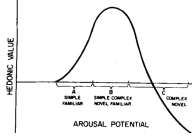


Exploring Aesthetics Through Surprise Gained from Diffusion

JOSHUA (HYUN JO) JANG, University of Toronto, Canada

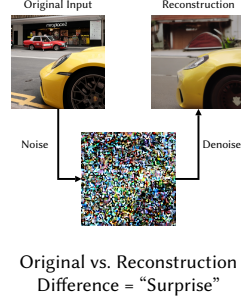
Classical Theory of Aesthetics (Berlyne, 1970)



“Hedonic (aesthetic) value is maximised at a moderate amount of arousal potential (surprise)”

Can we test this hypothesis?

Diffusion-Based Surprise Measurement



Surprise (Our Measurement) vs. Human-Evaluated Aesthetics Score

Original Input			
vs.	Small Difference	Medium Difference	Large Difference
Diffusion Recon.			
Human Score	Low (2.5)	High (7.6)	Low (3.1)

Findings

We see evidence for Berlyne’s theory, albeit modestly and under specific definitions of surprise.

Fig. 1. An influential classical theory from 1970 suggests that a moderate amount of *surprise* gained from a visual stimulus (image) results in the highest perceived aesthetic value. How can we quantify surprise and empirically test this surprise-aesthetics relationship? We answer these questions by defining surprise as the difference between an original image and its *diffusion* reconstruction. Inspired by image anomaly detection methods, we believe that diffusion models can tell us how much an image is “unexpected” or “surprising”. Under certain conditions and when comparing with human-evaluated aesthetic scores, this measure of surprise reveals similar patterns to the classical theory.

Quantifying why certain images appear aesthetically pleasing remains a longstanding challenge in computer vision and cognitive science. While modern neural models achieve high predictive accuracy, they offer limited explanation of the cognitive factors that shape human aesthetic evaluations. We revisit psychologist Daniel Berlyne’s classical theory, which proposes that aesthetic value peaks at a moderate level of surprise (inverted-U pattern), and introduce a diffusion-based method for formalising and measuring this variable. Our method is inspired by anomaly detection techniques, where we add noise to an input image, reconstruct it using a diffusion model, and measure surprise as the difference between the original and reconstructed images across multiple noise levels, spatial scales, and feature definitions (SSIM, LPIPS, DreamSim). Evaluated on 10K images from the AVA Aesthetics Dataset, our measurements reveal that specific surprise formulations, specifically high-level semantic differences between an original image and its reconstruction, exhibit the inverted-U pattern predicted by Berlyne. Although the correlation is modest, our findings demonstrate that modern generative models can be used to explore and ultimately contribute to explaining human aesthetic evaluation.

Additional Key Words and Phrases: Image Aesthetics, Diffusion Models, Surprise, Cognitive Science

1 Introduction

Modern computational photography has achieved remarkable improvements in objective image quality in terms of noise, resolution, and distortion. Yet one of its significant goals is to produce images that are pleasing and meaningful to the viewer. After all, photographs are taken not only for scene measurement but also to record and share moments we want to remember. Accordingly, there have been attempts to algorithmically enhance aesthetic aspects such as colour [7] and composition [10]. However, unlike objective signal quality, where metrics like PSNR offer clear evaluation, the

notion of a “pleasing image” remains ambiguous and subjective. Many works rely on user studies, which are difficult to reproduce and offer limited transparency into the mechanisms behind them. A natural direction is to seek a cognitive metric that models the human aesthetic evaluation process. Early theories from cognitive science and philosophy, including those of Birkhoff [4] and Berlyne [2, 3], proposed relationships between order, complexity, novelty, and beauty. While conceptually compelling, these formulations were often vague (e.g., “beauty is proportional to order over complexity”) and unsuitable for rule-based algorithms. In contrast, computer vision methods largely treat aesthetics evaluation as a supervised learning problem. State-of-the-art vision models trained on datasets such as AVA [11] can predict human scores but typically operate as black boxes with limited interpretability. As a result, it is difficult to identify which cognitive components correspond to their outputs. To narrow this gap, we focus on a single cognitive variable of *surprise* and subject it to empirical evaluation. This idea is inspired by Berlyne’s classical theory, which suggests that maximum hedonic (aesthetic) value is maximised at a moderate amount of arousal potential (surprise) [3]. In other words, aesthetic value and surprise form an inverted-U relationship. Although influential, studies that attempt to formalise this notion of surprise have relied primarily on low-level statistics such as Shannon entropy [6, 12], which do not fully capture perceptual surprise in natural images. We revisit Berlyne’s hypothesis by leveraging modern *image diffusion models*, which are trained on a vast quantity of natural images and thus may implicitly encode expectations about “not surprising” images. Specifically, we adapt techniques from diffusion-based anomaly detection [1, 16] to compute a concrete surprise score. Our pipeline proceeds as follows. (1) We add noise to an input image and reconstruct a clean version using Stable Diffusion, a latent diffusion model trained

