# GANAesthetic : An experience of interactively exploring aesthetically pleasing images and incorporating the human perception of beauty to discover aesthetic latent dimensions

**Ton Hoang Nguyen (Bill)**
**Derek Lomas, Willem van der Maden, Ujwal Gadiraju, Garrett Allen EEMCS,**
**Delft University of Technology, The Netherlands**
22-6-2022

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

# GANAesthetic : An experience of interactively exploring aesthetically pleasing images and incorporating the human perception of beauty to discover aesthetic latent dimensions

**Ton Hoang Nguyen (Bill)**[2]
**Derek Lomas**[1] , **Willem van der Maden**[1] , **Ujwal Gadiraju**[2] , **Garrett Allen**[2]
[1]HCD, IDE, Delft University of Technology, The Netherlands
[2]EEMCS, Delft University of Technology, The Netherlands

## Abstract

**Despite the fact that climate change is becoming increasingly dangerous and prevalent, there is still a lack of public engagement. This can be explained by the fact that the media portrays climate change as an abstract concept. The message can be more effectively communicated through visual art because it is more likely to invoke emotional responses in individuals. By including human perception and rating data, the generative adversarial neural network (GAN) produces better image output. Therefore, this paper explores methods for using the human perception of beauty in order to improve StyleGAN2 outputs. In GANAesthetic, UI sliders allow users to explore satellite images interactively, that is, visually appealing satellite images generated from StyleGAN2. The GANAesthetic was determined to be the most appropriate methodology for the study. The choice of GANAesthetic over other approaches will be explained in this paper, as well as its implementation. The paper will also describe an experiment to discover aesthetic latent dimensions.**

## 1 Introduction

Art has been known for ages as a means of communication and expressiveness in emotions [1]. The famous painting by Pablo Picasso called *Guernica* was used to convey the message about the horrors of wars, which emotionally influenced people back then and are still relevant today [2]. Visual art can be an effective way to convey a message about the threat of climate change due to the fact that most of the information heard in the media about this matter is perceived as an abstract issue and not as a direct experience [3].

In current times, the discussion about climate change has been more prevalent due to the fact that it is affecting our planet. The increase in carbon dioxide concentrations intensifies the greenhouse effect, raising sea levels by 0.24 to 0.32 meters by 2050 [4]. Biodiversity and climate are inextricably linked, implying that the warming climate of Earth has an effect on the wildlife and ecosystem by introducing phenological shifts and extinctions of species [5]. Despite all the warnings, there is still a lack of active engagement from the public about this topic. This phenomenon is happening due to the fact that people often do not experience an emotional response to ordinary climate communications [3]. To solve this, the project *Landshapes* by Frederik Ueberschar aims to show the impact of climate change by showing the aesthetically pleasing satellite landscape images generated by StyleGAN2-ADA [6]. The satellite images generated by the model can be seen in Figure 1.

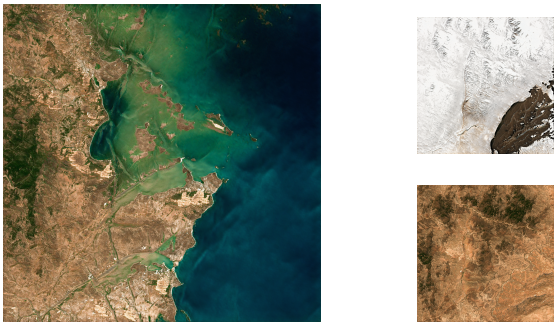

Figure 1: Satellite images from the project *Landshapes*

In recent years, there has been major development in AI, specifically in generative models such as generative adversarial neural networks (GAN), variational autoencoders (VAE) or TransGAN, where two transformers are used [7]. These models enable us to generate synthetic data with different variaties of features, from creating deepfakes, where the technique is to put the faces of public figures such as politicians or celebrities onto other people's faces, resulting in a high potential for deception, [8] to generating singing voices[9].

With the rise of GAN, the field of computational creativity has witnessed rapid progress. The system of generating art called Creative Adversarial Networks (CAN) is able to generate artistic paintings that would be creative rather than only emulating the images. That means generating images from the same distribution as the training set [10]. Studies have shown that moderately novel art attracts people more than art that is habitual since it reduces the arousal potential and desirability of that art [11]. The CAN attempts to generate artistic images that "do not have too little arousal potential and also

do not have too much because it activates the aversion system." [10].This is done by increasing *stylistic ambiguity* and deviating from style norms while not moving too far from what is accepted as an art.

After establishing and creating CAN, the same researchers conducted human subject experiments, taking human artists to evaluate AI-generated art and human-made art. The result of these experiments was that human artists could not distinguish the art generated by the CAN from the art created by contemporary artists based on likeness, novelty, complexity, ambiguity, and surprise.

Another study has created a survey experiment where they had 288 participants rate artworks by humans or an AI-based on variables such as originality, successful communication of ideas, etc. Human-created artworks received significantly higher ratings in composition, degree of expression, and aesthetic value. [12].This result indicates that there are objective differences in the art produced by humans and AI. This contradicts the study conducted by Elgammal et al. [10] where the number of participants was only 18 people. Therefore, they do not have enough statistical power to imply that the outcomes of the study are generalizable.

In the same experiment by Hong et al. [12] also discovered that participants that had negative bias towards AI gave lower artistic ratings to artworks created by AI compared to human-created artworks. To discover and explain the perceptual bias towards AI systems, one study conducted an experiment with 565 participants where the technique of priming effect was used [13]. Before starting the survey, it is indicated by whom the paintings were created. In this case, it was either an AI system or a human artist. For each painting, they were evaluated based on their likeness, beauty, novelty, and meaning. At the end of the survey, the manipulation check was introduced, where they asked the participants whether they remembered the identity of the painters. The origins of the paintings have been manipulated, and they had to guess the origins of the 4 paintings. The result of this study was that the artworks generated by the AI system were less well evaluated and significantly less liked than the paintings made by humans.

From these findings, it is evident that the idea of incorporating human rating data into the GAN training is to converge closer to the human perceptual beauty distribution. Thus, for example, AI systems by using human rating data will generate artistic images with higher composition, degree of expression, and aesthetic value ratings where these variables score higher for human-created artworks [12].

The overall goal is to make use of human rating data or human perception in order to improve the output of the GAN/Transformer so that it is aesthetically pleasing to the eye. This research specifically delved into finding the methodologies that could help GAN models generate more beautiful images by incorporating human perception. To accomplish such goals, there are a few research questions that need to be investigated:

1. *What existing methodologies could be used ?*
2. *How such methodologies could be useful for this particular research ?*
3. *What methodology is the best and most relevant for this research ?*
4. *Then why such methodology is the most preferable ? And how feasible is constructing such methodology ?*

Three exact methodologies have been found from the literature studies. The first one is to use humans as discriminators during adversarial training, the second is using deep learning models to assess whether the images are aesthetic or not; and lastly, creating an interactive aesthetic image editor with the interpretability learning of GANs, namely GANAesthetic.

Some methodologies that incorporate human perception of beauty have already been discovered, such as retraining GAN models multiple times on curated images by crowdworkers [14] or making use of visual and contextual features in order to create automatic aesthetic measures [15]. The two methodologies could be integrated into the GANAesthetic.

The contribution of this research is: (textiti) The exploration and analysis of other existing methodologies, such as using human-based discriminators during adversarial training, deep learning approaches for determining whether the images are aesthetically pleasing, and the interactive image editor with interpretability learning of GANs. (*ii*) The study on interactive image editing with GAN uses known UI components such as sliders to create the interface where users can edit synthesized images by modifying the sliders; perform an experiment where participants manipulate sliders with the semantics unknown and determine which ones contribute to the aesthetic the most.

This paper is structured as follows: Section 2 covers related work, in Section 3 individual methodologies and existing applications will be analyzed. The overview of the GANAesthetic pipeline and its implementation will be explained in Section 4. The experimental setup and the results will appear in Section 5. Section 6 reflects on the reproducibility and ethical considerations of this research. The discussion of the results from the experiments will be discussed in Section 7. The conclusion of the research and the future work are addressed in Section 8. StyleGAN2 training images are shown in Appendix A, while PCA latent dimensions are shown in Appendix B.

