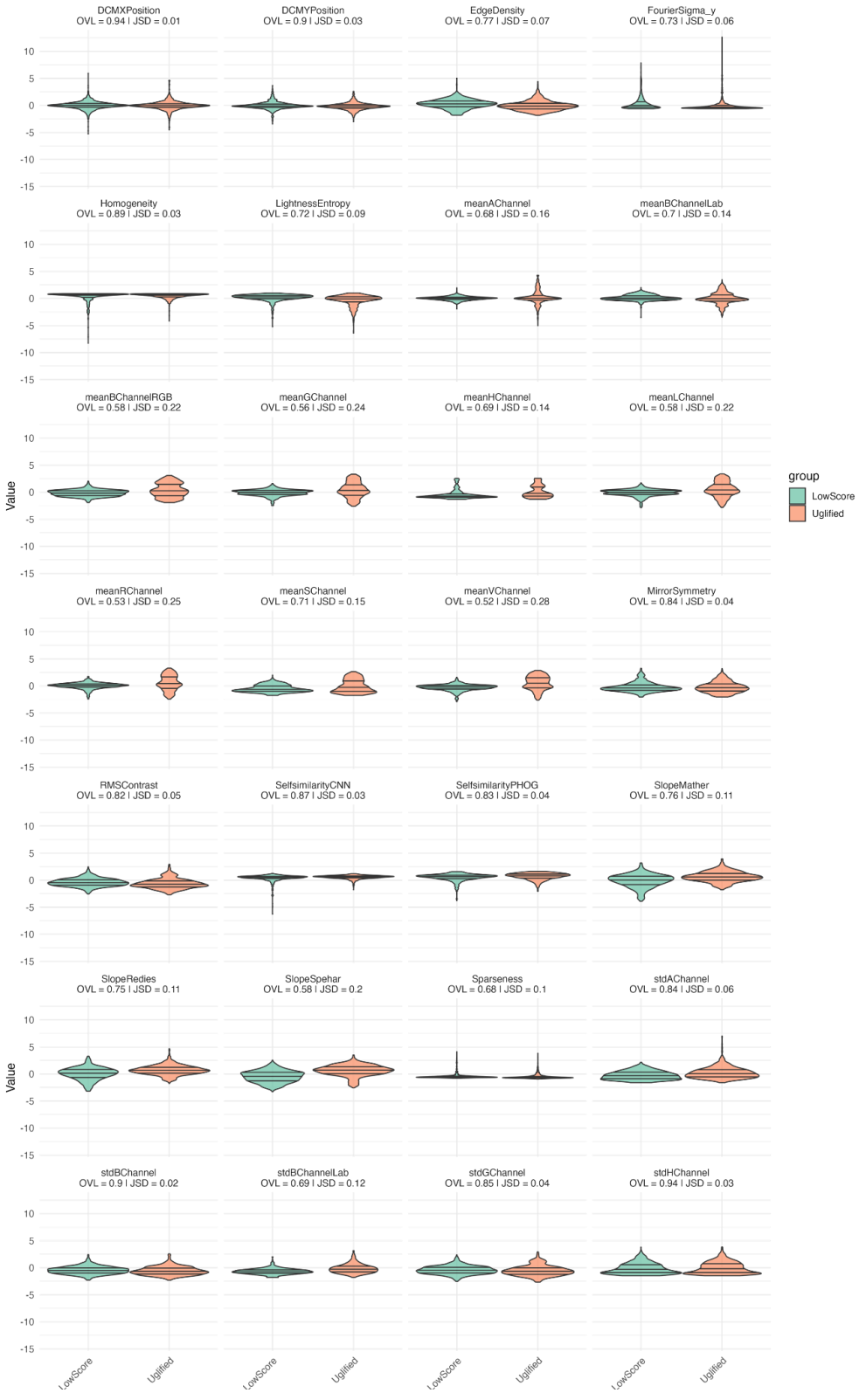


Figure 11: Same as Figure 10, metrics 9th to 36th of the QIP toolbox.

