Iterative training with human rated images to improve GAN generated image aesthetics: Effects of dataset size and training length

Betul Irmak Celebi
Supervisors: Willem van der Maden, Garrett Allen
Professors: Derek Lomas, Ujwal Gadiraju
EEMCS, Delft University of Technology, The Netherlands

23-6-2022

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

Abstract

Generative Adversarial Networks (GANs) brought rapid developments in generating synthetic images by mimicking structures in the training data. With the list of application of GANs growing drastically, it has lately become an exciting technology to explore for designers to communicate their ideas and arts through technology and create engaging experiences for humans. Nevertheless, translating human experiences to artificial intelligence and creating visually pleasant imagery is a challenging task due to complex semantics of human perception. To address this issue, we introduce an iterative training approach in which the generated images are curated by humans and the most pleasing ones are fed back into the network to retrain. Additionally, we do a factorial analysis to investigate how the aesthetic quality and the diversity are affected by the size of training data and training length. In experiments, we validate that this method can significantly improve the aesthetic quality of generated images regardless of the dataset size and training length, however the use of smaller datasets comes with a cost of reduction in the image diversity and novelty in the output images. The aesthetic bias towards certain contexts can also deteriorate the diversity and affect the model evaluations. On the other hand, no significant relationship has been found regarding the training length, however this could possibly be due to instabilities that happen during the model convergence progress.

1 Introduction

Since the emergence of Generative Adversarial Networks (GANs), introduced by Ian Goodfellow [8], the focal aim of generative image models has been synthesizing photo-realistic images. Although these models excel at learning the latent distributions in the empirical data to mimic and resemble the underlying structures, they know nothing about the human experience lying within the complex process of perception [10]. Automated metrics such as Frechet Inception Distance (FID) [22] or Inception Scores (IS) [11] have been in use to evaluate the quality of the novel images generated by these models, however, these metrics are not direct measures of human experience nor the aesthetic quality of the images.

Most work in the computational aesthetics community is related to analyzing and designing features to capture "aesthetic properties" through photographic guidelines and practices such as "the golden ratio, color harmonies or the rule of thirds" [9]. Often we can use these generalized calculations to measure the aesthetic quality of the generated images to estimate how likely they would create positive experiences. However, there are visual and semantics representations that are too complex to explicitly define as a mathematical formula. The most effective solution is to ask humans for their subjective feedback to evaluate the aesthetic quality of the images [17].

The AI and experience project Landshapes©by Frederik Ueberschaer explores how to utilize GAN models to generate intriguing landscape images as shown in 1, to evoke climate fascination in public[24]. The novel landscape images generated by the model were then shown to participants (n=5) to gather feedback on their emotional experience and how aesthetically pleasing they found the images[24]. A question that remains from this project is, how to inform the AI systems, namely GANs, about human feedback and improve the human experience.

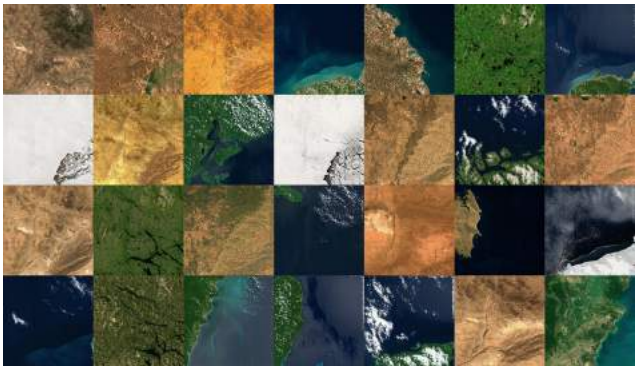


Figure 1: Sample of landscape images of various land shapes such as aquatic, grassland, desert and tundra biomes generated by the Landshapes model.

This paper aims to explore how to effectively keep humans in the loop and directly incorporate human-rated images in the iterative training process. To achieve that, we will be gathering crowdsourced ratings, evaluating their interactions with the images generated by the custom Landshapes GAN [24] and feeding the highly rated images back into the model. Furthermore, we will be creating multiple models trained with different sized subsets of the ranked images, and evaluating the human reactions throughout different training iterations to investigate the effect of the relationship of the dataset size and training iterations over the quality and variety of the output images. Precisely, the paper aims to answer the following questions:

1. To what extent can we improve the aesthetic quality of the model, through selecting the most pleasing outputs and retraining the GAN?
2. How does the relationship of the dataset size and the amount of training iterations affect the aesthetic improvements during the iterative training?

