Beauty in the Eye of Machine: Using Automated Measures of Aesthetic Beauty to Improve GAN Output of Satellite Images

Joseph Catlett
Supervisors: Willem van der Maden, Garrett Allen
Responsible Professors: Derek Lomas, Ujwal Gadiraju
EEMCS, Delft University of Technology, The Netherlands
June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

This paper aims to evaluate which automated measures of aesthetic beauty are the best predictors for human ratings of aesthetics and proposes that typicality and novelty may increase the correlation between the two. To study the correlation between these metrics, a literature study was performed to find a select amount of potentially good predictors, a pipeline was created to extract these values from each image within our dataset, a survey was conducted to vote for which images were considered most aesthetic, and finally regression analysis was performed to see which metrics offered highest correlation with the human rating data. From this we could see there were indeed a number of automated metrics that consistently scored high as predictors for the human aesthetic ratings and there was a slight improvement in the fit of the prediction model upon including novelty as a feature. However, at this moment, the improvement is not significant to conclude these features are better at predicting human ratings.

1 Introduction

Humans have been artistic creatures for millennia. There is something innate to the human experience that allows us to interpret and evaluate aesthetic beauty that has been somewhat of a mathematical challenge for some time. The field of computational aesthetics has been seeking to quantify this human experience, to reduce the complex function of human aesthetics to a simple equation of discrete metrics. In this paper, we hope to contribute to this decades old effort.

Being able to capture what makes images beautiful and automate the process of defining, categorising, and even improving the 'beauty' of an image has applications in a host of fields. AI (as well as other machine learning strategies) are already used in the generation of video-game characters, backgrounds, and levels [1], in VFX and image post-processing/editing [2; 3], image retrieval and categorisation [3] and has further potential application in any industry that relies on graphic imagery such as marketing, graphic design, or fashion.

However, the present research in the field of computation aesthetics focuses heavily on extracting visual and spatial features of an image and aggregating them in some way to compute some 'aesthetic score'. Despite there being a large body of work about what different features can be used in this process, there is not much consensus regarding which are the best predictors for human ratings in particular. We argue that there is something fundamentally missing from the approach of extracting features and scoring them, namely some measure of typicality and novelty. Which describe how images are perceived within a context of experience, being compared to the images seen before, the ones predicted to be seen next, and the environment at large. By adding this layer of aesthetic measure, we predict to see an increase in the correlation between automated aesthetic evaluation and human aesthetic ratings. This leaves us with two main sub-questions for this research paper:

1. Which existing automated measures of aesthetic beauty are the best predictors for human aesthetic ratings?
2. Does including the contextual approach of typicality and novelty improve the correlation between automated aesthetic rating and human aesthetic rating?

Both of these will be answered in the context of the LandShapes dataset which contains satellite images from Google's Earth Engine[1] as well as using images from the *This City Does Not Exist* GAN (General Adversarial Network) [4].

The structure of this paper will be as follows; Section 2 will begin with a strict and clear definition of the research question followed by a description of the methods and tools utilised to answer said question. Section 3 will be an overview of the literature survey done as well as the proposed algorithm and metrics to achieve the research goal. This will be followed by an in-depth discussion of the experiment's setup and results in Section 4. Finally in Sections 5 and 6, I will discuss the ethical implications of this research and conclude with a discussion of the results and possible future work.

2 Background Information

This section will briefly describe the history of aesthetic philosophy and outline the aesthetic theories which this paper will use as the foundation its research. We will also discuss the project with led to the conception of this research and in what context this question is being answered.

Aesthetic Theory

The philosophy and study of aesthetics has existed in explicit terms for over a century, being pioneered by Gustav Fechner in the 1870's [5]. Today, there exist a number of interpretations of the many experiments done in the field of aesthetics and there are many theories to describe the ways humans generate their aesthetic perceptions too. One of the most commonly cited is described in the book 'Aesthetic Measure' by George Birkhoff [6]. Birkhoff's theory is that the aesthetic measure, M, of an object can be quantified by the the ratio of its *Order*, O, and *Complexity*, C;

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

The definitions of order and complexity generally depend on the object being aesthetically graded, however order usually refers to some ability to encode the object's information in as few bits as possible, while complexity describes the intricacy and 'interestingness' of the image and the effort it takes to perceive it [6; 7].

However this theory was developed in 1933, and the goal of this paper is to expand upon and contribute to more recent, modern theories of aesthetics. This paper will be assuming the Unified Theory of Aesthetic Pleasure, developed by Paul

[1] https://earthengine.google.com/

Hekkert in 2014 [8]. Hekkerts theory recognises different levels to the cognitive processing of stimuli, dividing the process into three levels: perceptual (unity-in-variety), cognitive (typicality and novelty), and social (connectedness and autonomy). These levels refer to the visual and sensory stimuli of the object, the contextual nature of the object, and the contextual nature of the social environment, respectively.

Hekkert expanded upon his theory in 2016, along with Michael Berghman [9]. In [9], they developed a number of experiments demonstrating the degree to which each of these levels contribute to the aesthetic experience of different products and objects. This study was done via direct surveying of a participants' opinions on the aforementioned three levels.

This paper however seeks to find a way to automate the process of quantifying the second level, typicality and novelty, within the context of AI generated satellite imagery. It also attempts to use this automated quantification to predict the aesthetic ratings of human survey participants.