on a large natural image dataset. Its reconstructions approximate what is statistically “expected” under the learnt distribution. (2) We compare the original image and its reconstructions across multiple noise levels, scales, and feature representations. (3) Using the AVA Aesthetics Dataset as evaluation, we test whether these comparative results exhibit the hypothesised inverted-U relationship with human aesthetic scores. Our objective is not to develop a state-of-the-art aesthetics predictor but to provide empirical evidence toward an understanding of human aesthetics evaluation, grounded in Berlyne’s classical theory.

2 Related Work

2.1 Image Aesthetics Assessment

Recent work in computer vision has enabled the prediction of aesthetic quality, or image aesthetics assessment (IAA), directly from images, using convolutional to transformer-based models. Early approaches such as NIMA [15] used models such as VGG16 and MobileNet to predict aesthetic quality (human rating) distributions from large-scale IAA datasets, achieving up to 81% accuracy. Building on this, MUSIQ [8] introduced a vision transformer (ViT) architecture capable of handling multi-scale image representations. It was able to capture both local texture and global composition features, leading to significant accuracy improvements over CNN-based methods. More recently, AesCLIP [14] and VILA [9] make use of large-scale transformers, aligning image features with human aesthetic concepts encoded in text. These methods take advantage of implicit, human-like semantic information and have set new state-of-the-art results across multiple aesthetic benchmarks. However, these models are optimised primarily for predictive accuracy on benchmark datasets, not necessarily for explaining the cognitive factors behind human aesthetic evaluation. Their architectures are mostly opaque, making it difficult to disentangle specific factors that impact their predictions.

2.2 Classical Theories of Aesthetics

On the other side, researchers in cognitive science, psychology, and philosophy have proposed individual theories since the early 20th century to explain how humans measure aesthetic value from visual stimuli. A notable early example that inspired many subsequent theories is Birkhoff’s 1933 formulation, defined as:

$$M = \frac{O}{C},$$

where M, O, and C represent aesthetic value, order, and complexity, respectively. In Birkhoff’s description, “the complexity is roughly the number of elements that the image consists of, and the order is a measure for the number of regularities found in the image” [4]. Both terms are somewhat ambiguous, and later work attempted to formalise order using Shannon entropy and complexity using Kolmogorov complexity, with analyses performed on well-known paintings. A later generation of theories introduced the notion of optimal novelty, most notably through the work of psychologist Daniel Berlyne [2, 3]. Berlyne hypothesised that aesthetic value follows an inverted-U relationship with respect to what he termed *collative variables*: a broad category involving “novelty, surprisingness, incongruity, complexity, variability, and puzzlingness” [2]. Figure 1

shows the original curve proposed by Berlyne. Attempts to formalise Berlyne’s collative variables, particularly *novelty* (*surprise*), have predominantly relied on Shannon entropy [6, 12]. While this is explicit and computationally straightforward, it introduces several limitations. First, the probability of any given pixel is determined solely from the image itself, without incorporating any prior over natural images. Second, entropy-based measures cannot capture complex visual patterns or higher-level semantics. Consequently, their alignment with intuitive, real-world notions of surprise remains limited.