The structure of the paper is as follows: Initially, preliminary research on the related literature will be explained in section 2. The proposed method of iterative retraining will be introduced in section 3. Section 4 will be presenting the methodology including data preparation of the generated and rated images, and how the subsets of this dataset are used to retrain and investigate the Landshapes GAN model. Furthermore, the evaluation methods of the aesthetic quality via crowdsourced ratings and the resulting diversity of the output images is also explained. Section 5 will be a displaying the experimental results, while 6 will be a discussion on these

results, further elaborating on the parameters, such as bias in annotation or ratings, that might have led to our findings. Section 7 will reflect on the ethics and the reproducibility of the research. Finally concluding thoughts on the research and possible future work will be discussed in section 8.

2 Background

Within a few years after [8] introduced GANs, many extensions of the GAN architecture serving for different data synthesis purposes have emerged. NVIDIA's StyleGAN is an alternative GAN formulation, which shows state-of-the-art performance at generating high resolution photo-realistic images [5]. StyleGAN adapts a style based generator network architecture, which can distill high level features and smoothly combine them with "mixing and interpolation operations" to create new images [13]. This architecture consists of millions of trainable parameters, thus relies on vast amount of data and training iterations to prevent overfitting and generate high quality results, which can take days to weeks of training.

[12] introduces a novel method to train StyleGAN with limited amount of data and avert discriminator overfitting. The proposed method incorporates StyleGAN with an "adaptive discriminator augmentation (ADA)" pipeline, that helps stabilizing training given only a few thousand images. This mechanism applies augmentation operations to the images, such as pixel blitting, color or geometric transformations, before being evaluated by the discriminator during the training.

Landshapes model was trained using StyleGAN-ADA through transfer learning for 750 iterations using a pre-trained model provided by NVIDIA. This method allows using the weights of the pre-trained model instead of initializing at random values, and fine-tuning them to repurpose the model according to the a customized dataset [19]. Ueberschaer used a custom dataset of 4,000 satellite images of 1024×1024 resolution gathered through Google's Earth Engine[1] at random longitude and latitudes. Then the aesthetic quality of the images were assessed with the feedback of the audience.

There have been previous studies on how to incorporate the human interactions with GANs via human rating predictors as the loss function of the generator as presented in [17] and [25]. However no existing literature on how to directly include human rated pictures in the iterative training process has been found. Intuitively, when the model is further trained with images that are rated as "more pleasant to the eye", it should start mimicking these distributions and start generating more pleasing imagery. On the other hand, how this method will affect the diversity of the generated images should be further analysed.

3 Iterative training with selected generated images

Human annotation of the generated images as "pleasing" or not is a central step in the iterative training process. Even though the aforementioned augmentation pipelines help stabilize and generalize the models, they do not not increase the variability in the training dataset [25]. The manual effort to construct a labeled dataset remains as the bottleneck process, thus it is important to explore the minimal amount of data needed to lead to significant improvements.

Since the context of the empirical data fed back into the GAN is the same as the initial training data (satellite imagery) we expect to get significant results in the aesthetic quality of the images even with a few hundred of images. However, a small amount of data can lead to mode failures such as mode collapse, where the generator starts outputting the same or very similar images for different input vectors [7]. This leads to low diversity in the generated images.

On another note, as the size of the dataset grows, the complexity of the dataset also increases [4]. The changes in the data complexity, causes changes in the training length for the model to converge and stabilize. It is important to further explore the relationship between the dataset size and the training length, and understand how it affects the aesthetic quality and authenticity of the generated images. We would like to promote diversity and novelty, while also keeping the computation and annotation cost at minimum.

In this study, we will be investigating how the dataset size and training length can affect the changes in aesthetic quality, when a pretrained generative model is further trained with its generated images that resulted positive human interactions. Note that, it is assumed that the pretrained model to be improved, has a low enough FID score, thus can generate high fidelity and realistic looking images. This means, the pure aim of the experiment is to improve the artistic quality of realistic looking images generated by the pretrained model.