Foundational Project

This research uses data from Frederik Ueberschaer's Master Thesis, LandShapes [10]. In this project, Ueberschaer sought to answer three questions: can artistic experiences that provoke positive emotional engagement foster climate change awareness and action, can GANs be used to produce such experiences, and can playful and interactive components strengthen the engagement with this experience.

The role of emotions, both negative and positive, can greatly influence the perception and support of political policy, specifically with regard to climate action [11]. From [11], it was found that emotions such as worry, hope, and interest were the most powerful emotional explanations of variance in support for national climate policy, and that images that incited these emotions had noticeable influence over support for policy. This last point proposes the concept that imagery can be a powerful source of emotional stimulation. It has also been shown that the act of motivational state appraisal, or analysing with regard to its pleasure-pain balance, can lead to emotional catharsis [12]. This study was done by exposing participants to the possibility of having to taste certain potentially good or bad tasting foods. However, this concept of appraisal being the source of emotions has been extended to the appraisal of artwork [13].

This then creates the context and argument for this paper. Ueberschaer has created a dataset of images and a trained GAN with the intent of stimulating emotional responses. There is strong evidence that emotional responses can be dependent on aesthetic appraisal, therefore we can see why being able to automatically evaluate aesthetics can be impactful in the context of producing emotionally, politically, and environmentally engaging content.

3 Methodology

This section will provide an outline of the visual and spatial features used in the measuring of aesthetic beauty. A brief argument for their relevance will be provided and I will also discuss any algorithms used to process them, the libraries and tool-kits used, and an overview of the pipeline implemented to combine all of these features.

The idea behind this experiment was to separate the visual, spatial, and contextual features of an image to argue that visual and spatial features are not enough to predict human aesthetic ratings. Visual features are the easiest to understand; both as concepts and descriptors, but also with regard to their relevance to aesthetics.

- **Saturation:** Adults generally have a preference for higher saturated colours, especially within a western context [14].
- **Luminance:** There is evidence to support a link between brighter hues and preference for an image [14]. Beyond this however, it is also a common metric for image quality in photography, art, and other visual industries [15; 16; 14].
- **Contrast:** Contrast emphasises traits such as hue, colour combinations, and luminance and as such has evidential support for correlation with higher human rating [14]. Similar to luminance, contrast is often used as a quality measure in a host of visual fields [15; 16; 14].
- **Sharpness;** Having edges that are well defined and clear was identified as a good predictor of high human ratings of aesthetics [16; 14].
- **Colour Histogram:** There is limited evidence of generality with regard to colour preference across gender, age, culture, or even within these groups [14]. However, from [14], certain colour combinations have been observed to be commonly used within photography, visual arts, and have been linked to predicting human preference. There is were also informal surveys done with the images produced by LandShapes conducted by members of the LandShapes team. In it, participants were asked to explain what made the images aesthetically pleasing and in meetings with the conductors of the survey, a common remark was that of the combination of land and sea was a popular responds, making the combination of shades of green and blue a potentially successful indicator of high aesthetic value within this context.

Spatial features are slightly more abstract and may require some further definition. Spatial features refer to the underlying components of an image that are not immediately visible to the observer but still have an impact on the appraisal of said image.

- **Rule of Thirds & Diagonal Dominance:** These are common practices of image composition, most often used in photography and visual arts. An example of the Rule of Thirds and Diagonal Dominance can be seen in Fig. 2 and 1. The Rule of Thirds is supposed to be a harmonizing placement technique [19]. It is regarded as one of the most important compositional rules used in painting and photography [20; 19]. Diagonal Dominance is not as widely recognized but is still a popular technique in photography as it creates a leading line for the eyes to fixate on, drawing the observer to areas of interest [21].

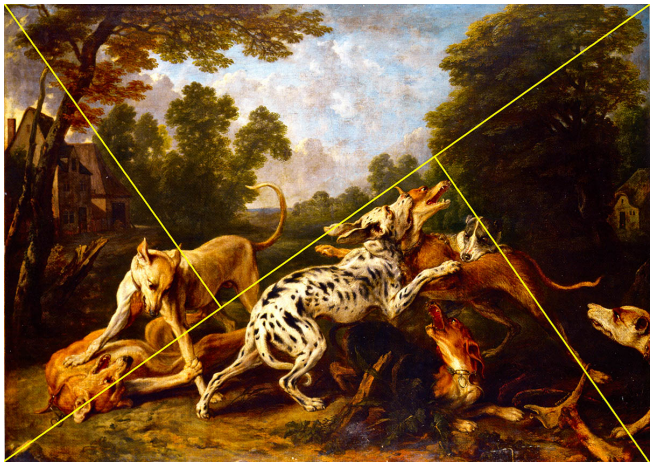

Figure 1: Painting, "Dogs fighting in a wooded clearing", with overlay showing how the salient regions follow lie mostly along diagonals [17].