2.3 Diffusion-Based Image Anomaly Detection

Modern generative models have introduced new ways to quantify surprise. In this work, we interpret surprise as the magnitude of anomalous regions in an image relative to an expected, “normal”, “expected” baseline. This perspective aligns closely with diffusion-based anomaly detection methods, which leverage the reconstruction behaviour of denoising diffusion models trained exclusively on normal data [1, 16]. These approaches add noise to an input image and generate a reconstruction through a denoising diffusion probabilistic modelling (DDPM) method. During this process, regions that deviate from the learnt distribution tend to be overwritten with their statistically normal counterparts. To measure the discrepancy between the input and its diffusion reconstruction, existing methods adopt different comparison strategies. [1] computes a pixel-wise L1 difference followed by median smoothing to obtain a binary anomaly mask, whereas [16] measures cosine distance ResNet-34 feature vectors to construct a spatial anomaly map. Diffusion-based anomaly detection has demonstrated strong performance across both medical imaging domains, such as brain MRI [1], and natural images, including the VisA visual anomaly dataset [16]. Given these results, we treat diffusion reconstruction error as a viable and general mechanism for estimating image surprise. We expand on this idea in our method, where we interpret diffusion-detected anomaly as a proxy for surprise in evaluating their relationship with human aesthetic evaluations.

3 Method

Our primary objective is to empirically investigate Berlyne’s theorised inverted-U relationship between surprise and real human aesthetic evaluations. However, Berlyne’s theory does not precisely define which *kinds* of surprise contribute to aesthetic value. Therefore, while we use a diffusion model as our main backbone, we focus on identifying certain conditions and formulations of surprise that follow the inverted-U pattern from a variety of options. Specifically, for a given input image, we produce 27 individual *surprise scores*, each representing a different calculation of surprise, across three noise levels, three spatial scales, and three difference metrics. As shown in Figure 2, the end-to-end pipeline of using diffusion for reconstruction, measuring surprise scores, and relating them to real human aesthetic scores is a three-stage process.

3.1 Step 1: Generating Diffusion Reconstructions

We use a pretrained image diffusion model as the backbone for generating denoised reconstructions. Specifically, we adopt Stable