4 Method

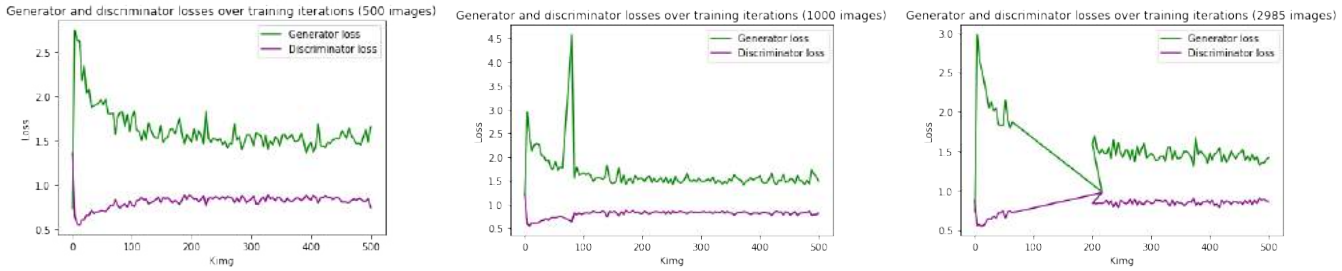
To examine what size of dataset and number of iterations would bring significant improvement in how pleasing the images are, a methodology of two main parts has been designed. The first and most crucial part is annotation of the generated images as "pleasing" or not and creating subsets of different sizes to test the effect of the size. In the second part, the Landshapes model is retrained with the resulting datasets. The resulting models are then compared on how pleasing the output images are as well as the model performance on the diversity of the generated images.

4.1 Dataset setup

In order to investigate the effect of dataset size on the aesthetic quality of the images, the size of the training sets to retrain the Landshapes model should be varied. For this task, a generic dataset of "aesthetically pleasing" pictures should be created which then can be sampled from, to create different sized subsets of this dataset.

Initially, a set of 6,000 1024×1024 dimension images were randomly sampled from the images generated by the original Landshapes model which were renamed. These images consist of biomes such as coastlines, mountain ranges, pure landmass of green fields, ice fields, and rocky fields. All images were shuffled to remove the seed number information from the images. These images were then curated in collaboration with four students from the Industrial Design Engineering faculty of TU Delft for a binary annotation ("pleasing"

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



(a) (b) (c)

Figure 2: Plots of generator and discriminator loss throughout the training progress for models trained with (a) 500 images, (b) 1000 images, and (c) 2985 images. Note that no records of statistics between iterations 60 and 220 were saved for (c) due to problems with the training environment.

or "not pleasing") process. The image dataset was split per individual for faster annotation, and then the positively voted images were merged in one dataset.

To test the reliability of the annotation group, a set of the same 200 images were distributed and annotated by each student and Krippendorff's α was computed[2]. Krippendorff's α is a coefficient in range [0,1] that measures the agreement among workers that take part in the labeling process [14], by finding the proportion of existing disagreements in the labels D_0 and expected disagreements D_e for the given data according to how many labels there are (see Equation 1). It is accepted to require $\alpha \geq 0.8$ for a robust labeling, however irresolute conclusions can be still be drawn at $\alpha \geq 0.667$, which is the lowest limit for the soundness of the annotations [15]. The group had a score of 0.864, which was deemed reliable for the labeling process of the images.

$$\alpha = 1 - \frac{D_0}{D_e} \tag{1}$$

In the end, a total of 2985 images were picked as "pleasing". This is rather a small amount for GANs, due to the lack of time and resources to annotate more images as mentioned in Section 3. However, since the original network model was trained using $\sim$4000 images, we hypothesize the amount of images needed to improve the aesthetic quality without leading to mode failure should be below the number of this number.

From the dataset of 2985 images, three levels of randomly sampled subsets were created. To observe the changes in the aesthetic quality of the images, we started with a small set of images, and gradually increased it for the experiments. The sizes of the subset levels that were adopted for the experiment are shown in Table 1.

4.2 Retraining the Landshapes model

Retraining and fine tuning the Landshapes model is a straightforward process. After creating the subsets as explained in Section 4.1, these subsets are used to train and tune the pretrained Landshapes model with transfer learning. For this step, the StyleGAN2-ADA repository that was developed and introduced in [12] has been used.

[2]Python implementation adapted from [14] https://github.com/emerging-welfare/kAlpha has been used.

Subset size	500, 1000, 2985
Iterations (kimg)	80, 200, 500

Table 1: Subset levels with amount of images and number of iterations used for evaluation and comparison. As a result the combination of the subset size and iterations, nine fine-tuned networks were used in the experiment. Although 1000 kimg is ideal for repurposing pretrained models via transfer learning [13], we hypothesize that the significant results can be acquired with small amounts of iterations since we are training a model within the same empirical domain.