## 2 Related work

GAN has been a hot topic in the research since the first published paper by Ian Goodfellow et. al. in 2014 [16]. It contains two components, namely the generator and the discriminator. The generator synthesizes the fake data and the discriminator determines whether the data created by the generator belongs to the real data distribution. Both of these components are playing min-max adversarial game where the generator tries to fool the discriminator meanwhile the discriminator decides whether the data is real or fake.

$$\begin{aligned} &E_{x\sim p_{data}(x)}[logD(x)] \\ &\quad + E_{z\sim p_z(z)}[log(1-D(G(z)))] \end{aligned} \tag{1}$$

The above equation (1) denotes the loss function that equals to $\min_G \max_D V(D,G)$. During training we want for discriminator to maximize $logD(x)$, in other words we want to maximize the probability for the discriminator to determine correct

class for real samples and generated samples. Conversely, the generator should minimize $log(1 - D(G(Z)))$. The training ends whenever the generator's probability distribution is equal to the discriminator's probability distribution that is $p_{data} = p_{generated}$. This means that the discriminative probability distribution is equal to $\frac{1}{2}$ since it cannot longer distinguish between real and generated data.

If the desire is to generate the images of specific features, it is not possible with the normal GAN. One way to tackle this problem is to train the generator and discriminator to be conditioned on modalities like class label [17]. These type of GANs are known as Conditional GANs (CGAN).

To perform a conditioning, the $y$ serves as an additional information that is fed into both networks. The $y$ is concatenated with the vector $z$ from the random noise distribution $p_z(z)$ into the generator. For the discriminator, it receives two inputs from the real training data and $y$.

### *A. PG-GAN*

The Progressively-Growing GAN (PG-GAN) is using new training methodology where instead of training all the layers of the discriminator and generator at once, it starts from low resolution and progressively grow to higher resolutions by adding new layers during the training [18].

In the starting phase, the generator (G) and the discriminator (D) have low resolution of $4 \times 4$ images. As the training advances, the new layers are added to upsample the resolution to $8 \times 8$ in G and to downsample the resolution in D. This is repeated until the resolution of $1024 \times 1024$ is reached.

The result is that "it speeds the training up and greatly stabilizes it, allowing us to produce images of unprecedented quality" [18].

### *B. StyleGAN2-ADA*

PG-GAN provides high-resolution images with high image quality, however it lacks control over the synthesized images. Therefore, NVIDIA created the new improved extension of the PG-GAN called StyleGAN [19].

The focus of the StyleGAN is on introducing the style-based generator which offers "unsupervised separation of high-level attributes from stochastic variation in the generated images, and enables intuitive scale-specific mixing and interpolation operations" [19]. The important new feature to know is that instead of directly feeding the vector $z$ from the latent space $\mathcal{Z}$ into the synthesis network, it goes through the mapping network $f : \mathcal{Z} \rightarrow \mathcal{W}$ where the dimensionality of both latent spaces is 512. The network consists of 8 layers of multi-layer perceptron and generates style vector $w$, later fed into the synthesis network. This intermediate latent space allows us to have factors of variation more disentangled and will be relevant in Section 4.

The issue with the StyleGAN is that its generated images contained blob-like artifacts, this phenomena occurs at $64 \times 64$ resolution and get worse in higher resolution. The cause has been spotted in the instance normalization used in AdaIN. Thus the new StyleGAN2 was introduced and it solves the problem with weight modulation [20].

The project *Landshapes* was trained on StyleGAN2-ADA with roughly $4040$ high quality RGB images in $1024 \times 1024$ resolution. The satellite images were gathered through Google's Earth Engine[1] at random locations with using QGIS[2].

The reason of choosing StyleGAN to begin with is that it provides generation of high quality images in high resolution. The StyleGAN2-ADA provides training with limited data while avoiding discriminator overfitting by incorporating adaptive discriminator augmentation (ADA) [21].

## 3 Other Relevant Existing Approaches

In this section, the theoretical analysis and the description of both approaches will be conducted. The motivation behind on why using the approaches could be potentially useful on improving the beauty of the images outputted by GAN will be discussed here.

### 3.1 Human-based discriminator

Instead of using neural networks as discriminators we could use humans instead, the idea is that while determining whether the data generated from the generator is real or fake, humans can add an additional assessment such as how beautiful the images are.

The experiment of using this approach was conducted by Japanese researchers where crowdworkers evaluate the generated samples of how natural and human-like the synthesized speech is [22]. The difference between training a basic GAN and HumanGAN is that the HumanGAN is able to capture the human perception distribution which in theory is wider than the real-data distribution with training computer-based discriminators.

The human-based discriminator outputs a posterior probability about "to what degree is the input perceptually acceptable" where the value ranges from 0 to 1. The function during training can be seen in (2).

$$V(G, D) = \sum_{n=1}^{N} D(G(z_n)) \tag{2}$$

For training the generator, the gradient-based iterative method is used. However, the problem is that the discriminator function is not differentiable since it is replaced by humans. To solve the issue, the natural evolution strategies (NES) fix the problem by computing the approximate gradients by using data perturbations [23].

The result of the experiment can be seen in Figure 2, the darker area denotes lower probability range and vice versa. With using vanilla GAN, the generator is trained on the certain set of real data which implies it only covers the trained data distribution. The HumanGAN demonstrated that it represents the human's perception distribution which is much wider than the real data distribution. This can be applied to the case of discovering human's perceptual aesthetic distribution.

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

[2] https://www.qgis.org/en/site/

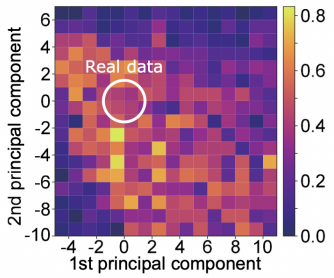


Figure 2: Color map representing posterior probabilities in speech naturalness [22]

| Performances of the models on certain datasets | | | |
|---|---|---|---|
| Models | AVA1 | AVA2 | CUHKPQ |
| VGG-16$_{\text{Composite}}$[30] | - | 85.40% | - |
| Resnet-50$_{\text{Composite}}$[30] | - | 90.01% | 94.1% |
| FCN$_{\text{Croppings+Skips}}$ [31] | - | 91.01% | - |
| TCN [32] | 82.3% | - | - |
| ILGNet [33] | 79.25% | 85.62% | - |

Table 1: Evaluation and comparison of models

### 3.2 Deep learning approach for human rating data prediction

There was an annual image recognition software contest called ILSVRC where it contained roughly around 1.2 million images for training the machine learning model [24]. Few neural networks have shown significant improvement in performance such as AlexNet, VGG-19 and resnets [25; 26; 27]. With using these neural nets, we can create models that could predict human ratings on the set of images.

There are already some existing research and developments on this topic. Most of these neural networks are binary classifiers that predict whether the image is aesthetic or not.

*A. Datasets*

The AVA [28], which stands for Aesthetic visual analysis, is the largest set of images that are rated, from 0 to 10, based on the aesthetic quality of the images. This dataset contains around 250 000 images and there are two approaches of dividing this dataset into higher and lower aesthetic quality images for training and testing, namely AVA1 and AVA2.

1. With AVA1, the score of 5 is the threshold that distinguishes the low and high aesthetic quality of the images. Meaning that the images rated from 0 to 4 are considered low quality and from 6 to 10 are considered high quality. The images that are rated as 5 are omitted from the training and testing the model.

2. The other technique of splitting the images into training and testing set called AVA2 is to sort the images according to their mean aesthetic quality score. The way of dividing the images into high aesthetic quality and low aesthetic quality is by taking top 10% of the mean score and label them as high quality and likewise for the low aesthetic quality by taking bottom 10%.

Some researcher also used CUHKPQ [29] for training which is the dataset of images that are assessed by professional photographers that take into account composition lighting, color arrangement. camera setting and topic emphasis.