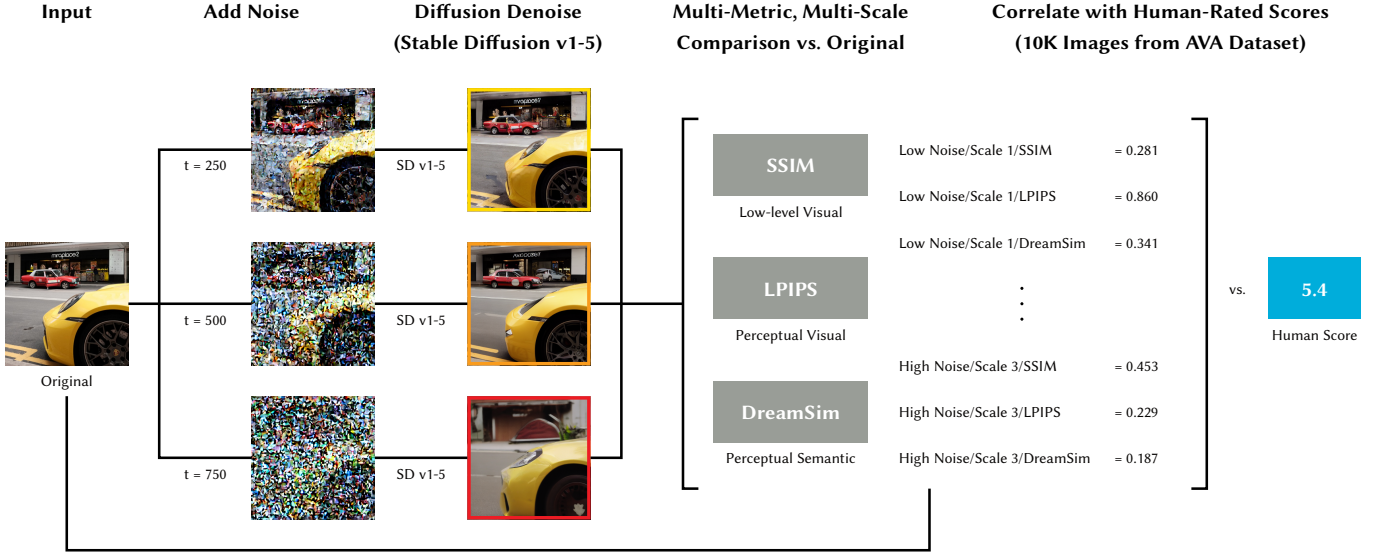


Fig. 2. Our method is composed of three major stages: (1) From an input image, we generate three versions with progressively stronger noise. Then, we reconstruct the original using Stable Diffusion v1-5, a latent diffusion model (LDM). This results in noise-free, *diffusion expected* alternatives of the original. (2) For each original-reconstruction pair, we measure the difference between the two images across multiple spatial scales (on a Gaussian pyramid) and metrics (low to high-level, SSIM/LPIPS/DreamSim). Each noise/scale/metric combination produces a mean scalar *surprise score*, and there are 27 total combinations. (3) Finally, after computing all surprise scores for all images in our target dataset (10K subset of AVA), we evaluate which surprise score follows the hypothesised inverted-U relationship with human-evaluated aesthetic scores.

Diffusion v1-5, a latent diffusion model (LDM) trained on approximately 600M image-caption pairs from the LAION-5B dataset [13]. Because LAION-5B consists of commonly encountered photographs, we assume that, at this scale, the model implicitly encodes a “non-surprising” prototype of typical subjects, compositions, and colours. Prior work suggests that adding a single level of static noise is insufficient for capturing anomalies across different spatial scales [16]. Following this finding, we apply three noise levels corresponding to $t=250$, 500, and 750 steps of the forward diffusion process (out of the 1000 steps used during SD v1-5 training). Each level destroys progressively more information from the input. We then perform unconditional denoising using the pretrained UNet of SD v1-5. This is achieved via passing a blank string onto the SD text encoder. We run only 50 reverse-diffusion steps rather than the full 250/500/750 to balance reconstruction fidelity and inference speed, determined via manual tuning. For each input image, this procedure produces three reconstructions, each representing a different degree of information loss.

3.2 Step 2: Original vs. Reconstruction Comparisons

3.2.1 Spatial Scale. Next, we compare each original image to its reconstructions across multiple spatial scales using a Gaussian pyramid: Scale 0 (512x512), Scale 1 (256x256), and Scale 2 (128x128). Rather than aggregating these scales, we treat them independently to allow fine-grained analysis of surprise.

3.2.2 Difference Metric. Prior diffusion-based anomaly detection methods [1, 16] rely on purely visual comparison features, from L1 pixel differences to CNN embeddings. They demonstrate strong

performance in their respective tasks in anomaly detection. To expand on this further, we evaluate three complementary metrics, each capturing differences at a distinct representational level.

- SSIM [17] serves as our lowest-level metric, measuring luminance, contrast, and local structural differences.
- LPIPS [18], which computes distances between AlexNet or VGG-16 features (similar to [16]), captures mid-level perceptual differences.
- DreamSim (CLIP backbone) [5] emphasises high-level, semantic differences between the two images.

SSIM and LPIPS natively provide spatially localised difference maps, whereas DreamSim outputs only a global scalar. To allow for consistent spatial interpretability during experimentation, we compute all differences on a sliding-window grid of 64x64-sized patches and then average the results into a final scalar score.

3.3 Step 3: Surprise Scores vs. Human Aesthetic Scores Analysis via AVA

The AVA Aesthetics dataset [11] contains approximately 255K photographs sourced from the DPChallenge online photography platform. Each image is annotated with roughly 210 user ratings on a 1-10 aesthetic quality scale, where the mean score, score histogram, and variance are provided. This dataset serves as our primary benchmark. For every image, we compare the 27 surprise scores produced by the previous step against their human-evaluated mean aesthetic score. We randomly sample 10K image-mean score pairs and compute all surprise metrics for this subset. All processing was conducted on a single NVIDIA RTX 5060 Ti 16GB GPU over 40 hours.

4 Results & Evaluation

4.1 Reconstructions and Detected Surprise



Fig. 3. Visualisation of an original input image, versions with noise added, and their reconstruction using SD v1-5. We observe that as the noise added becomes stronger, the reconstruction deviates further from the original.



Fig. 4. Difference heatmaps across spatial scales for the low-level SSIM metric. Notice that highlighted regions are generally blob-like, and subtle visual differences (e.g., car wheel) are detected.

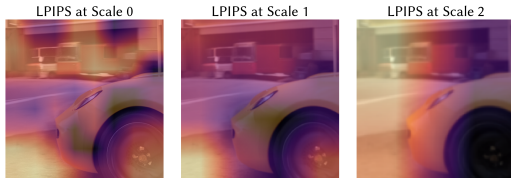


Fig. 5. Difference heatmaps across spatial scales for the mid-level LPIPS metric. It shows slightly clearer segmentation over SSIM but is overall similar.