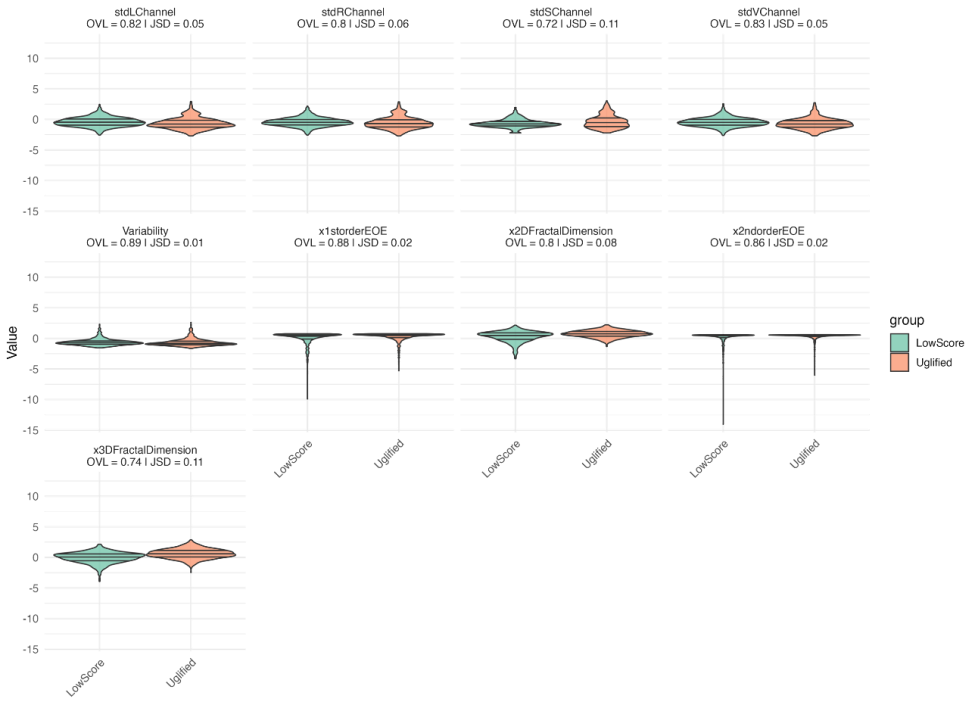


Figure 12: Same as Figure 10, metrics 37th to 45thof the QIP toolbox.

Code Availability

All QIP metrics were computed using the Python-based QIP toolbox of Redies *et al.*[44], while all subsequent analyses and visualizations of these metrics were performed using custom R[58] scripts, which are made publicly available. The second component of the OSF repository project, titled *Validation Software*, contains:

- All numerical tables and R code[58], executed in RStudio[59], for reproducing the statistical analyses and generating the figures in the main manuscript and Supplementary Material (MSC statistical validation and complete analysis of QIP metrics.zip).
- The code required to generate all corresponding figures (MSC_analysis_results_and_figures.zip).
- Automatically generated textual descriptions (captions) for images in the MSC Database produced using the GPT-4o[60] vision-language model (MSC dataset captions.zip)

Acknowledgements

C.A.P. and X.O. were supported by the Ministerio de Ciencia e Innovación, Gobierno de España MCIN/AEI/10.13039/501100011033: grants PID2020-118254RB-I00 and TED2021-132513B-I00, by the Agencia de Gestió d'Ajuts Univesitaris i de Recerca (AGAUR) through grant 2021-SGR-01470, and CERCA Programme / Generalitat de Catalunya. B.R. is supported by Grant PID2022-143257NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A Way of Making Europa". O.P. was funded by a Maria Zambrano Fellowship for the attraction of international talent for the requalification of the Spanish university system—NextGeneration EU (ALRC). A.J. was funded by the FI fellowship AGAUR 2022 FI-SDUR 00248 (Secretaria d'Universitats i Recerca, Generalitat de Catalunya, and Fons Social Europeu).

Author contributions

C.A.P., O.P., and B.R. conceived the study. C.A.P., O.P., and A.J. developed the codebase and conducted the experiments. O.P., C.A.P., and A.J. performed the statistical analyses. C.A.P. and O.P. wrote the initial draft of the manuscript. Project supervision was carried out by C.A.P. and X.O. The manuscript was revised by O.P., C.A.P, B.R. and X.O. Data curation was performed by C.A.P. and O.P.

Competing Interests

The authors declare no competing interests.