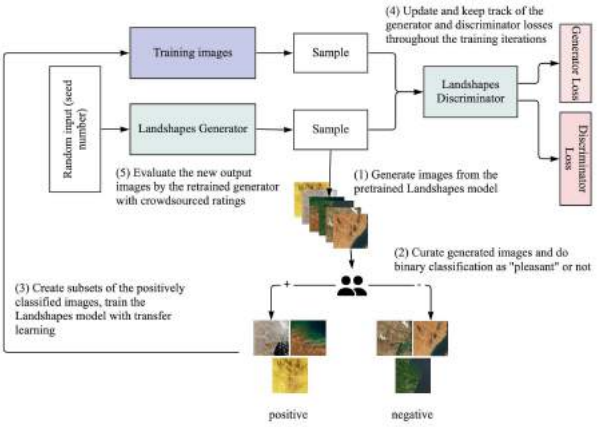


Figure 3: An overview of the retraining pipeline. We first start by generating images (1), which are then curated and annotated (2). We use the positively annotated images as the empirical data (3) to fine tune the pretrained generator and discriminator (4). We then evaluate samples of the tuned generator with crowdsourced ratings (5).

The retraining process as depicted in Fig. 3 has been applied using each image dataset, via transfer learning for 500 kimg iterations. Although the model was trained three times, we saved intermediate *.pkl* snapshots during the training process every 40 iterations to investigate the effect of training duration. Losses during the training were also trained as shown in 2. In each of the training, it is observed that there is a steep change in discriminator and generator losses until the 80 kimg. Around 200 kimg, we reach to a convergence. Thus

checkpoints at 80, 200 and 500th iterations of each training cycle were picked for evaluation as shown in Table 1.

As a result nine fine tuned models and the Landshapes model, as the baseline, were used for a comparative analysis. The images generated by each network were curated to crowdsourced workers to gather ratings and assess their perceived aesthetic quality. Then a set of qualitative and quantitative analysis was done to assess the diversity and authenticity of the generated images.

4.3 Network and Environment Properties

Models were trained partially on the Google Cloud virtual machine instances with a Tesla P1000 GPU and on the High Performance Computer nodes of the TU Delft Industrial Design Engineering faculty, using a single NVIDIA RTX3090 as the training GPU. The Landshapes model was trained for ∼60-80 hours, depending on the time it took to reach 500kimg iterations.

For the experiments PyTorch implementation of StyleGAN2-ADA[3] from by NVIDIA Research Labs was adapted. This repository provides a user friendly pipeline for customizing and training networks and monitoring the training progress with different configurations. *11gb-gpu* configuration has been utilized to maximize the GPU usage. To promote diversity within our dataset, we enabled augmentation parameters for vertical and horizontal mirroring as well as *'bgc'* parameter for pixel blitting, geometrical and color adjustments for the adaptive discriminator augmentation. As mentioned in 4.2, intermediate snapshots every 40 kimg were saved to provide qualitative and quantitative insights into the training progress. The γ coefficient of the R1 regularization, which is applied to regulate the gradient penalty during the training of the discriminator [18], was set to 10. This value was determined in regard to the proportion of the amount of pixels (1024×1024) and the mini batch size (4 per GPU) as suggested in [12]. The remaining parameters were set to the default values.

4.4 Aesthetic quality evaluation

After the training cycle, 60 images at random were produced from each of the nine fine tuned models and the baseline model, Landshapes, to assess and compare the perceived aesthetic qualities. For this, a behavioral experiment with human subjects was carried out on the crowdsourcing platform Prolific[4] in collaboration with [20]. 40 images were sampled from each set, and a survey of 100 multiple choice questions was created. In each question, the participant was shown four pictures from four distinct models picked at random, where they are asked to pick the one they find the "most pleasant to the eye". The reason the multiple choice strategy was opted for rather than a scaled rating method is due to the fast pace and simplicity of the four choice method to compare several model in one questionnaire.

To acquire a ranking, each model was scored according to the sum of how many images were picked by participants per individual model, and the amount of times an image per model was displayed. To formalize this, let P denote the set of participants and C_p the set of images a participant $p \in P$ has chosen with the respective model identifier. Given that 40 images from each model was displayed, the score of a model g was calculated as shown in equation 2.