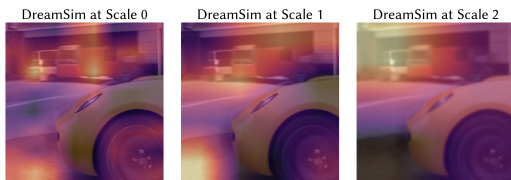


Fig. 6. Difference heatmaps across spatial scales for the high-level DreamSim metric. Visual differences that do not affect semantics are mostly ignored (e.g., car tyre). We also observe much clearer segmentation of regions.

Figures 4, 5, and 6 visualise an example comparison between an original image and a reconstruction at noise level $t=750$ (example reconstructions can be seen in Figure 3). The corresponding surprise measurements are depicted as heatmaps, separated by spatial scale and metric type. As expected, increasing the scale in the Gaussian pyramid leads to more coarse, blob-like highlighted regions, while scale 0 is sharper but noisier. Similarly, as the metric increases in abstraction, the heatmaps display reduced noise and more coherent contrast. At scale 0, SSIM tends to produce noisy activations, where even minor pixel-level discrepancies result in large differences. LPIPS is similar, though its highlighted areas lack clear separation. DreamSim produces more segmented heatmaps. For instance, the red taxi and road marking regions are emphasised, while semantically less meaningful differences such as the road itself are not. It also ignores small visual differences that do not change semantics, such as the car wheel. We consistently observed that DreamSim’s highlighted regions most closely matched those humans would intuitively identify as different. These heatmaps serve only as qualitative illustrations. For quantitative analysis, we use the mean difference as the final surprise score from each (noise level, scale, metric) combination, shown as labels in Figure 2.

4.2 Comparisons with Human Aesthetic Scores

Our analysis is grounded in Berlyne’s prediction that aesthetic value should follow an inverted U-shaped relationship with surprise. Accordingly, we do not expect linear or monotonic correlations, but a peak at moderate surprise levels. Because human scores are inherently noisy and influenced by many confounding variables, and because we generate only a single reconstruction per image without averaging, we use a quantile-binning strategy to reveal trends. We divide the 10K AVA mean scores into three quantile bins and compute the mean surprise within each bin for all 27 surprise score variants, producing 27 bar plots. To quantify the strength of the inverted-U pattern, we compute a simple U-metric. Let x_i denote the human aesthetic score for image i and y_i the corresponding surprise score. Define the quantile threshold $Q_{1/3}$ and $Q_{2/3}$ of the empirical distribution of x_i .

We partition the dataset into three bins:

$$\mathcal{L} = \{ i \mid x_i \leq Q_{1/3} \},$$

$$\mathcal{M} = \{ i \mid Q_{1/3} < x_i \leq Q_{2/3} \},$$

$$\mathcal{H} = \{ i \mid x_i > Q_{2/3} \}.$$

For each bin, we compute the mean surprise:

$$\mu_L = \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} y_i, \quad \mu_M = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} y_i, \quad \mu_H = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} y_i.$$

We define our U-metric as:

$$U = \mu_M - \frac{1}{2}(\mu_L + \mu_H).$$

We additionally impose the constraint: $\mu_M \geq \mu_L$ and $\mu_M \geq \mu_H$.

Statistical significance is checked using a two-sided permutation test by randomly shuffling AVA scores and recomputing the U-metric 500 times, then computing the probability of our original U-metric. A threshold probability of 0.05 is used.

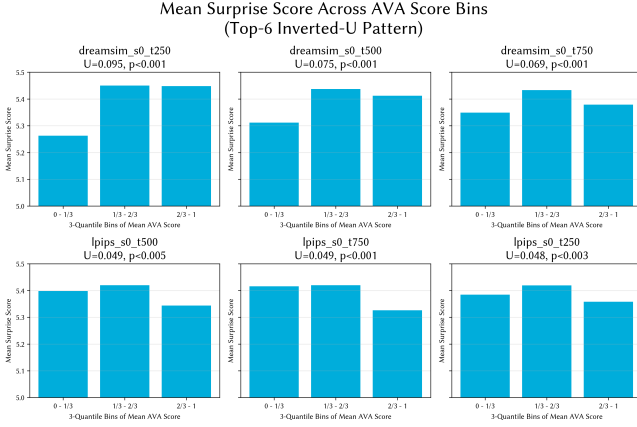


Fig. 7. Bar chart visualisations of our custom U-metric, used to measure how much a specific surprise score relates to human aesthetic scores in an inverted-U pattern. Here, we show the top six surprise scores and their noise/scale/metric combination in this regard. We notice that the top three were all at scale 0 (original) using the high-level DreamSim metric; i.e., we identify that semantic differences between an image and its diffusion reconstruction tend to follow Berlyne’s theory.