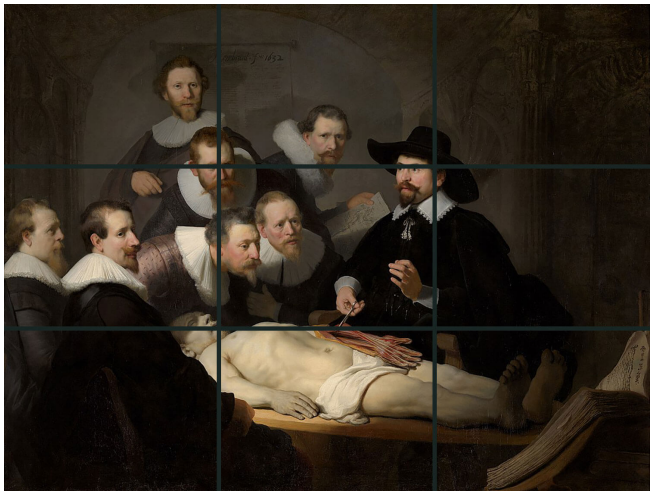

Figure 2: Rembrandt's "The Anatomy Lesson of Dr. Nicolaes Tulp" oil painting, shown adhering to the "Rule of Thirds" principal or salient objects and contours following gridlines and intersection points [18].

- **Entropy:** This is a measures of image complexity, describing how much informational data it takes to encode an image. While there is little evidence demonstrating a correlation between human ratings of aesthetics and entropy, it is a common measure of spatial information within an image [22; 15; 23; 16; 14]. Having images with more salient regions of high detail/interestingness, contrasting regions of complexity, and images having a general threshold of 'intrigue' has been linked to be more popular among observers of photography and geometric spatial patterns [22; 16; 14].

- **Symmetry:** Having well balanced and symmetrical images with continuity and repetition make for a more coherent impression and smoother interpretation experience, leading to increase in preference for these images [24; 25]. However it is worth saying, overly predictable images that are dull or monotonous become boring, so having a balance with asymmetrical variety is required [9]

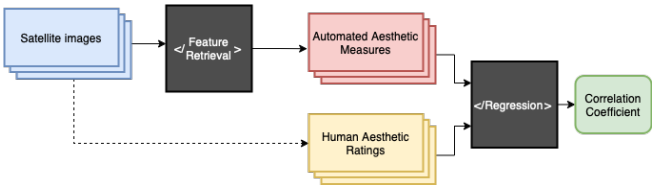

Figure 3: This diagram illustrates the pipeline from dataset to final correlation coefficient.

- **Line Orientation Ratios:** Studies done on paintings have demonstrated a preference for horizontally dominated images (objects and lines lean more horizontally than vertical) [14]. This same study also showed that horizontal and vertical lines are both preferred over diagonal lines[2].

Finally, beyond these more fundamental visual and spatial features, we have the contextual features, namely typicality and novelty.

- **Typicality:** When the brain observes something it perceives within the frame of reference of past experiences. Items that are similar to other stimuli or that are inline with what was expected allow for smoother processing. This smoother processing allows for more appreciation [9]. There is also a phenomenon known as the 'Mere Exposure Effect', where appreciation can emerge from the sheer frequency of experience [26].

- **Novelty:** However, in contrast, having experiences that are new, unexpected, and novel, allow us to deconstruct and analyse, enriching the experience as a whole [9].

Once these features were decided upon, a software pipeline was produced to gather, process, and display the relevant data for interpretation., illustrated in Fig. 3. Firstly, the dataset is iterated through and for each image the mentioned visual and spatial features are extracted. The image name and corresponding feature data are organised into rows within a csv file with each column representing a single feature. This data is pre-processed into a well formatted dataframe, including image resizing and normalisation of columns.

Then, dimensionality reduction is performed using Principal Component Analysis (PCA) and using the features with the most variance are used for K-Means Clustering. The clusters are used as feature profiles (representing images of similar complexity, content, colours et cetera). The distance between any single image to the centroid of its own profile is the measure of novelty and the average distance between the image and all centroid clusters is a measure of typicality. This new feature value is added to the corresponding csv row in a new column.

[2]This is not to be confused with *Diagonal Dominance* which is the act of placing objects of interest/eye-catching objects along the diagonal of an image, not the object itself necessarily being orientated diagonally.

The human aesthetic rating is also added to the rows in a new column, measured by the number of votes each image received as the "most pleasing to the eye". This data was gathered from crowdsourced workers where each participant filled in a survey, selecting from a panel of 4 options which image was the most aesthetically pleasing [27]. These votes were summed and normalised into a measure of human aesthetic rating for each image. This value, appended to the rows of the images in the csv, is used in a linear regression[3] to measure the correlation between the features and the human ratings. This shows us the reliability of the fit of these features as a whole for human aesthetic ratings as well as the individual significance per feature within the model.

A second linear regression is also performed without the typicality and novelty features. the general fit and individual feature scores of this model was then compared to the model including these contextual features.

OpenCV[4] was used for much of the image processing and gathering of many of the spatial and visual features. NumPy[5], SciPy[6], scikit-learn[7], pandas.py[8] were used to process and cluster the data. Specifically the StandardScalar, PCA, KMeans, and TSNE libraries were used to scale the images, perform dimensionality reduction and clustering, and represent the clusters in a human-readable format, respectively. Statsmodels[9] was used to conduct the actual linear regression modelling.

4 Experimental Setup and Results

This section will discuss how the environment was setup for testing, how the output was obtained and then what data analysis techniques were used to gather the final results.

To generate these feature profiles used in the computation of typicality and novelty, we used K-Means Clustering. This meant that we first had to reduce the dimensionality of the feature space. We decided to utilise PCA (Principal Component Analysis). Both of these processes required prior data analysis.