*B. Deep learning models*

Couple of classifiers with state-of-art performances on the aesthetic assessment can be seen in Table 1. The models were evaluated on the testing set from the AVA1, AVA2 and CUHKPQ datasets and evaluted with an accuracy metric which measures the ratio of correct predictions over the total number of instances evaluated.

With Triple Column Network (TCN), it works with 3 channels that perform an transformations on the images such as cropping images or using saliency maps and then later concatenated and inputted into the network [32]. The ILGNet uses multiple inception modules with directly connecting the layers of local features to the layer of global features [33]. Both were evaluated on the AVA1 dataset and it is evident that TCN performs better with an accuract of 82.3%.

Both Resnet-50$_{\text{Composite}}$ and FCN achieved the highest accuracies with 90.01% and 91.01%, respectively. The Fully-Connected network uses VGG16 architecture with added skip connections and accepting a triple of image croppins as input [31]. Residual neural networks are notorious of dealing with deep networks with introducing skip connections which deal with vanishing gradients and mitigate the Degradation problem [27]. As we can see here, both highly performed models on AVA2 used skip connections.

Additionaly, the Resnet-50$_{\text{Composite}}$ which uses 3 neural networks running in parallel to extract unique features from three different aspects, including the global view, local view and scene-aware information and in the end aggregate them as composite features for Support Vector Machine (SVM) to classify [30]. It achieved an accuracy of 94.1% on CUHKPQ dataset.

The motivation behind using deep learning approach for human rating data is being able to assess the satellite images and classify them either as the high or low aesthetic images without performing complex feature engineering such as understanding visual (i.e. saturation, luminance and etc.) and contextual features (i.e. novelty and typicality). Neural networks are often in the literature referenced as ”black-box models” due to the fact that it is merely impossible to get the approximation function and are well-known as non-identifiable models. With StyleEx, it is possible to understand the decisions of the aesthetic classifier by discovering and visualizing multiple factors of variation that affect its prediction [34]. This is useful for the factor analysis.

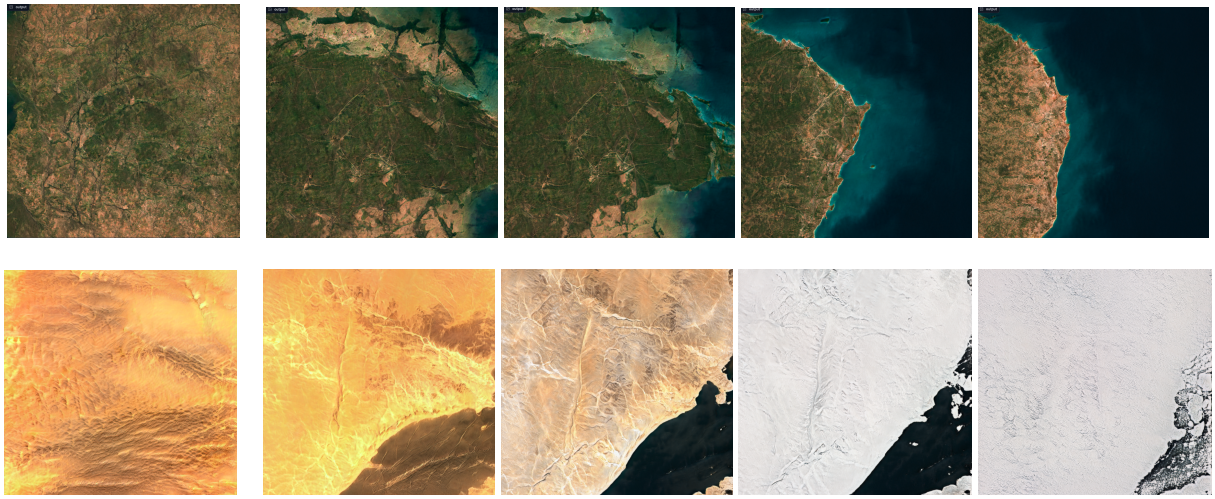

Figure 3: The images are arranged in two rows, showing traversals in a particular latent direction with layer-wise editing. The first row is $E(v_1, 2-5)$ and the second row is $E(v_4, 1-18)$. The explanation of the notation can be seen in the equation 4.

## 4 Methodology

This section will discuss the approach of determining linear functions, that maximizes an amount of variance, for each latent dimensions and bring an overview of GANAesthetic.

### *A. Identifying important latent directions with GANSpace*

There is various research done on how to interpret learned representations of the deep generative models with state-of-art methods [35]. From the literature study, there are two approaches to interpretability learning of GANs, namely supervised learning and unsupervised learning.

The standard way to interpret the GAN latent dimensions in a supervised manner is to use modalities such as class labels or attribute predictors to probe the representation of generators [36].

With an unsupervised approach, no training and off-the-shelf classifiers are needed in order to define the attributes in the images, which can be a time-consuming process and require expensive supervision. Instead, the solution is to identify interpretable latent dimensions using simple mathematical techniques [37; 38].

The particular unsupervised technique that will be explored is GANSpace, which uses PCA to identify important latent directions [37].

In Section 2, the intermediate latent space $\mathcal{W}$ was described as a space with disentangled factors of variation. The synthesis network of StyleGAN2 consists of 18 layers where each layer contains its own $w_i$ to "enable powerful *style mixing*, the combination of features of various abstraction levels across generated images" [19; 37]. In other words, the synthesis network contains $L$ $(L = 18)$ intermediate generator layers $G_1...G_i...G_L$ where the output of each layer is defined as $y_i = G_i(y_{i-1}, w)$. From studies of the interpretability learning of GANs, it is observed that different layers of the synthesis network produce different semantics in terms of abstraction level [19; 37; 38; 39]. For example, with the FFQHQ dataset, the shallow layers of the synthesis network bring high-level aspects such as pose, general hair style, and face shape, while the deeper layers control lower-level features such as color scheme and microstructure [19].

In GANSpace, to identify important latent semantic directions is to use Principal Component Analysis (PCA) which is commonly used as a dimensionality reduction method [37]. To identify important latent directions is by finding new variables known as principal components that are linear functions maximizing the amount of variance and are uncorrelated with each other [40].

The procedure starts with sampling $n$ random latent vectors $z_{1:n}$ from the noise distribution $p(z)$ which are then inputted to the mapping network to get the corresponding $w_i = M(z_i)$ values. The next step is to perform PCA on the $w_{1:n}$ to get eigenvectors that form basis $\mathbf{V}$ with $v_{1:m}$ principal components for the latent space $\mathcal{W}$. This gives us an equation described in (3) where the PCA coordinates $x$ are edited before feeding $w'$ into the synthesis network. The entries of $x$ are initially equal to 0 until they are modified by the user.

$$w' = w + \mathbf{V}x \tag{3}$$

The paper also introduces layer-wise edits where only certain layers have their modified $w$, leaving other layers' inputs unchanged [37]. The notation of the usage of particular principal component $v_i$ from layer $j$ to $k$ is shown in (4).

$$E(v_i, j-k) \tag{4}$$

The example of the traversals along principal axes are shown in Figure 3. The important to note that the maximum amount of principal components is 512 since that is the dimensionality of intermediate latent space $\mathcal{W}$. This implies that the basis $\mathbf{V}$ is a full-rank matrix that has dimension of $512 \times 512$.

*B. GANAesthetic*

The oveview of the GANAesthetic pipeline can be seen in Figure 4. The procedure starts with data collection and training StyleGAN2. Afterwards, using GANSpace in order to obtain important semantic latent directions and with Gradio and Google Collab to construct the user interface with the sliders. Finally, experiments with incorporating human perception are conducted.