Figure 7 presents the top six (noise level, scale, metric) combinations with the highest, significant ($p < 0.05$) U-metrics. The top three in particular follow a specific pattern, where the spatial scale is 0 (original), and the metric is always DreamSim (high-level). In other words, our key finding is that surprise formulations that most closely follow Berlyne’s inverted-U pattern rely on semantic differences between an original image and its diffusion reconstruction.

4.3 Qualitative Examples

We present a few hand-selected examples illustrating the observed relationship between surprise and aesthetic score in Figure 8. When an image and its reconstruction are nearly identical (low surprise), we generally observe a low aesthetic score. This tendency is especially apparent for images lacking a clear subject, such as the leftmost pair in Figure 8. On the other hand, when the reconstruction diverges dramatically from the input, such as in the rightmost pair, the aesthetic score is also low. Unconventional subjects or compositions, including extreme close-ups of mundane objects, were common examples. Finally, images exhibiting moderate levels of difference tend to achieve the highest aesthetic ratings. In the centre pair of Figure 8, the primary discrepancy is the red bridge but not the background, and this corresponds to a high aesthetic score.

5 Discussion & Limitations

We observe that certain definitions of surprise, specifically the reconstructions from SD v1-5 at the original scale using the DreamSim metric, indeed exhibit the inverted-U pattern theorised by Berlyne. However, as shown in Figure 7, the effect size is small, as the y-axis spans 5.0 to 5.5 for a scale that is 0-10. While the inverse-U pattern is statistically significant ($p < 0.001$), its magnitude is limited. This outcome is somewhat expected. Aesthetic evaluation is a complex cognitive process shaped by many variables, including subjective,

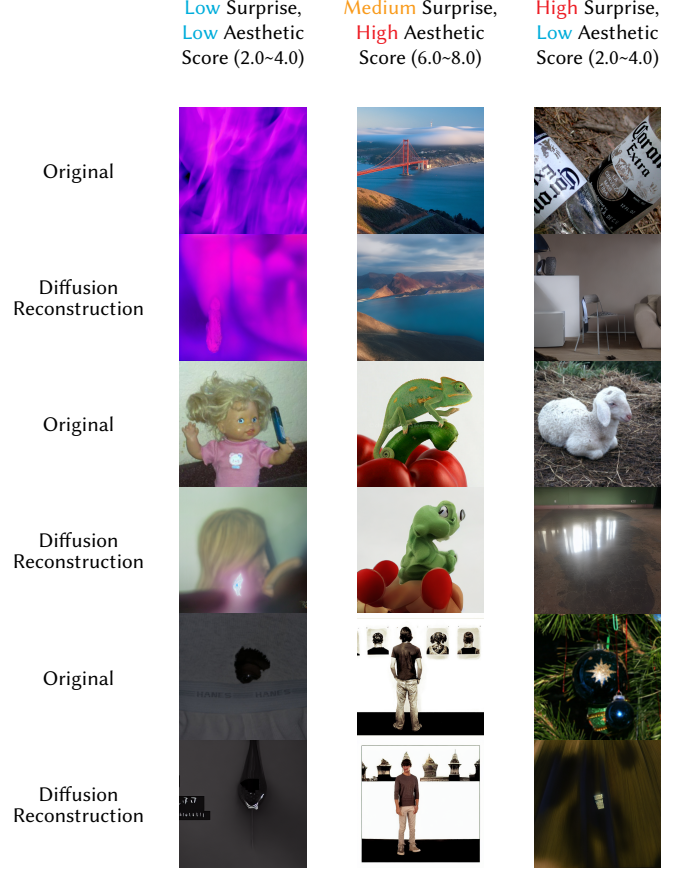


Fig. 8. Hand-picked example images and their high-noise ($t = 750$) reconstruction that satisfied our inverted-U surprise-aesthetics relationship. For some examples, such as the top-left purple pattern and the top-centre bridge scene, their surprise and aesthetic score are somewhat intuitive.