Principal Component Analysis

There are a total of 12 features extracted from the images, excluding typicality and novelty. Using each of these as a dimension would produce an unnecessarily complex, high dimensional space for the clustering. PCA reduces the output space to the features with the most variance by selecting a subset of features with the highest variance. Fig 4 displays the reduction in dimension with respect to variance. Maintaining a variance of 80% is a heuristically popular approach [28], and by applying this here, we get a dimensionality reduction of $12 \rightarrow 7$.

[3]Ordinary Least Squares Regression
[4]https://pypi.org/project/opencv-python/
[5]https://numpy.org
[6]https://scipy.org
[7]https://scikit-learn.org/stable/
[8]https://pandas.pydata.org
[9]https://www.statsmodels.org/stable/index.html

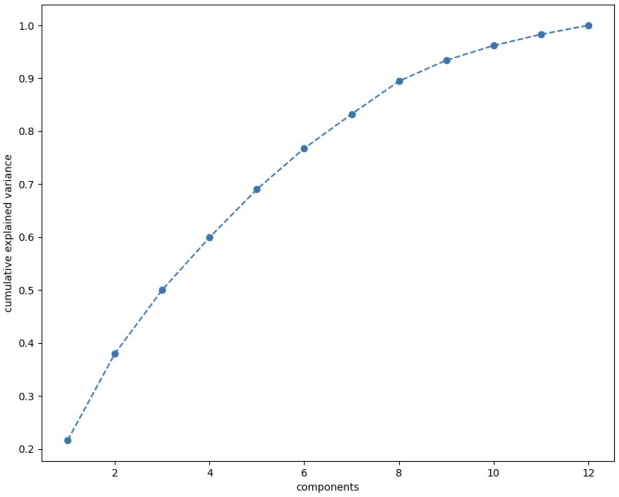

Figure 4: Graph demonstrating the increase in variance with regard to number of components (dimensions) clustered upon.

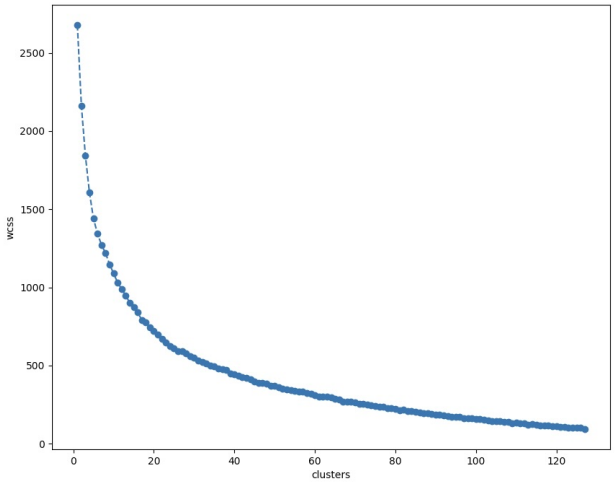

Figure 5: Graph showing the decline in *Within Cluster Sum of Squares* for all $k \in [1, 128]$.

K-Means Clustering

When performing K-Means Clustering, it is important choose a k-value (number of clusters) that minimises the *Within Cluster Sum of Squares* (WCSS), or the average squared distance from all elements to their cluster's centroid for all clusters $k \in [1, n]$, where n some number less than the number of images being sampled. With the dataset of 256 images, k-means clustering was done for all values of $k \in [1, n]$, where $n=128$. This generated the graph shown in Fig 5. The optimal k-value is one that strikes a balance between minimising WCSS and k. From this graph, we chose $k=7$. [10]

[10]This value was also later confirmed by running the entire program and generating a regression model for all k values $k \in [6, 14]$ and observing $k=7$ indeed produced the highest *Adjusted R-Squared Value*, which is a measure of how fitted the model is.

OLS Regression Results

Dep. Variable:	aesthetic	R-squared:	0.424
Model:	OLS	Adj. R-squared:	0.396
Method:	Least Squares	F-statistic:	14.93
Date:	Thu, 16 Jun 2022	Prob (F-statistic):	2.60e-23
Time:	11:21:49	Log-Likelihood:	552.04
No. Observations:	256	AIC:	-1078.
Df Residuals:	243	BIC:	-1032.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3623	0.056	6.418	0.000	0.251	0.474
histograms	-0.3622	0.047	-7.788	0.000	-0.454	-0.271
entropy	-0.0105	0.003	-3.109	0.002	-0.017	-0.004
straight_diagonal_line_ratio	-0.0416	0.024	-1.733	0.084	-0.089	0.006
horizontal_vertical_line_ration	-0.0741	0.036	-2.064	0.040	-0.145	-0.003
diagonal_dominance	-6.257e-07	5.27e-06	-0.119	0.906	-1.1e-05	9.75e-06
symmetry	-9.757e-05	5.75e-05	-1.698	0.091	-0.000	1.56e-05
rule_of_thirds_power_points	-2.687e-05	4.35e-05	-0.617	0.538	-0.000	5.89e-05
rule_of_thirds_gridlines	6.759e-05	4.69e-05	1.441	0.151	-2.48e-05	0.000
sharpness	1.083e-06	1.08e-06	1.004	0.316	-1.04e-06	3.21e-06
contrast	0.0002	6.23e-05	2.809	0.005	5.23e-05	0.000
luminance	0.0002	5.49e-05	2.866	0.005	4.92e-05	0.000
saturation	0.0002	3.97e-05	4.910	0.000	0.000	0.000

Figure 6: Output of the *Ordinary Least Squares* Regression excluding *typicality* and *novelty* as predictive features. In red you can see the *Adjusted R-Squared* value, which reflects the fit of the model on a scale between 0 and 1. In general, a higher value indicated higher correlation. In blue you can see the *p-value* per feature. This value represents how significant the feature was in the predictive model. Heuristically a p-value below 0.05 is deemed statistically significant.