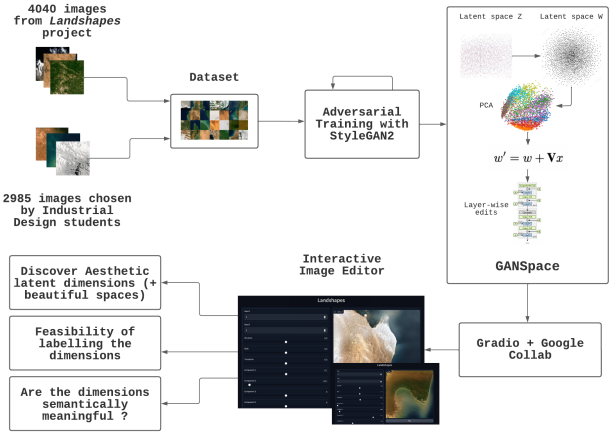


Figure 4: An overview of GANAesthetic pipeline

*C. Why GANAesthetic is the most preferred over other methodologies ?*

Choosing GANAesthetic over human-based discriminators and aesthetic classifiers is because it aids in discovering beautiful regions in the latent spaces and serves as a tool to help cognitive scientists and design engineers study factors contributing to the aesthetic or study semantic representations in perceptual spaces. In addition, it fits well with the agenda to create a fascination with the climate in users through interactivity.

Improving generated images by iterative retraining of StyleGAN2 with human rated images has proven to be effective [14]. Constructing GANAesthetic for improved models will enhance the experience of interactively exploring beautiful satellite images.

The Gibb's sampling with people is the technique that could be applicable for extracting semantic representations from high-dimensional latent spaces [41]. It is a continuous-sampling paradigm that presents the human participants with sliders representing certain semantic latent dimensions, in this case, PCA components.

To find the best latent dimensions and the range of coordinate values that characterize the aesthetic, participants move sliders $(s_1, ..., s_n)$ to select combinations of sliders and their values that maximize the aesthetic's utility. The utility value is calculated by the formula $U_m = l_m + n_m$, where $l_m$ represents the utility for stimulus $m$ (in this case, $m$ = aesthetic) and $n_m$ represents the noise component that contains participant-level noise resulting from sensory and cognitive processes, as well as population-level noise resulting from individual differences in utility functions.

The Gibb's sampling itself is the sampling from the $n$-dimensional probability distribution and uses the method of Marko Chain Monte Carlo (MCMC). This implies that it is a time-dependent sampling algorithm. Let $p(z_1, ..., z_n)$ be a target distribution over an $n$-dimensional state space from which it is desired to sample from, the Gibb's sampler starts with the vector state z $= (z_1^i, ..., z_n^i)$, where $i = 1$. Afterwards, iteratively conditionally update the coordinates by sampling from $p(z_k^{i+1} | z_1^{i+1}, ..., z_{k-1}^{i+1}, z_{k+1}^i, ..., z_n^i)$. The coordinate updates are done by the participants instead of having conditional class probability, where the participants move the sliders $(s_1, ..., s_n)$ corresponding to the latent dimensions $(z_1, ..., z_n)$ to maximize the aesthetic of the synthesized images [41].

$$p(aesthetic) = p(z_k^i | z_{-k}) = \frac{e^{\gamma l(z_k^i, z_{-k})}}{\sum_j \gamma l(z_k^j, z_{-k})} \tag{5}$$

The equation that represents the probability distribution over slider locations can be seen here (5). Each of the points on the slider is described as $z_k^i$ and the other fixed dimensions as $z_{-k}$. The utility value of each point on the slider is associated with a utility such as $U_m = l(z_k^i, z_{-k}) + n_m$ with $m$ = aesthetic and the noise being Gumbel distributed $n_m \sim Gumbel(\mu, \gamma^{-1})$ [41].

This approach could potentially help to discover beautiful perceptual regions in the latent spaces, where we let users travel in those spaces to evoke climate fascination. Response surface methodology (RMS) can also be used to visualize the aesthetically-pleasing perceptual regions in the latent spaces by collecting human rating data on the factors and their values.

## 5 Experimental Setup and Results

This section outlines the implementation of parts in the GANAesthetic pipeline in Figure 4 and conducting an experiment on discovering aesthetic latent dimensions with human participants.

### 5.1 Training StyleGAN2 with Satellite images

With StyleGAN2-ADA[3], NVIDIA switched from Tensorflow to Pytorch because training performance with Pytorch on the NVIDIA Tesla V100 GPUs is 5%–30% faster than the Tensorflow version, and inference is up to 35% faster in high resolutions, which is desirable in the context of *Landshapes*.

Unfortunately, it is not possible to convert from network pickle files (.pkl) for StyleGAN2-ADA to Pytorch format (.pt) when converting weights for the GANSpace to use. The solution for this is to train StyleGAN2 model in Tensorflow and then use the script that converts weights into Pytorch format.

[3]https://github.com/NVlabs/stylegan2-ada-pytorch

The model is trained with 4040 satellite images in $1024 \times 1024$ resolution generated from the Google Earth Engine and 2985 out of 6000 images, randomly sampled from the pre-trained *Landshapes* model, were chosen as aesthetically pleasing by 4 students from the Industrial Design faculty of TUDelft. This implies that the StyleGAN2 model is trained on 7025 high quality RGB images. All of the training is done in Google Collab, a free cloud service provided by Google that allows access to NVIDIA GPUs such as the T4 or P100.

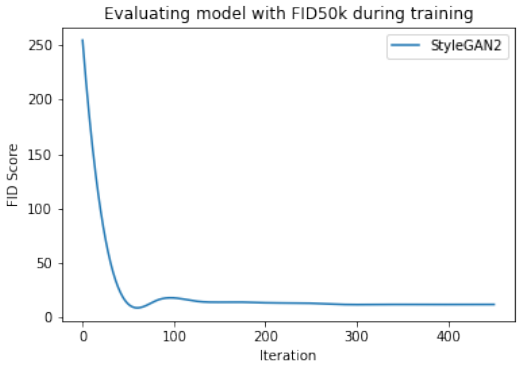


Figure 5: Training performance of StyleGAN2 with FID metric

The metric used is Fréchet inception distance (FID), which quantitatively assesses the quality of the images based on a comparison of the distribution of the generated images with the distribution of the real images. The training process can be seen in Figure 5. The outputs of the StyleGAN2 can be seen in Appendix A, the best and final FID score is 11.5706.

## 5.2 Interactive Image Editor

To create GANAesthetic[4] interface that users can edit synthesized images with the sliders, Gradio[5] offers an easy and fast way to create an app for demo of machine learning models with nice and friendly interface. The GANSpace[6] is used for identifying latent semantic directions in unsupervised manner provides and provides an existing implementation in Python which is then integrated with Gradio.

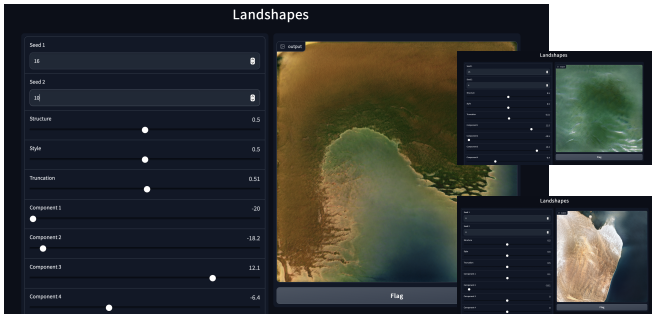


Figure 6: Interactive Image Editor with StyleGAN2

[4]https://github.com/HahaBill/ganaesthetic-landshapes
[5]https://gradio.app
[6]https://github.com/harskish/ganspace

The generated sliders refer to each of the principal components, where using multiple sliders implies changing the coordinates $x$ (see equation 3). The created interface can be seen in Figure 6, where each of the components implies the principal components, and with changing seeds, it is possible to generate different synthesised images and perform editing on those.

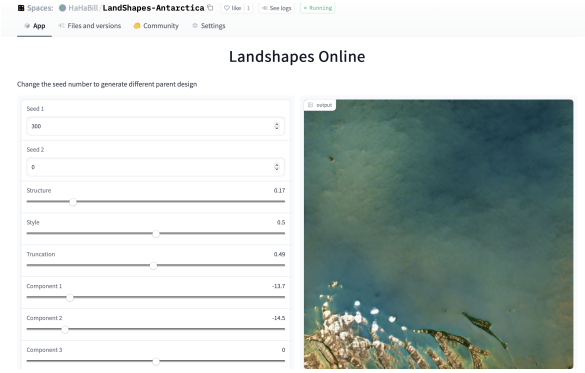


Figure 7: Interactive Image Editor on HuggingFace

The interactive image editor can be hosted on the server with using HuggingFace[7], it works with Gradio which can be seen in the Figure 7. On the HuggingFace, using the sliders and generating synthesized photos takes roughly 10 seconds, and on Google Collab, it takes around 1 second with NVIDIA GPU T4 or P100.