References

1. Chatterjee, A. & Vartanian, O. Neuroaesthetics. Trends Cogn Sci 18, 370-5 (2014).
2. Vessel, E. A., Ishizu, T. & Bignardi, G. in The Routledge International Handbook of Neuroaesthetics (eds. Skov, M. & Nadal, M.) 102-131 (Routledge, London, 2022).
3. Carandini, M. et al. Do We Know What the Early Visual System Does? The Journal of Neuroscience 25, 10577-10597 (2005).
4. Reber, R., Winkielman, P. & Schwarz, N. Effects of Perceptual Fluency on Affective Judgments. Psychological Science 9, 45-48 (1998).
5. Goodale, M. A. & Milner, A. D. Separate visual pathways for perception and action. Trends in Neurosciences 15, 20-25 (1992).
6. Farzanfar, D. & Walther, D. B. Changing What You Like: Modifying Contour Properties Shifts Aesthetic Valuations of Scenes. Psychological Science 34, 1101-1120 (2023).
7. Iigaya, K., Yi, S., Wahle, I. A., Tanwisuth, K. & O'Doherty, J. P. Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features. Nature Human Behaviour 5, 743-755 (2021).
8. McManus, I. C. The aesthetics of simple figures. British Journal of Psychology 71, 505-524 (1980).
9. Mallon, B., Redies, C. & Hayn-Leichsenring, G. Beauty in abstract paintings: perceptual contrast and statistical properties. Frontiers in Human Neuroscience 8 (2014).
10. Spehar, B., Clifford, C. W. G., Newell, B. R. & Taylor, R. P. Universal aesthetic of fractals. Computers & Graphics-UK 27, 813-820 (2003).
11. Tinio, P. P. L. & Leder, H. Natural scenes are indeed preferred, but image quality might have the last word. Psychology of Aesthetics, Creativity, and the Arts 3, 52-56 (2009).
12. Bertamini, M., Rampone, G., Makin, A. D. J. & Jessop, A. Symmetry preference in shapes, faces, flowers and landscapes. PeerJ 7, e7078 (2019).
13. Jacobsen, T. & Höfel, L. Aesthetic Judgments of Novel Graphic Patterns: Analyses of Individual Judgments. Perceptual and Motor Skills 95, 755-766 (2002).
14. Rhodes, G. The Evolutionary Psychology of Facial Beauty. Annual Review of Psychology 57, 199-226 (2006).
15. Bar, M. & Neta, M. Humans Prefer Curved Visual Objects. Psychological Science 17, 645-648 (2006).
16. Bertamini, M., Palumbo, L., Gheorghes, T. N. & Galatsidas, M. Do observers like curvature or do they dislike angularity? British Journal of Psychology 107, 154-178 (2016).
17. Clemente, A., Penacchio, O., Vila-Vidal, M., Pepperell, R. & Ruta, N. Explaining the curvature effect: Perceptual and hedonic evaluations of visual contour. Psychology of Aesthetics, Creativity, and the Arts (2023).
18. Vartanian, O. et al. Impact of contour on aesthetic judgments and approach-avoidance decisions in architecture. Proceedings of the National Academy of Sciences 110, 10446-10453 (2013).
19. Geller, H. A., Bartho, R., Thömmes, K. & Redies, C. Statistical image properties predict aesthetic ratings in abstract paintings created by neural style transfer. Frontiers in Neuroscience Volume 16 - 2022 (2022).
20. DPChallenge Community. (Challenging Technologies, LLC.) http://www.dpchallenge.com/ (2016). Accessed 2017-04-19