OLS Regression Results

Dep. Variable:	aesthetic	R-squared:	0.452
Model:	OLS	Adj. R-squared:	0.421
Method:	Least Squares	F-statistic:	14.22
Date:	Wed, 15 Jun 2022	Prob (F-statistic):	1.40e-24
Time:	13:43:10	Log-Likelihood:	558.43
No. Observations:	256	AIC:	-1087.
Df Residuals:	241	BIC:	-1034.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3580	0.060	5.964	0.000	0.240	0.476
novelty	0.1030	0.033	3.137	0.002	0.038	0.168
typicality	-0.0389	0.047	-0.826	0.410	-0.132	0.054
histograms	-0.3699	0.047	-7.939	0.000	-0.462	-0.278
entropy	-0.0068	0.004	-1.834	0.068	-0.014	0.001
straight_diagonal_line_ratio	-0.0646	0.026	-2.505	0.013	-0.115	-0.014
horizontal_vertical_line_ration	-0.0747	0.035	-2.123	0.035	-0.144	-0.005
diagonal_dominance	1.261e-06	5.21e-06	0.242	0.809	-9e-06	1.15e-05
symmetry	-0.0001	5.69e-05	-2.090	0.038	-0.000	-6.83e-06
rule_of_thirds_power_points	-4.082e-05	4.28e-05	-0.953	0.342	-0.000	4.36e-05
rule_of_thirds_gridlines	4.596e-05	4.95e-05	0.928	0.354	-5.16e-05	0.000
sharpness	-5.001e-07	1.35e-06	-0.371	0.711	-3.16e-06	2.16e-06
contrast	0.0002	6.75e-05	2.848	0.005	5.92e-05	0.000
luminance	0.0001	5.83e-05	1.856	0.065	-6.61e-06	0.000
saturation	0.0002	3.93e-05	5.222	0.000	0.000	0.000

Figure 7: Output of the *Ordinary Least Squares* Regression including typicality and novelty as predictive features. The colour highlighted values correspond to the same meaning as Fig. 6

Regression

With these two pre-processing operations complete and the typicality and novelty values for each image appended to the dataframe, we can now complete the Ordinary Least Squares (OLS) regression analysis. OLS was run on the data twice, once including typicality and novelty as a predictive feature, and once excluding them. All other variables and features were kept consistent across iterations. Figures 7 and 6 depict the output of the OLS regression excluding and including typicality and novelty, respectively.

Results

When we compare the output of the output of the OLS regression without *typicality* and *novelty* versus with, as depicted in Figures 6 and 6, we can see that there is an increase in the model fit from $r_{without} = 0.396$ to $r_{with} = 0.421$, which is an an increase of 6.31%. Since the only difference between the two tests was the inclusion of *typicality* and *novelty*, we can safely assume this increase in predictability comes from these features.

5 Discussion

While there was indeed a positive increase in fit, both models only have a moderate fitting score. We should be careful categorising effects as weak, medium, and strong based off of the Adj. R-Squared Value because it depends entirely on the research field in which we are conduction our research [29]. However, despite this, we do need to interpret these values, and one common approach to define r $\geq$ 0.75 as substantial, 0.75 $>$ r $\geq$ 0.50 as moderate, 0.5 $>$ r $\geq$ 0.25 as weak, and anything below this as unsubstantial [30]. With both models falling into the moderate category, it is unclear as to whether there is enough statistical evidence to support the claim contextual features provide a significant improvement over visual and spatial features in predicting the human ratings of aesthetic beauty for satellite images.

This is further supported when we analyse which features are considered statistically significant. In both Figures 6 and 7, there a number of visual and spatial features deemed statistically significant, more so than typicality and novelty. Saturation, contrast, horizontal and vertical line ratio, and colour histogram are significant across the two tests. There are also other features that are measured as significant in individual tests. This, alongside the fact *typicality* is not even close to being a significantly predictive feature ($p_{typicality} = 0.410 \gg 0.05$), it could be argued the improvement simply comes from the addition of another *"general feature"*. In other words adding another visual or spatial feature could provide a similar or even better improvement.

This is somewhat in line with the predictions made by Hekkert and Berghman [9]. In their study, they state "effects on the perceptual level are quite substantial, and the relative importance of unity and variety differs across surveys," [9]. Numerous times throughout the study, they state that the perceptual level (visual and spatial features) offer the most variance and therefore account for most of predictive qualities of the model. Which is indeed what we see in our study with the limited difference in model fit between the inclusion and exclusion of the cognitive level or contextual features

We also see another prediction of theirs coming true; "The perceptual qualities of unity and variety maintain the largest effect. By contrast, whether a design is considered typical turns out to be less important for aesthetic appreciation when controlled for qualities at the perceptual and social level," [9]. Again, Hekkert and Berghman state that perceptual qualities are the strongest predictors, but also that typicality is less im-

portant for aesthetic appreciation. They mention the lack of statistical significance for typicality throughout the study. In our own study, typicality was also one of the lowest statistically significant features.