## 5.3 Discover aesthetic latent dimensions

Three different images of seeds generated by the StyleGAN2 are used in the experiment. The images are in Appendix B, where the first image depicts a *coastline area*, the second a *forest and desert area*, and the third depicts an *arctic area*.

Each of the images shows ordered rows from $c_0$ to $c_{13}$ denoting principal components. The sequences of the synthesized images are changing of $x$ coordinates in the range of $[-2, 2]$.

The aim of experiment is to discover what semantic latent dimensions corresponds to the aesthetic the most. The participants were provided with 3 images and instructed to pick 3 rows of the images that they thought were most beautiful/pleasing to the eye. As such, the format for the answers was: "(Image 1: C1, C2, C3), (Image 2: C1, C2, C3), (Image 3: C1, C2, C3)".

The experiment's results are shown in Table 2, and the total number of participants was 56. It can be noticed that $c_3$ was the winner followed by $c_1$, $c_4$, $c_6$ and $c_9$ as the most votes for the COASTLINE. For the FOREST/DESERT, $c_0$ came out on top, with $c_1$, $c_4$ and $c_7$ coming in second, third and fourth, respectively. Finally, in the ARCTIC, the score$c_3$ and $c_4$ are equal, additionally $c_2$ and $c_5$ were ranked among the highest.

[7]https://huggingface.co

Table 2: A total of 56 participants picked 3 principal components of each of the images (COASTLINE, FOREST/DESERT, ARCTIC) that were most beautiful or pleasing to the eyes.

| | COASTLINE | FOREST/DESERT | ARCTIC |
|---|---|---|---|
| $c_0$ | 9 | 41 | 4 |
| $c_1$ | 17 | 29 | 6 |
| $c_2$ | 7 | 8 | 17 |
| $c_3$ | 31 | 3 | 20 |
| $c_4$ | 16 | 28 | 24 |
| $c_5$ | 7 | 4 | 24 |
| $c_6$ | 18 | 2 | 8 |
| $c_7$ | 8 | 21 | 4 |
| $c_8$ | 6 | 12 | 11 |
| $c_9$ | 18 | 5 | 14 |
| $c_{10}$ | 9 | 5 | 11 |
| $c_{11}$ | 5 | 6 | 6 |
| $c_{12}$ | 10 | 1 | 4 |
| $c_{13}$ | 4 | 0 | 11 |

# 6 Responsible Research

Integrity and reproducibility are addressed in this section. The experiment and its findings represent high-level of integrity, ethics and transparency. The explanation of why there is a need for reproducibility of research methods and why it plays a crucial part in the scientific community are discussed in this section.

## 6.1 Integrity

The Dutch scientist Diederik Stapel has been exposed for his wrongdoings by manipulating collected data and fabricating experiments. At least 30 research papers used fraudulent data that he provided and thus unfortunately negatively affected multiple researches and studies [42].

The research integrity is crucial for strengthening the validity of the research and providing trust and confidence in the methods used, the findings in the results and absence of data trimming and falsification. The negative results should be highlighted and published otherwise nothing is learned and scientists will repeat failed experiments [43].

In this study, the methodology for an experiment to discover aesthetic latent dimensions was clearly defined. In addition to putting in their time, participants were not asked for any personal information and no attention checks were performed. While the participants' cultural backgrounds were diverse, most of them were between the ages of 21 and 26. As a result, the collected data may reveal a generational bias on what is considered beautiful. There was no modification or fabrication of the collected data, and it is all publicly available on GitHub.

The StyleGAN2 that was trained on the satellite images from the Google Earth Engine was evaluated with FID, which is a widely used metric in the literature in the context of generative models.

GANSpace provided the source code and documentation on how to calculated important latent directions with PCA. The code was used and adapted with Gradio in order to create interactive image editor.

## 6.2 Reflection on Reproducibility

Reproducibility in computational research is essential since it allows next generation of scientists to build on the previous generations' achievements and improve the existing system, additionally it prevents credibility crisis [44].

In recent years, artificial intelligence has been widely discussed and booming in the industries, however, there is still a problem of replication crisis, including the difficulty of reproducing the results of a publication due to missing source code, training data, and hyperparameter settings [45].

Thus, it is my duty as a responsible researcher to ensure that all the results and experiments are fully reproducible in order for the scientific community to check their credibility. The resources and implementations can be found in my GitHub page [8] and an interactive image editor on HuggingFace [9]. The GitHub repository holds all the python scripts necessary to develop the interactive image editor, collected human rating data from the experiment and links to Google Collab for faster performance of the interactive image editor. When sampling latent vectors in order to perform PCA, it downloads the StyleGAN2 model from the Google Drive. NVIDIA's script for training StyleGAN2 is also available in the repository. The training dataset of satellite images was generated from the Google Earth Engine, which is publicly available on the internet.

Deep learning models have been well-known for having a great number of parameters and hyperparameters. It is important to note that machine learning frameworks does not guarantee fully reproducible results [46; 47; 48]. CUDA convolutions operation and using GPUs can be the source of the randomness in the training. The solutions are to use CPU-only training or set a global random seed.

# 7 Discussion and Future Improvements

The purpose of this section is to analyze and discuss the results presented in Section 5 with the possibility of suggesting improvements and addressing future work.

*A. Assessment of StyleGAN2 model*

450 iterations (kimg) produced a FID score of 11.5706 for our final StyleGAN2 model with satellite images. Appendix A shows the synthesized images, where it can be seen that some of them contain anomalies such as black blobs, as seen in the top-leftmost image, or unnaturally looking areas on the landscape images. Based on the original StyleGAN2 paper, the FFHQ dataset was trained with $1024 \times 1024$ resolution and 70k images with a final FID of 2.84. The LSUN Car dataset with $512 \times 384$ resolution and 893k images was trained with StyleGAN2 and achieved the FID score of 2.32

[8] https://github.com/HahaBill/ganaesthetic-landshapes

[9] https://huggingface.co/spaces/HaHaBill/LandShapes-Antarctica

[20]. It is evident from these findings that the improvements would require a greater amount of training data and a longer training period.

Another metric that could be used is called Perceptual path length (PPL) which measures how entangled are the factors of variation since "interpolation of latent-space vectors may yield surprisingly non-linear changes in the image" [19]. This metric is certainly appropriate for our case, since each latent dimension can indicate a particular factor or feature, but it results in a 42-minute evaluation time for one GPU. On the other hand, FID takes 21 minutes on a single GPU. For the experiment of discovering aesthetic latent dimensions, the model required to be trained as soon as possible.

*B. Performance of Interactive Image Editor*

The interactive picture editor's biggest flaw is that it takes a long time to generate synthesized images when you move the sliders. It took 10 seconds on HugginFace. This was improved by using a Google Collab interface to employ one NVIDIA GPU to accelerate an inference from the synthesis network, with an average time of roughly 1 second. This might be improved by using additional GPUs or transforming each of the slider frames into videos and storing them in the database.

*C. Analysis of the experiment of discovering aesthetic latent dimensions*

The results of the experiment can be seen in Table 2. With the COASTLINE, $c_1, c_3, c_4, c_6$ and $c_9$ were among the highest scored latent dimensions for the aesthetic. In the first place, $c_3$ is transitioning from green and teal waves with light brown terrain to more of an open sea with green vegetative land. $c_6$ and $c_9$ are similar with $c_3$, however with different structure. Some participants favored the golden brown terrain with the $c_1$.