21. Flickr Community. (SmugMug, Inc.) https://www.flickr.com/ (2003). Accessed 07/09/2016
22. Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A. & Jain, R. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 1349-1380 (2000).
23. Murray, N., Marchesotti, L. & Perronnin, F. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR2012) 2408-2415 (2012).
24. Joshi, D. et al. Aesthetics and Emotions in Images A computational perspective. IEEE Signal Processing Magazine 28, 94-115 (2011).
25. Parraga, C. A. et al. The Minimum Semantic Content (MSC) image dataset, software files, and supplemental material (OSF -Open Science Framework) https://doi.org/10.17605/OSF.IO/ZGSVJ (2025).
26. Parraga, C. A., Muñoz González, M., Otazu, X. & Penacchio, O. in European Conference on Visual Perception (ECVP2022) 139-139 (Nijmegen, The Netherlands, 2022).
27. Parraga, C. A., Penacchio, O., Muňoz Gonzalez, M., Raducanu, B. & Otazu, X. Aesthetics Without Semantics. arXiv:2505.05331v2 (2025).
28. Datta, R., Joshi, D., Li, J. & Wang, J. Z. Studying aesthetics in photographic images using a computational approach. European Conference on Computer Vision (ECCV2006) 3953, 288-301 (2006).
29. Brachmann, A. & Redies, C. Computational and Experimental Approaches to Visual Aesthetics. Frontiers in Computational Neuroscience 11 (2017).
30. Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. Annual Review of Neuroscience 24, 1193-1216 (2001).
31. Field, D. J. Relations between the statistics of natural scenes and the response properties of cortical cells. Journal of the Optical Society of America A 4, 2379-2394 (1987).
32. Ruderman, D. L. & Bialek, W. Statistics of natural images: Scaling in the woods. Physical Review Letters 73, 814-817 (1994).
33. Parraga, C. A., Troscianko, T. & Tolhurst, D. J. The human visual system is optimised for processing the spatial information in natural visual images. Current Biology 10, 35-38 (2000).
34. Parraga, C. A., Troscianko, T. & Tolhurst, D. J. Spatiochromatic properties of natural images and human vision. Current Biology 12, 483-487 (2002).
35. PdPhoto Community. http://pdphoto.org/ (2003). Accessed 07/09/2016
36. Photos Public Domain Community. Free stock photos, textures, images, pictures & clipart for any use including commercial. http://www.photos-public-domain.com/ (2010). Accessed 07/09/2016
37. McManus, I. C. et al. The Psychometrics of Photographic Cropping: The Influence of Colour, Meaning, and Expertise. Perception 40, 332-357 (2011).
38. Laeng, B., Øvervoll, M. & Ala-Pettersen, E. A. Original art paintings are chosen over their "color-rotated" versions because of changed color contrast. Perception 54, 780-814 (2025).
39. Nakauchi, S. & Tamura, H. Regularity of colour statistics in explaining colour composition preferences in art paintings. Scientific Reports 12, 14585 (2022).
40. Nascimento, S. M. C. et al. The colors of paintings and viewers' preferences. Vision Research 130, 76-84 (2017).
41. Pitie, F., Kokaram, A. C. & Dahyot, R. Automated colour grading using colour distribution transfer. Computer Vision and Image Understanding 107, 123-137 (2007).
42. Ryabov, A. (MATLAB Central File Exchange, 2022).
43. The Math Works Inc. Computer Software (The Math Works, Inc., 2022).
44. Redies, C. et al. A toolbox for calculating quantitative image properties in aesthetics research. Behavior Research Methods 57, 117 (2025).
45. Efron, B. & Tibshirani, R. J. An Introduction to the Bootstrap (Chapman and Hall/CRC, 1994).
46. Endres, D. M. & Schindelin, J. E. A new metric for probability distributions. IEEE Transactions on Information Theory 49, 1858-1860 (2003).
47. Ungerleider, L. G. & Haxby, J. V. 'What' and 'where' in the human brain. Current Opinion in Neurobiology 4, 157-165 (1994).
48. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1, 1-47 (1991).

49. Lu, X., Zhe, L., Hailin, J., Jianchao, Y. & James, Z. W. in Proceedings of the 22nd ACM international conference on Multimedia (ACM, Orlando, Florida, USA, 2014).
50. Ma, S., Liu, J. & Chen, C. W. in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 722-731 (2017).
51. Kong, S., Shen, X., Lin, Z., Mech, R. & Fowlkes, C. in ECCV 2016 (eds. Leibe, B., Matas, J., Sebe, N. & Welling, M.) 662-679 (Springer International Publishing, Amsterdam, The Netherlands, 2016).
52. Schober, P., Boer, C. & Schwarte, L. A. Correlation Coefficients: Appropriate Use and Interpretation. Anesth Analg 126, 1763-1768 (2018).
53. Janse, R. J. et al. Conducting correlation analysis: important limitations and pitfalls. Clinical Kidney Journal 14, 2332-2337 (2021).
54. Penacchio, O., Otazu, X., Wilkins, A. J. & Haigh, S. M. A mechanistic account of visual discomfort. Frontiers in Neuroscience 17 (2023).
55. Penacchio, O., Haigh, S. M., Ross, X., Ferguson, R. & Wilkins, A. J. Visual Discomfort and Variations in Chromaticity in Art and Nature. Frontiers in Neuroscience Volume 15 - 2021 (2021).
56. Bartho, R., Thoemmes, K. & Redies, C. Predicting beauty, liking, and aesthetic quality: A comparative analysis of image databases for visual aesthetics research. arXiv:2307.00984 (2023).
57. Zeng, H., Cao, Z., Zhang, L. & Bovik, A. C. A Unified Probabilistic Formulation of Image Aesthetic Assessment. IEEE Transactions on Image Processing 29, 1548-1561 (2020).
58. *R* Core Team. (R Foundation for Statistical Computing, Vienna, Austria, 2025).
59. Posit Team. (Posit Software PBC, Boston, MA, 2025).
60. Hurst, A. et al. GPT-4o System Card. arXiv:2410.21276 (2024).