This lack of significant improvement could be because of the underlying relationship between the visual and spatial features and the contextual features. In Section 4, it was explained how through PCA and K-Means clustering, profiles were generated and these were the foundation for typicality and novelty. However, this means that the visual and spatial features provided the underlying structure used to produce the contextual values. Any underlying patterns or correlations in those features may be imposed onto the contextual features, and therefore the contextual features simply become a summary for the visual and spatial features.

One aspect of this study which may interfere with the reliability of the results is with the way in which participants were asked to select images. When being presented with 4 images, voting for one image does not necessarily mean the other 3 are not aesthetically pleasing. Even in the case some image n_i is chosen 100% of the time, that does not mean image n_j, n_k, or n_l are ugly or bad. This may be a limiting factor to the increase in model fit, and we predict that using a different human rating system, such as a 7 point Likert Scale, may provide a more suitable environment for this type of regression.

There are other contextual aspects of rating satellite images that can be integrated attempt to mitigate this interference. These include (but are not limited to) the order in which the images are shown to participants, considering the level of prior exposure or familiarity to satellite imagery, locality of participants and the similarity of locality to the images being displayed.

With regard to which of the visual and spatial features were considered statistically significant, it is interesting to see that which ones exceeded the threshold differed between the inclusion and exclusion of typicality and novelty. This was an unexpected result.

The significance of saturation, contrast, and luminance is in line with a lot of the literature on aesthetics today [14]. Another concurring finding was that of the spatial qualities such as entropy and line rations. These were consistently hovering around the significant threshold as predicted by Palmer, Schloss, and Sammartino [14].

However, one contrasting finding with Palmer, Schloss, and Sammartino is the strength of the colour histograms. In [14], it is argued that there is no discernible patterns across cultures, age groups, gender, or any other group that has a reliable prediction of what colours are preferred or favourable. However, in our study, colour histograms is the most significant feature. This most likely due to the limited context in which the 'favourable colours' are being selected. The images limit the participants ability to choose, and here colours are related to specific concepts or objects (blue is water, green is fields or plains, yellow and orange sand, grey is urban et cetera). This does corroborate the study done by current members of the LandShapes project, where interviewees reported land and sea combinations as being among the image qualities they found aesthetic.

6 Responsible Research

The following section will discuss responsible research practices and how we maintained an ethical and integral approach to this research project. During this project, The Netherlands Code of Conduct for Research Integrity was used as a reference for ethical practices and responsible research principles [31]. Many of them are general and broad, however some are quite relevant for this paper and these will be discussed in detail.

Data Integrity

Data manipulation is an umbrella term that refers to the fabrication of data points or results, trimming or omitting data, manipulating the data such as 'cherry-picking' or simply misrepresenting the data used or observed. Since my project is using large datasets it is important to be transparent and honest about the nature of how these images were selected, manipulated (if at all), and used, with specific reference to how the images were displayed to rating participants and the measured images consistency.

As mentioned in Section 2, the dataset used is a combination of images from Frederik Uebershcauer's, LandShapes [10] and from the openly available, online GAN *This City Does Not Exist* [4]. The data was not trimmed or filtered in anyway to manufacture some result, and the only manipulation performed were scaling operations; the visual and spatial features were kept consistent for both the human participants and the automated ranking system. Neither was there any manipulation of the automatic aesthetic ratings nor the human ratings.

Plagiarism

As is the case with any research and software project, ensuring the correct usage and crediting of softwares, data, existing research, and resources is imperative. All software libraries and tools used to produce this were open-source and license free. I have also explained and linked to all of them in Section 3. To ensure absolute transparency the source code used for this entire project is also available on GitHub[11].

With regard to the images used, they were taken from the LandShapes project which were obtained from Google Earth Engine. As mentioned in *"Section 1: Provision of the Services, 1.1 Limitations of Use"* of the Google Earth Engine Terms of Service, the images and licenses are free to use for research and educational purposes [32].

Many research papers were used to familiarise ourselves with every aspect of this topic, including technical comparisons of image processing techniques and compression algorithms, theoretical discussions of aesthetics, cultural and anthropological studies of biases and preferences to visual stimuli, among others. However, much of it did not manifest in this project. That which did, has been explicitly remarked on in the relevant sections throughout the paper and cited in IEEE format, as dictated by *IEEE Reference Guide* [33].

[11] https://github.com/jcatlett99/Research_Project.git

Reproducibility

It is important to produce and describe a study such that that setup and processing are able to be reproduce, allowing for validation and verification of research. Obviously, a large part of this is being descriptive and transparent within this research paper. Every attempt has been made to do so, furthermore, this paper is made publicly available and the associated researchers' are contactable through the affiliated institutions to allow for critique, inquiry, and possible future edits. As much of the project is made available for this same reason. As mentioned in the previous section, the codebase is available on GitHub.

The dataset is not included as this dataset is curated by Frederik Ueberschauer for LandShapes [10]. However, despite being in the context of satellite imagery, this project should be reproducible on any context of images (so long as the dimensional requirements for the image processing are met). Not providing these dataset also prevent individuals from using the images obtained from this project for commercial or any other usage that violates Google Earth Engine's Terms of Service.