$$\frac{\sum_{p \in P} \sum_{c \in C_p} 1_{\{c=g\}}}{|P| * 40} \tag{2}$$

The margin of error for each score was calculated through adapted Fleiss' Kappa [6], κ, calculated per question as a weight in the margin of error formula. κ is a measure of inter-rater agreement, that detects the agreement amongst two or multiple raters, in the categorical scale [23] (e.g. different models to be compared). This way the variance per a given question would affect the estimated error less, if there is more agreement between the raters. Given a set of questions Q_g in which images generated by the model g were displayed, the error has been calculated as shown in equation 3, where P_e is the expected probability of agreement, and P_o is the observed agreement for the outcome of a question. Thus the variance for each image affects the error in proportion with the disagreement level. This gives us a better idea how significant and representative the final scores are per model.

$$MAE_g = z_\gamma \cdot \sqrt{\frac{\sum_{q \in Q_g} (1 - \kappa_q) \cdot \sigma_q^2}{|Q_g|}} \tag{3}$$

where

$$\kappa = \frac{P_O - P_e}{1 - P_e} \tag{4}$$

Finally, to test the significance and inter-dependence of the ratings, Chi-squared (χ^2) test was applied. and ANOVA test were applied to the results, and the p-value has been calculated.

4.5 Diversity and authenticity evaluation

Evaluation of image quality of the generative models is a notoriously difficult task, due to the subjective nature of human perception and the lack of a metric that directly measures the domain specific authenticity and diversity of models trained with unlabeled data [1]. For this reason both qualitative and quantitative reviews were carried out to investigate the diversity and authenticity of the output images.

Quantitative assessment

Quantitative evaluation of the resulting images were determined using the precision and recall distributions (PRD) [21]. Although FID [22] and IS [11] scores are the most common metrics to assess GAN performance on how well the fake pictures resemble the training data, they are unable to characterize different shortcomings and failure cases since they yield a combined one-dimensional value [21]. For the experiment the kNN recall and kNN precision implementations of PRD, based on the works of [16] were computed.

[21] has introduced a novel definition of precision and recall. PRD disentangles the difference between the training data distribution Q and learned distribution P into two separate scores. *Precision* measures to what extent Q can be

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

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

generated by a sample of P, and symmetrically *recall* measures how much of P can be generated from a portion of Q. In other words, precision measures how similar the generated images are to the training data, thus how realistic they are. In contrast, *recall* measures how well the generated images cover the variances ins the empirical data distribution, which gives the information on how diverse the produced images are. These values can quantify the degree of GAN failures such as overfitting, mode dropping and mode invention, which give a notion of the authenticity of the generated images [3].

In addition to PRD, The generator and discriminator losses were monitored every checkpoint during training, in order to gain an understanding of the training stability and performance. From the loss plots, convergence and instances of mode collapse can be identified, which can be related to issues related to diversity.

Qualitative assessment

A manual assessment of a sampled set of generated images from each checkpoint, to monitor the categorical distributions of different biomes (e.g. coastline, grassland, forest, desert, tundra) amongst the images have been conducted. This is one of the most intuitive ways of post hoc evaluation for produced generator models [3]. Although this is a very simple and straightforward solution, it comes with some limitations since limited amount of images can be evaluated at a short amount of time. Given nine models to compare, this becomes an expensive task, as well as irreproducible.

For this reason, a pixel-wise k-means clustering method used in was applied to the 500 sampled images from each model using scikit-learn[5], to systematically assess the class imbalances and diversity within the output images. The images were separated into 5 clusters. As a metric for diversity assessment, *inertia* of the clusters have been recorded. Inertia is the sum of squared distances of the points within a cluster, to the closest cluster centroid [2]. This means, if the inertia is higher, the cluster consists of less similar images, meaning the inter-biome variation is high. Additionally, to assess the balances and distributions between different biomes, *cluster sizes* have been used as a metric. This gives an idea of which biomes and land shapes have been heavily represented in a given sample of images.

5 Results

Table 2 gives an overview of the results acquired for the evaluation metrics including scores, precision and recall values, explained in 4.4 and 4.5 for nine of the produced models and the baseline model. The generation results for each retrained model are shown in Appendix A.2. The generated images show a good variation in color and context, however there is still some overrepresentation of certain land shapes that are visible in even small samples. As seen in the generated images, there has been changes in the saturation, color variation, land shape interpolations and mixing, and landmass to water mass ratio over the course of training.