Continuing on to the next image, FOREST/DESERT, the transition from the earthy regions to the water mass accounts for the majority of the highest-scoring principal components $c_0$, $c_1$, $c_4$ and $c_7$. By creating additional waves in the sea area with a wide piece of land with some residents, the component $c_0$ won the majority. In addition, the red orange land can be found in all four components.

The result of scores for the principal components in context of ARCTIC were more spread out. There is the most scored highly cluster consisting of $c_2, c_3, c_4$ and $c_5$, additionally the lower scored cluster of $c_8, c_9$ and $c_{10}$.

It was overwhelming and mentally exhausting for some participants to look at and process all the latent dimensions. The result is similar to the result of GANSlider, where participants were given an image and asked to reconstruct it with sliders [49]. The finding of the study was that with increased number of sliders (latent dimensions) implies significantly higher task difficulty, workload and user actions. The conclusion of the study was that it is recommended to use at most 3-5 sliders.

Overall, the experiment has shown that certain dimensions, clusters of dimensions and factors such as water mass are more preferred and pleasing to the eyes than the others.

## 8 Conclusions and Future work

The main goal of this research is to use methodologies that incorporate human rating data and perception of beauty to make GAN outputs more aesthetically pleasing. Several methodologies have been explored, from the human-based discriminator method to predicting human rating data on a set of images with aesthetic classifiers. After the literature study of all possible options, the most useful and relevant methodology for the study is to create an interactive image editor of StyleGAN2 by moving the UI sliders. This led to the creation of the GANAesthetic, which is a pipeline for the development of a system that delivers an experience of interactively exploring aesthetically pleasing satellite imagery from the project *Landshapes* and discovering aesthetic latent dimensions, beautiful regions in the latent spaces, semantic representation of those dimensions, etc.

A novel method of interpreting and discovering important semantic latent directions, GANSpace, was explored and, with the help of its available source code, implemented. The Gradio allowed us to create a nice and friendly user interface that was integrated with the GANSpace and used trained StyleGAN2 to generate and edit satellite images with UI sliders. The findings of the experiment of discovering aesthetic latent dimensions with the human perception of beauty were that some dimensions are more aesthetically appealing than others and increasing number of controllable sliders result in increase workload and mental exhaustion.

The GANAesthetic was shown in this paper to be successful and has a large potential for future scientists and designers to perform experiments and to study semantic representations or discover beautiful regions in perceptual spaces.

Future work will include implementing and carrying out Gibb's sampling with people in order to identify ranges of coordinates in each dimension that represent aesthetic values or semantic representations of the dimensions such as color, contrast, geometrical modification, etc. Using ranges of coordinates obtained from the Gibb's sampling to find out the correlation between automated measures of aesthetic beauty and human perceptual beautiful spaces. Additionally, a resource surface methodology (RSM) allows us to study the relationship between several latent dimensions and the response variable, in this case, aesthetic.

## References


[1] Melissa Dolese, Aaron Kozbelt, and Curtis Hardin. Art as communication. *The International Journal of the Image*, 4:63–70, 01 2014.

[2] Richard Rhodes. Guernica: Horror and inspiration. pages 1–25, 11 2013.

[3] Liselotte Roosen, Christian Klöckner, and Janet Swim. Visual art as a way to communicate climate change: a psychological perspective on climate change–related art. *World Art*, pages 1–26, 10 2017.

[4] Rebecca K. Priestley, Zoë Heine, and Taciano L. Milfont. Public understanding of climate change-related sea-level rise. *PLOS ONE*, 16(7):1–12, 07 2021.

[5] Sattar Quratulann, Maqbool Ehsan, Ehsan Rabia, and Akhtar Sana. Review on climate change and its effect on wildlife and ecosystem. *Open Journal of Environmental Biology*, pages 008–014, 08 2021.

[6] Frederik Ueberschär. Ai for experience: Designing with generative adversarial networks to evoke climate fascination. pages 1–156, 02 2021.

[7] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up, 2021.

[8] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. Deep learning for deepfakes creation and detection: A survey, 2019.

[9] Feiyang Chen, Rongjie Huang, Chenye Cui, Yi Ren, Jinglin Liu, and Zhou Zhao. Singgan: Generative adversarial network for high-fidelity singing voice generation, 2021.

[10] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. Can: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms, 2017.

[11] D. E. Berlyne. Aesthetics and psychobiology. *Journal of Aesthetics and Art Criticism*, 31(4):553–553, 1973.

[12] Joo-Wha Hong and Nathaniel Ming Curran. Artificial intelligence, artists, and art: Attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Trans. Multimedia Comput. Commun. Appl.*, 15(2s), jul 2019.

[13] Martin Ragot, Nicolas Martin, and Salomé Cojean. Ai-generated vs. human artworks. a perception bias towards artificial intelligence? In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–10, New York, NY, USA, 2020. Association for Computing Machinery.

[14] Irmak Celebi. Improving gan generated image aesthetics through iterative training with human rated images: Effects of the choice of dataset size and the number of training iterations, 2022.

[15] Joseph Catlett. Beauty in the eye of machine: Using automated measures of aesthetic beauty to improve gan output of satellite images, 2022.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[17] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2017.

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2018.

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2019.

[21] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data, 2020.

[22] Kazuki Fujii, Yuki Saito, Shinnosuke Takamichi, Yukino Baba, and Hiroshi Saruwatari. Humangan: generative adversarial network with human-based discriminator and its evaluation in speech perception modeling, 2019.

[23] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information, 2018.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2014.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[28] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012.

[29] Xiaoou Tang, Wei Luo, and Xiaogang Wang. Content-based photo quality assessment. *IEEE Transactions on Multimedia*, 15(8):1930–1943, 2013.

[30] Xin Fu, Jia Yan, and Cien Fan. Image aesthetics assessment using composite features from off-the-shelf deep models. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3528–3532, 2018.

[31] Konstantinos Apostolidis and Vasileios Mezaris. Image aesthetics assessment using fully convolutional neural networks. In *MMM*, 2019.

[32] Nishi Doshi, Gitam Shikhenawis, and Suman K Mitra. Image aesthetics assessment using multi channel convolutional neural networks, 2019.

[33] Xin Jin, Jingying Chi, Siwei Peng, Yulu Tian, Chaochen Ye, and Xiaodong Li. Deep image aesthetics classification using inception modules and fine-tuning connected

layer. In *2016 8th International Conference on Wireless Communications Signal Processing (WCSP)*, pages 1–6, 2016.

[34] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, and Inbar Mosseri. Explaining in style: Training a gan to explain a classifier in stylespace, 2021.

[35] Bolei Zhou. Interpreting generative adversarial networks for interactive image generation, 2021.

[36] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. *arXiv preprint arXiv:1906.10112*, 2019.

[37] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls, 2020.

[38] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans, 2020.

[39] Zhenliang He, Meina Kan, and Shiguang Shan. Eigengan: Layer-wise eigen-learning for gans, 2021.

[40] Sidharth Mishra, Uttam Sarkar, Subhash Taraphder, Sanjoy Datta, Devi Swain, Reshma Saikhom, Sasmita Panda, and Menalsh Laishram. Principal component analysis. *International Journal of Livestock Research*, page 1, 01 2017.

[41] Peter M. C. Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. Gibbs sampling with people, 2020.

[42] Mieke Verfaellie and Jenna McGwin. The case of diederik stapel, 2011.

[43] Devang Mehta. Highlight negative results to improve science. 2018.

[44] V.C. Stodden. Reproducible research: Addressing the need for data and code sharing in computational science. *Computing in Science and Engineering*, 12:8–13, 01 2010.

[45] Matthew Hutson. Artificial intelligence faces reproducibility crisis. *Science (New York, N.Y.)*, 359:725–726, 02 2018.

[46] Pytorch. Reproducbility. https://pytorch.org/docs/stable/notes/randomness.html, 2019.

[47] DeterminedAI. Reproducbility. https://docs.determined.ai/0.12.12/topic-guides/reproducibility.html#reproducibility, 2020.

[48] Sewade Ogun. Cross validation and reproducibility in neural network training. https://ogunlao.github.io/2020/05/08/cross-validation-and-reproducibility-in-neural-networks.html#cross-validation-applied-to-neural-network, 2020.