Ethical Inclusion of People

Since we involved crowdsourced workers we also had to consider the ethical ramifications of having external participants on the research but also on them as participants. Every survey worker was clearly informed about what they would be asked to do, what data was going to be collected, and the general purpose of that data. They were asked to give explicit consent before participating. The only data collected was the rating they assigned to an image or the preferred image from a subset of four images (depending on which version of the survey they filled in). No identifiable data was collected. The survey was created using Qualtrics[12] by Moshuir Rahman [27]. Rahman also recruited the participants and hosted the survey via Prolific[13]. Every participant was compensated at a rate of €10:00/hour for filling the survey, the standard rate for Prolific workers [27].

We also needed to consider how using this data may effect the research, for example if there were bots made for the survey, people did not the read the questions and answered flippantly, or there was extreme outliers such as someone purposefully answering alternatively. To combat this, we had attention checks throughout the survey to ensure people were reading the questions and paying attention, not clicking randomly. We also removed any results that proved to statistical outliers beyond some reasonable doubt [27].

Bias

When engaging in any scientific research it is important to try an mitigate the interference of bias as much as possible, but especially when the content is something as inherently subjective as human perceptions of aesthetic beauty.

As previously mentioned, the dataset was not interfered with nor curated in any way, removing any possible skewing of results in this regard. One area that is susceptible to bias is the selection of features to use as the automated aesthetic features. The method of selection was a literature study; many papers were read, the arguments and statistics were compared and candidates were proposed. Then we discussed the legitimacy of the features before encoding them into the program. This way multiple people contributed to selection to try and mitigate any one persons' personal bias toward a specific feature.

This bias is also present in the actual encoding of the features' extraction and usage. The code has been made available in order to allow public viewing and discourse to make critical analysis of software decisions throughout the process. Other members of the research team were also kept up to date and consulted to again reduce bias as much as possible.

Bias is inherent to the human condition and especially with regard to our own personal interpretation of the aesthetic experience of the world. There was a great amount of consideration about this throughout the project at all times was mitigated as much as possible.

7 Conclusions and Future Work

We set out to investigate how using automated measures of aesthetic beauty can improve GAN output of satellite images. This was to be done by answering two sub-questions; "which existing automated measures of aesthetic beauty are the best predictors for human aesthetic ratings?", and "does including the contextual approach of typicality and novelty improve the correlation between automated and human aesthetic rating?"

This was to be done by having participants vote for which images they believed to be the most aesthetic. Then, using the same dataset, extract a number of popular visual and spatial features, use them and the nature of the dataset to compute the contextual features of typicality and novelty. Finally, we can perform statistical analysis on the outputted values, in the form of a Ordinary Least Squares Regression model.

From this model, we found that when including contextual features the most predictive visual and spatial features were (in order of statistical significance), saturation and colour histograms, novelty, contrast, straight to diagonal line ration, horizontal to vertical line ratio, and symmetry [14]. Typicality was not statistically significant, and was in fact the 2nd last in terms of significance, behind only diagonal dominance.

When we excluded the contextual features, we saw similar results with some variation. Luminance and entropy are regarded significant while straight to diagonal line ratio and symmetry cease to be significant. If we were to take the intersection of these two results to answer the first sub-question, we would say the most statistically significant visual and spatial image features for predicting human aesthetic ratings are colour histograms, saturation, contrast, and horizontal to vertical line ratio.

[12] https://www.qualtrics.com/uk/

[13] https://www.prolific.co

[14] $p=0.000$, $p=0.002$, $p=0.005$, $p=0.013$, $p=0.035$, $p=0.038$, respectively

Comparing the overall fit of the models (via Adjusted R-Squared Value), the was indeed an improvement of fit when we include the contextual features, increasing from from $r_{without} = 0.396$ to $r_{with} = 0.421$. However, both of these values can be considered only moderately predictive at best [30], and as discussed in Section 5, may be only be due to the relationship between the contextual features and underlying visual and spatial features upon which they were built on. It is therefore still unclear about how well *typicality* and *novelty* are at improving the the correlation between automated ratings of aesthetics and human aesthetic ratings.

With regard to how these findings can be used to improve GAN output, the most obvious and applicable integration would be through a curation technique. Have the GAN produce *N* images, have the images rated by this automatic process, select the *n* most aesthetic images to feed back into the GAN for retraining.

The ambiguous relationship between the contextual and visual-spatial features produces an interesting follow up to this work. As discussed in Section 5, there a number of alternative contextual qualities that could also have been used. It would be interesting to also introduce these to the model, and do comparisons between them to answer the question "Which contextual image features are the best predictors for human aesthetic ratings?"

As well as this, doing a comparative study within contexts outside of satellite images for both contextual and visual-spatial features can illuminate the consistency and reliability of these features as human aesthetic rating predictors in a broader context.

Finally, we predict interference between the 'voting' method of aesthetic categorisation used in Rahman's survey [27], and believe that the introduction of a Likert Scale, so that every image can be given an independent aesthetic grade could offer a better environment to conduct this study.

References

[1] R. Rodriguez Torrado, A. Khalifa, M. Cerny Green, N. Justesen, S. Risi, and J. Togelius. Bootstrapping conditional gans for video game level generation. *2020 IEEE Conference on Games (CoG)*, pages 41–48, 2020.

[2] F. Yang, J. Ren, Z. Lu, J. Zhang, and Q. Zhang. Rain-component-aware capsule-gan for single image de-raining. *Pattern Recognition*, 123:108377, 2022.