[5]https://scikit-learn.org/stable/

Images	Kimg	#Picked	Error	Precision	Recall
Baseline	0	396	±15.8	-	-
500	80	495	±15.4	0.635	0.346
	200	500	±15.3	0.624	0.229
	500	453	±15.1	0.647	0.138
1000	80	542	±16.9	0.572	0.365
	200	482	±16.3	0.556	0.250
	500	547	±16.0	0.573	0.131
2985	80	549	±16.0	0.336	0.327
	200	451	±14.6	0.334	0.201
	500	549	±14.8	0.535	0.188

Table 2: Overview of results of the evaluation metrics for the models produced and used for comparison. Picked denotes the raw ratings per each model, gathered from crowdsourced workers. Raw ratings were acquired by counting how many times in total images belonging to a model were picked. The scores in Fig. 4 were calculated as the ratio of the times a GAN was picked over amount of times it was shown. The estimated error in terms of ratings have also been included in the table. Furthermore recorded precision and recall metrics are given on the rightmost columns.

5.1 Perceived aesthetic quality

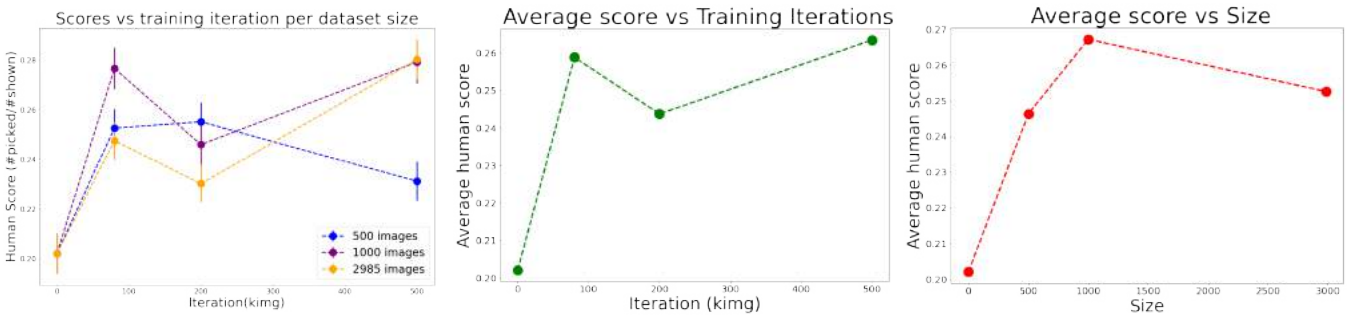
49 out of 50 survey participants were included in calculation of the results and comparison of the ratings. The χ^2 value for the survey, treating all different GANs as independent categories, has been calculated as 43.620 with 9 degrees of freedom, with the p value lower than 0.0001 showing a high statistical significance of the ratings. Fleiss' κ was calculated as 0.2, meaning the agreement between the raters was rather low, due to the subjectivity of the aesthetic perception of each individual.

Fig. 4 pictures the relationship between the aesthetic scores, number of iterations and the size of dataset. Regardless of the dataset size and iteration combination, all produced models outperformed the baseline model, verifying our initial hypothesis that the iterative learning with highly rated images would increase the perceived aesthetic quality as mentioned in Section 3. The model with the largest dataset and longest training obtained the highest score.

To investigate the individual effect of the independent variables dataset size and number of iterations on the dependent variable, perceived aesthetic quality, the average on both variables was plotted in Fig.4b and Fig.4c and a two-way ANOVA test was carried out. The test revealed that, there was a statistically significant difference in the obtained aesthetic scores between the groups. Simple main effect analysis showed that dataset size has a statistically significant effect on the aesthetic quality of the images (P=0.0289).

Taking a closer look at the factors at an individual level, Fig. 4b shows the reported perceived aesthetic quality first increases at the 80th iteration, and then has a slight decrease at 200 iterations and increases again at the 500th iteration. On the other hand, Fig. 4c displays a inverted U-shaped trend with the increasing dataset size, in which the model trained with 1000 images obtains the highest score on average.

As for the interaction plot on Fig. 4a, models trained with subset level 1000 and 2985 follow the same trend seen in Fig. 4b, with a sharp increase in iteration 80 and sharp decrease at



(a) Overall aesthetics scores calculated per model over different iterations.

(b) Average scores of models through different iterations (80,200,500 kimg).

(c) Average scores of models over different dataset sizes (500,1000,2985 images).

Figure 4: Plots showing the effect of dataset size and number of training iterations on the aesthetic scores calculated via crowdsourced ratings. Human scores were calculated with the ratio of how many times images belonging to a GAN was picked over how many images belonging to that GAN were shown in total. With 40 images represented each GAN in the survey and 49 participants in total, each GAN was displayed in the study 1960 times. All produced models have outperformed the baseline Landshapes model after training.

iteration 200 which then increases at iteration 500. The scores for the model trained with 500 images increase until thr 200th iteration and then follows a slower decrease compared. Observation of further training iterations might be needed to observe the same trend seen in 1000 and 2985 image datasets.