[49] Hai Dang, Lukas Mecke, and Daniel Buschek. GANSlider: How users control generative models for images using multiple sliders with and without feedforward information. In *CHI Conference on Human Factors in Computing Systems*. ACM, apr 2022.

# A Generated images from StyleGAN2

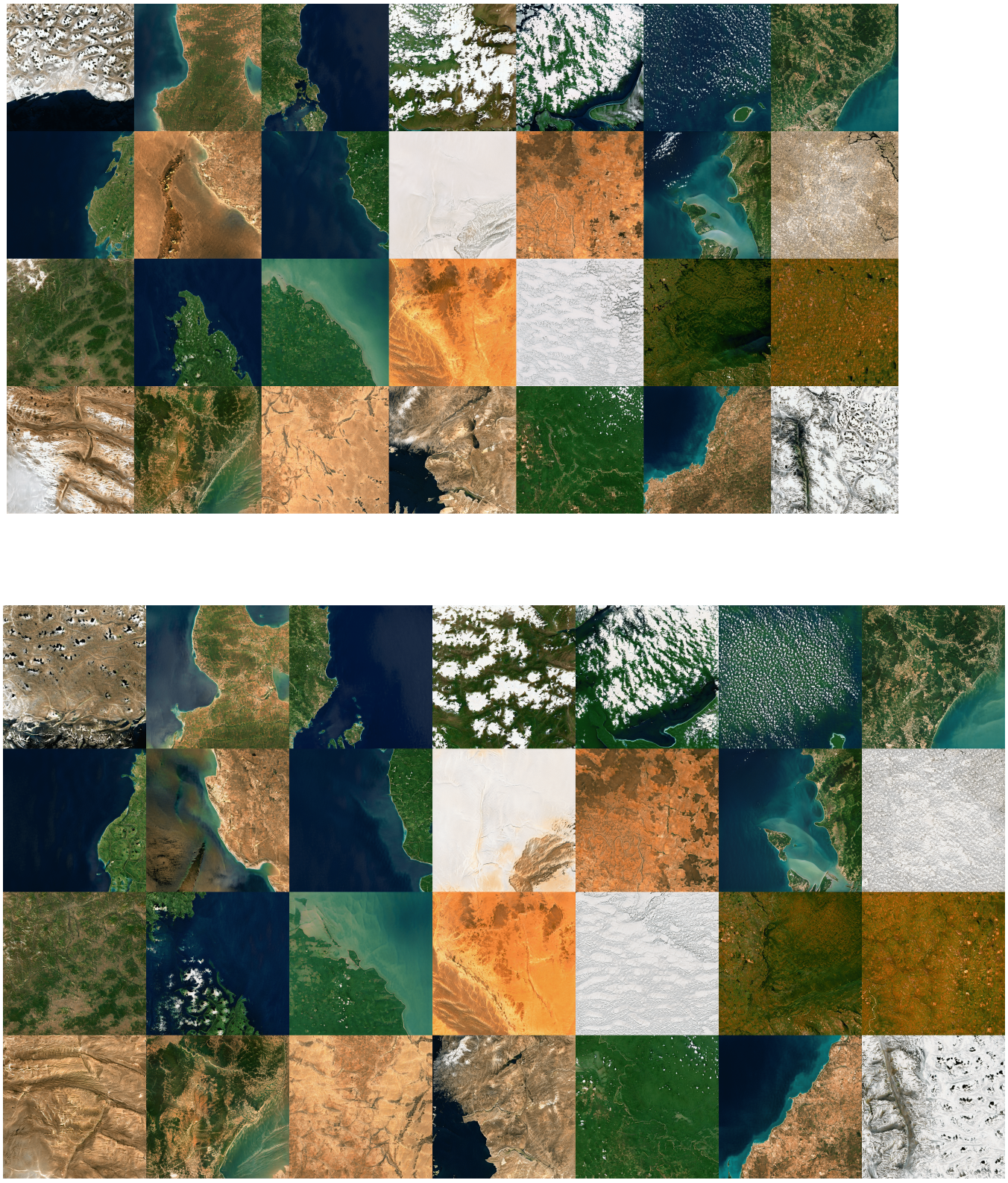

Figure 8: Images generated from the StyleGAN2 model from the training, the first image corresponds to the model with FID score of 23.3241 and the second image with FID score of 11.5706. The second image shows better improvements than the first image such as more structure, water lines, density of clouds and etc.