[3] T. O. Aydın, A. Smolic, and M. Gross. Automated aesthetic analysis of photographic images. *IEEE Transactions on Visualization and Computer Graphics*, 21(1):31–42, 2015.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. pages 8110–8119, 2020.

[5] G. T. Fechner. *Zur experimentalen Aesthetik*. Abhandlungen der Mathematisch-Physischen Classe der Königl. Sächsischen Gesellschaft der Wissenschaften. Bei S. Hirzel, 1871.

[6] G. D. Birkhoff. *Aesthetic Measure*. Harvard University Press, 1933.

[7] V. Douchová. Birkhoff's aesthetic measure. *AUC PHILOSOPHICA ET HISTORICA*, 2015:39–53, 2016.

[8] P. Hekkert. *Aesthetic responses todesign: a battle of impulses*, page 277–299. Cambridge Handbooks in Psychology. Cambridge University Press, 2014.

[9] M. Berghman and P. Hekkert. Towards a unified model of aesthetic pleasure in design. *New Ideas in Psychology*, 47:136–144, 2017.

[10] F. Ueberschär. Ai for experience: Designing with generative adversarial networks to evoke climate fascination. Master's thesis, Delft University of Technology, 2021.

[11] N. Smith and A. Leiserowitz. The role of emotion in global warming policy support and opposition. *Risk Analysis*, 34(4):937–948, 2014.

[12] I. Roseman and A. Evdokas. Appraisals cause experienced emotions: Experimental evidence. *Review of General Psychology*, 18(1):1–28, 2004.

[13] P. J. Silvia. Emotional responses to art: From collation and arousal to cognition and emotion. *Review of General Psychology*, 9(4):342–357, 2005.

[14] S. Palmer, K. Schloss, and J. Sammartino. Visual aesthetics and human preference. *Annual Review of Psychology*, 64(1):77–107, 2013.

[15] P. Bhandari, M. Jaiswal, V. Puranik, S. Kirtane, and D. Pukale. Image aesthetic assessment using deep learning for automated classification of images into appealing or not-appealing. *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020.

[16] T. Aydin, A. Smolic, and M. Gross. Automated aesthetic analysis of photographic images. *IEEE Transactions on Visualization and Computer Graphics*, 21(1):31–42, 2015.

[17] F. Snyders. Example of golden triangle method on a painting. compositional elements fall within the triangles, 2018. 2018. Accessed: May 27, 2022. [Online]. Available: https://en.wikipedia.org/wiki/Golden_triangle_(composition)#/media/File:Snyders_Dogs_fighting_demonstrating_Golden_Triangle_composition_method.jpg.

[18] H. Symons. How to create the rule of thirds to improve your art. Haydn Symons, 2022. Accessed: Apr. 18, 2022. [Online]. Available: https://www.haydnsymons.com/blog/how-to-create-the-rule-of-thirds/.

[19] S. Amirshahi, G. Hayn-Leichsenring, J. Denzler, and C. Redies. Evaluating the rule of thirds in photographs and paintings. *Art & Perception*, 2(1-2):163–182, 2014.

[20] B. Gooch, E. Reinhard, C. Moulding, and P. Shirley. Artistic composition for image creation. In *Eurographics Workshop on Rendering*, pages 83–88, 2001.

[21] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing photo composition. *Computer Graphics Forum*, 29(2):469–478, 2014.

[22] F. Bertacchini, P. Pantano, and E. Bilotta. Shaping the aesthetical landscape by using image statistics measures. *Acta Psychologica*, 224, 2022.

[23] B. Spehar, C. Clifford, B. Newell, and R. Taylo. Universal aesthetic of fractals. *Computers & Graphics*, 27(5):813–820, 2003.

[24] R. Arnheim. *Art and Visual Perception, Second Edition: A Psychology of the Creative Eye*. Art Psychology. University of California Press, 2004.

[25] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, and R. von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological Bulletin*, 138(6):1172–1217, 2012.

[26] R. F. Bornstein. Exposure and affect: Overview and meta-analysis of research. *Psychological bulletin*, 106(2):265, 1989.

[27] M. Rahman. How can crowdsourced workers effectively rate landscape artwork images produced by generative adversarial network transformers?, 2022. [Unpublished].

[28] E. Kaloyanova. How to combine pca and k-means clustering in python? 365 Data Science, 2020. Accessed: Jun. 02, 2022. [Online]. Available: https://365datascience.com/tutorials/python-tutorials/pca-k-means/.

[29] A. Field. *Discovering statistics using IBM SPSS statistics: North American Edition*. SAGE Publications, Inc, 2018.

[30] M. Sarstedt and E. Mooi. Regression analysis. *Springer Texts in Business and Economics*, pages 193–233, 2014.

[31] KNAW, NFU, NWO, TO2-federatie, Vereniging Hogescholen, and VSNU. Netherlands code of conduct for research integrity, 2018. Accessed: May 03, 2022. [Online]. Available: https://www.universiteitenvannederland.nl/files/documents/Netherlands%20Code%20of%20Conduct%20for%20Research%20Integrity%202018.pdf.

[32] Google earth engine terms of service, 2019. Google, 2019. Accessed: June 10, 2022. [Online]. Available: https://earthengine.google.com/terms/.

[33] IEEE Referecence Style, IEEE Periodicals:1-17, 2018. Accessed: June 10, 2022. [Online]. Available: https://ieeeauthorcenter.ieee.org/wp-content/uploads/IEEE-Reference-Guide.pdf.