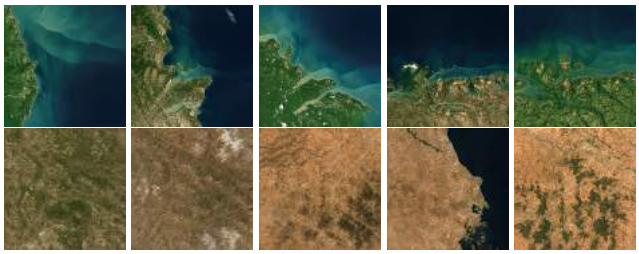


Figure 5: Top row shows the top 5 most selected images. Bottom row shows the 5 least selected (0 times picked) images. All images with the highest ratings are coastal images with light blue tones, while the least picked images are land mass images with low saturation and color variation.

When the image ratings are examined individually, there is a consistent pattern of high ratings for the coastal images with high color variation regardless of the dataset and training length. On the other hand, images with pure land mass and unsaturated brown tones were picked the least. Fig. 5 suggests there was a clear bias towards certain biomes in the way the raters aesthetic perception.

5.2 Diversity and fidelity results

Precision and recall results

As shown in Appendix A.1 and Table 2 there was a significant drop in recall over time, while precision had a slight increase for all training cycles. This resulted in high precision and low recall for all the models produced, which means a possible mode dropping might have occurred during training leading to less diversity and more contextual imbalances amongst output images.

A common trend in 2 is that as dataset grows, the precision values at the identical iterations get smaller. This is an expected behavior as [4] has shown that as dataset complexity grows, which is directly proportional to dataset size, the perceptual quality starts off lower during the training since training data replication is lower. This is due to the higher intrinsic dimensionality and larger diversity of latent structures represented within a more complex dataset.

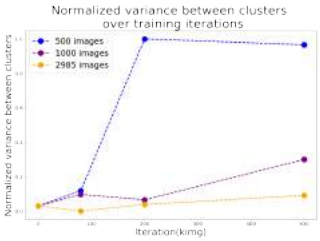
When we take a closer look at the recall to reason about the diversity, no significant differences can be observed between the models that were trained with different number of images, except for a slightly higher recall result for the model trained with the largest dataset.

Clustering results

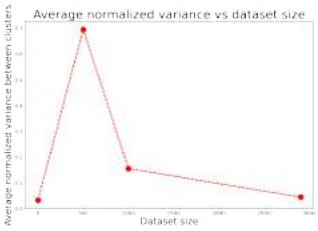
Output images were clustered into 5 different biome categories represented in Fig. 7. The most common land shapes regardless of the models were forests, coastal images and mountain ranges. The variance and inertia of the clusters that were recorded are plotted in Fig. 6.

Fig.6a shows the variance between the size of the clusters for each produced model, which tells how equally each land shape is represented in the output data. The baseline outputs the most balanced land shape representation in the generated samples since it shows the lowest variance within clusters. Initially, all models show little change in the variance compared to the baseline, however the we observe a large jump in the variance in the model trained with the smallest dataset. Looking at the averaged variances, Fig. 6b clearly shows that there is an logarithmic relationship between the dataset size and variance.

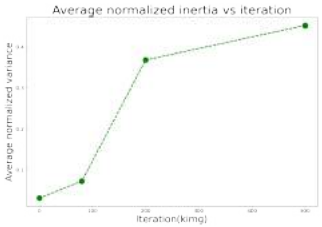
The inertia, which can be related with mode collapse, follows a similar pattern with the variance. Interestingly, the inertia first increases at 80 kimg for all models, implying the inner cluster variability is large, and there's a lot of dissimilar images of the same land shape category as depicted in Fig. 6d. As training progresses, there is a dramatic drop in the cluster inertia below the baseline, suggesting possible mode collapse or overfitting issues. Fig. 6e displays that the model trained with 500 image dataset shows the highest rate of overfitting while the other two models have a similar amount of overfitting, yet still way below the baseline.



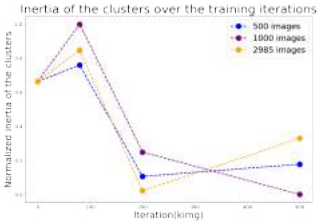
(a) Normalized variance of cluster sizes over different iterations (80,200,500 kimg) per model.