top-down factors far beyond the scope of our measurement. Our analysis isolates only particular variations of a single component in this broader context. For example, lighting alone can heavily influence aesthetics, likely affecting some of our qualitative examples as well, yet we do not address it in our method. Furthermore, there is no guarantee that the expectations encoded by a diffusion model align with human expectations. As a result, our surprise scores may deviate from what viewers intuitively find surprising. Future work may consider *personalised surprise estimation*, perhaps through few-shot modelling that learns an individual’s notion of surprise from a small set of examples.

6 Conclusion

We examined Berlyne’s 1970 hypothesis that images containing a moderate level of surprise correlate with the highest aesthetic value. Using a diffusion-based surprise measurement method inspired by recent anomaly detection methods, we showed that this relationship does emerge weakly under specific formulations of surprise. While our approach is not intended to improve upon the state of the art in aesthetic assessment, it provides evidence that modern

generative models can be used to explore the cognitive processes underlying visual preference. We hope that this work contributes to an understanding of how humans ultimately evaluate aesthetics.

References

- [1] Finn Behrendt, Debayan Bhattacharya, Julia Kruger, Roland Opfer, and Alexander Schlaefer. 2023. Patched Diffusion Models for Unsupervised Anomaly Detection in Brain MRI. *Proceedings of Machine Learning Research* 227 (2023), 1019–1032.
- [2] D. E. Berlyne. 1966. Curiosity and Exploration. *Science* 153, 3731 (1966), 25–33. <http://www.jstor.org/stable/1719694>
- [3] D. E. Berlyne. 1970. Novelty, complexity, and hedonic value. *Perception & Psychophysics* 8, 5 (Sept. 1970), 279–286. doi:10.3758/BF03212593
- [4] George Birkhoff. 1933. *Aesthetic Measure*. Harvard University Press.
- [5] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. doi:10.48550/arXiv.2306.09344
- [6] Peter beim Graben. 2025. Pragmatic information of aesthetic appraisal. *Cognitive Neurodynamics* 19, 1 (Dec. 2025), 39. doi:10.1007/s11571-025-10225-1
- [7] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. 2018. Exposure: A White-Box Photo Post-Processing Framework. *ACM Transactions on Graphics* 37, 2 (April 2018), 1–17. doi:10.1145/3181974
- [8] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Montreal, QC, Canada, 5128–5137. doi:10.1109/ICCV48922.2021.00510
- [9] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. 2023. VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Vancouver, BC, Canada, 10041–10051. doi:10.1109/CVPR52729.2023.00968
- [10] Jiawan Li, Fei Zhou, Zhipeng Zhong, Jiongzi Lin, and Guoping Qiu. 2025. Towards Smart Point-and-Shoot Photography. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 28242–28251. doi:10.1109/CVPR52734.2025.02630
- [11] N. Murray, L. Marchesotti, and F. Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Providence, RI, 2408–2415. doi:10.1109/CVPR.2012.6247954
- [12] Jaume Rigau, Miquel Feixas, and Mateu Sbert. 2008. Informational Aesthetics Measures. *Computational Aesthetics* (2008).
- [13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. doi:10.48550/arXiv.2112.10752
- [14] Xiangfei Sheng, Leida Li, Pengfei Chen, Jinjian Wu, Weisheng Dong, Yuzhe Yang, Liwu Xu, Yaqian Li, and Guangming Shi. 2023. AesCLIP: Multi-Attribute Contrastive Learning for Image Aesthetics Assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, Ottawa ON Canada, 1117–1126. doi:10.1145/3581783.3611969
- [15] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing* 27, 8 (Aug. 2018), 3998–4011. doi:10.1109/TIP.2018.2831899
- [16] Justin Tebbe and Jawad Tayyub. 2024. Dynamic Addition of Noise in a Diffusion Model for Anomaly Detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Seattle, WA, USA, 3940–3949. doi:10.1109/CVPRW63382.2024.00398
- [17] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861
- [18] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, 586–595. doi:10.1109/CVPR.2018.00068