## B PCA latent dimensions

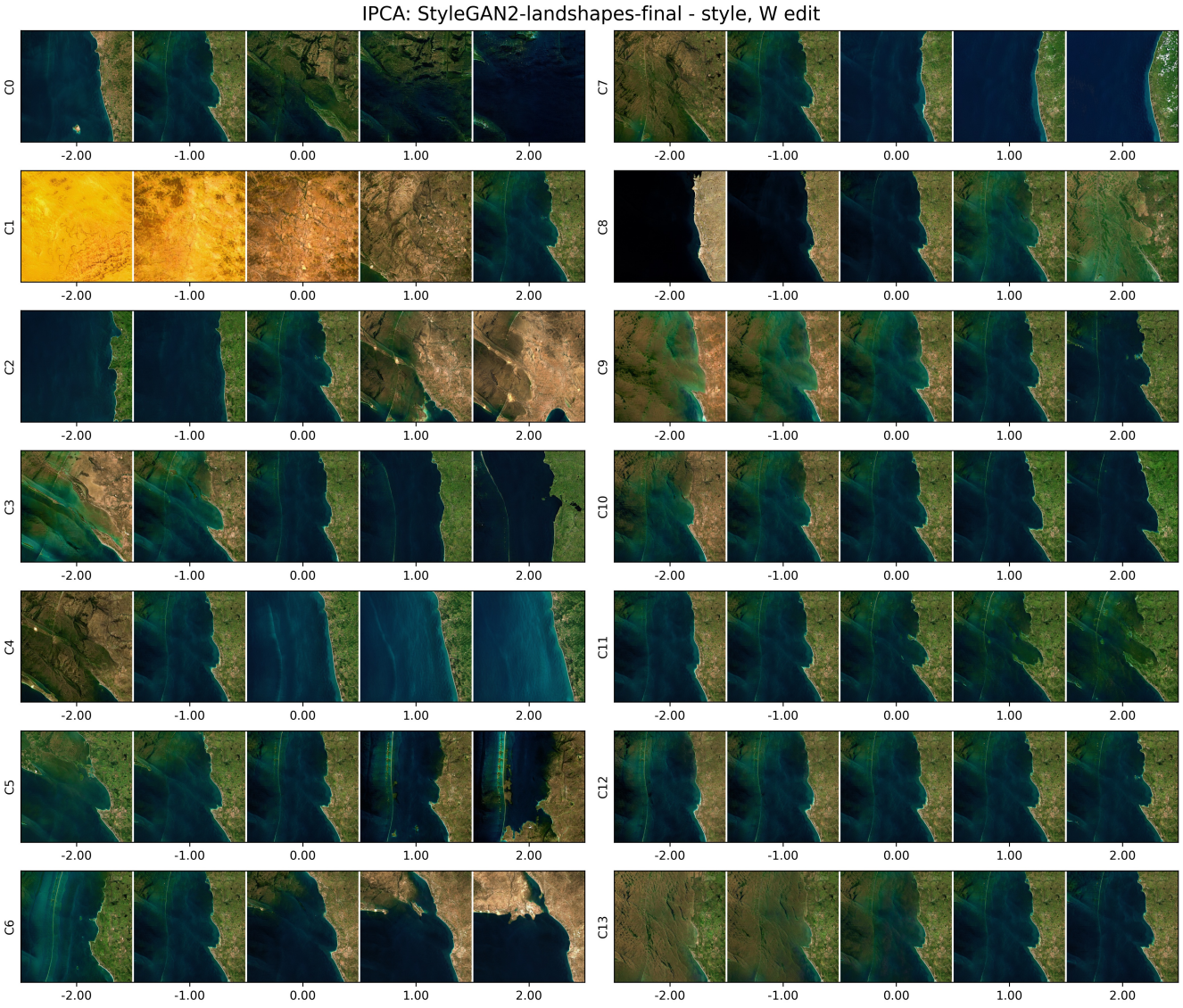


Figure 9: COASTLINE

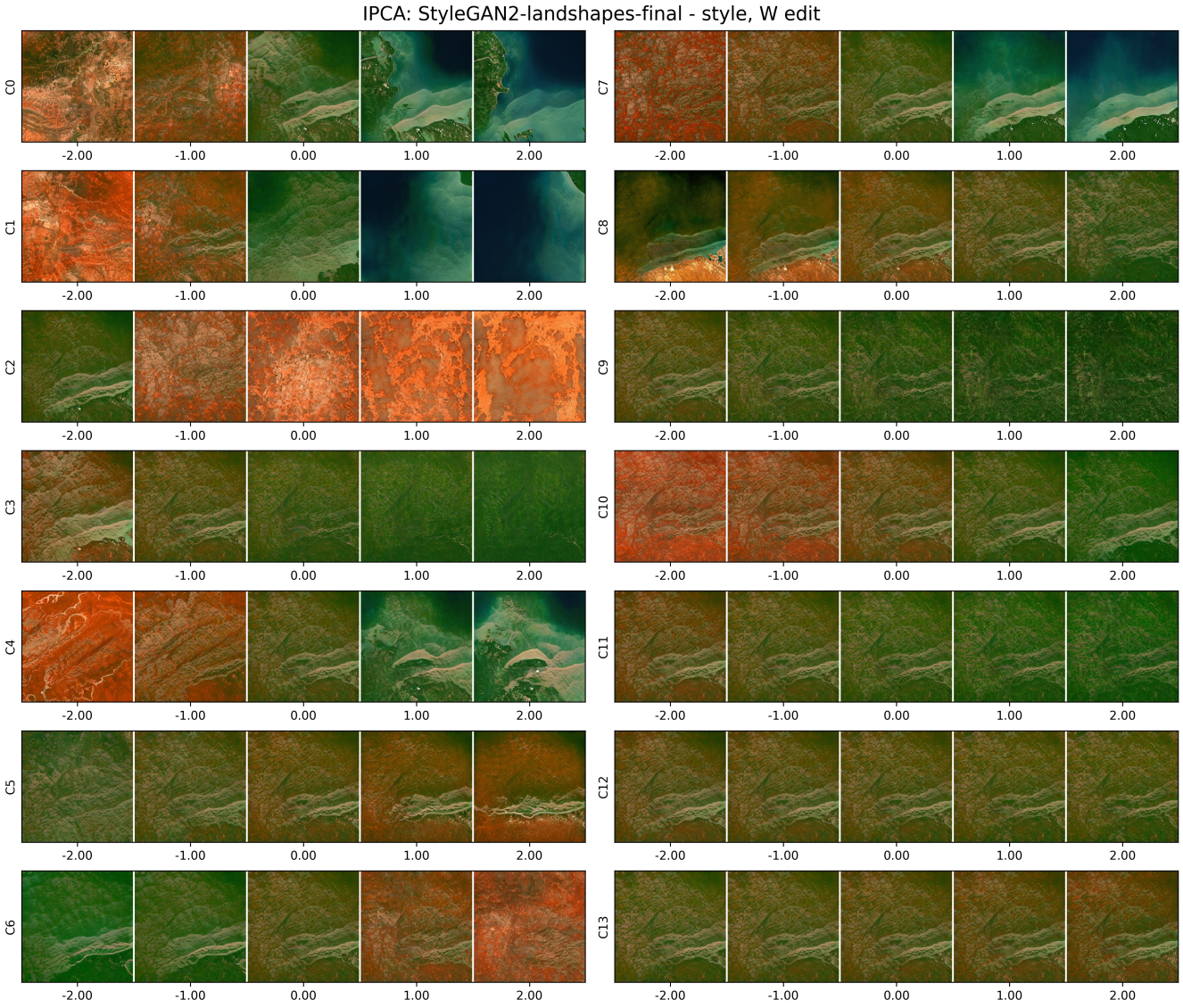


Figure 10: FOREST/DESERT

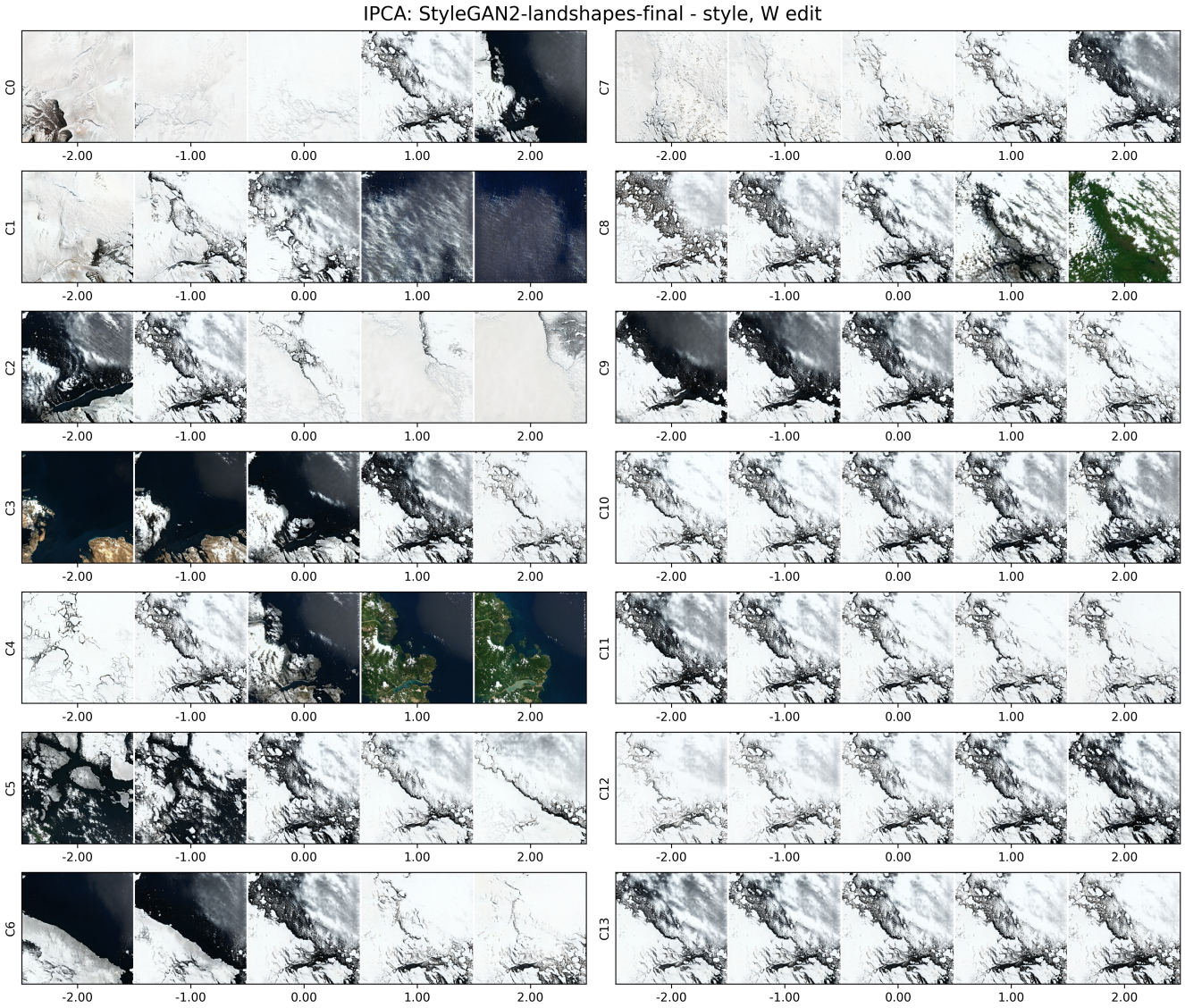


Figure 11: ARCTIC