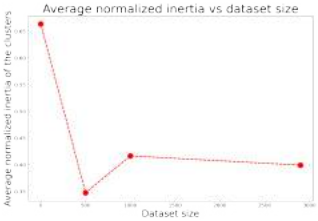
(b) Average variance of clusters over different dataset sizes (500, 1000, 2985 images).



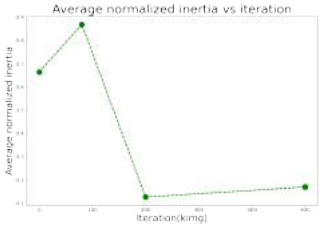
(c) Average variance of clusters over different iterations (80,200,500 kimg).



(d) Normalized inertias of clusters through different iterations (80,200,500 kimg).



(e) Average normalized inertias of clusters over different dataset sizes (500, 1000, 2985 images).



(f) Average inertia of clusters over different iterations (80,200,500 kimg).

Figure 6: An overview of the metrics that were calculated over the clustering results. Inertia plotted 6d shows how similar the images are, belonging in the same cluster. Higher the inertia means higher the dissimilarity, showing that there are various representations of the same biome in the sampled images. 6a shows the variance between the cluster sizes, so how balanced the biome representations are. An increasing variance means there is overrepresentation of one or more biomes compared to the rest. 6b plots the averaged variances of models trained with different dataset sizes and suggests that as training dataset gets smaller, the class imbalances increase significantly. All metrics were normalized between the ranges [0,1].

Figure 7: Examples of images from 5 different clusters representing different biome shapes. From left to right: forest, desert, coastal, mountain range and glaciers.

6 Discussion

The results show that, human curation method can drastically improve the aesthetic quality of the GAN generated images even with rather small datasets, but this can introduce significant aesthetic biases leading to reduction in the diversity and variety of the generated samples. StyleGAN architecture is incredibly sensitive to the distributions in the training data, thus the properties of the training data can quickly affect the aesthetic quality and the diversity of the images, and this effect gets even larger for models trained with smaller datasets due to fewer intrinsic dimensionality.

However a smaller choice of dataset size introduces issues relating to diversity, as overfitting behavior has been observed (Fig. 6e and Fig. 6b). Our findings regarding the diversity and novelty of the images back up the works of [4] on the effect of dataset size over GAN replication. The observed overfitting and replication of the training data has decreased the dataset size increased, leading to more balanced data and novel land shapes representations.

It is important to note that simply relying on human evaluations comes with fundamental limitations. Curation and evaluation of images on their aesthetic quality is a highly challenging task, due to the subjectivity in human perspective. Aesthetic bias in the training data and the evaluation methods can highly affect the model ratings. Looking at the training dataset a large proportion of the images were aquatic and forest land shapes, which lead to more water masses becoming apparent in the output images over training iterations.

Looking at Fig. 5 there was a clear bias towards coastal images with large color and salient structure variation. When we manually inspect the models that had less imbalances between the class distributions, we see that there has been a lot of style and features mixing of different land shapes, introducing variation within the categories. This could possibly be the reason that the model trained with 2500 images for 500kimg obtained the highest human ratings, while also having a fairly good diversity compared to the other models.

An unexpected finding during evaluation was the effect of training length on the human scores. Both the human aesthetic scores and diversity had fluctuations over the intermediate checkpoints, whereas the expected behavior would be a direct correlation. One assumption to explain this surprising behavior is the instability of the training. During training, the generator and discriminator compete against each other to find an equilibrium. Looking at the loss plots in Fig. 2, although there is a convergence we can still observe some fluctuations in the losses during the mini-max interplay of the

Figure 8: Images of same seed number over the 80, 200 and 500 iterations from left to right. Over the course of the training, we can see the increase in the salient structures and the color variation. However before the image converges to the final state, there are unrealistic looking structures morphing in to each other while mixing the aquatic and forest land shapes.

models.

Additionally, during the manual inspection, we have come across with generated images that were with structures morphing into each other at the intermediate checkpoints as shown in Fig. 8. During the training, intermediate iterations are still learning how to mix features to create realistic looking images and converge to the final state, thus this might have caused the middle iterations with morphing images to obtain lower ratings.

7 Responsible research

It is a crucial task to carry a scientific research with the highest standards of quality and ethics. The methodologies, evaluation steps and discoveries in this paper aim at having a great level of transparency to ensure the objectivity as well as accuracy of this study. It is necessary to address the research reproducibility and integrity, which are two hallmarks of a reliable science.

7.1 Integrity

Research integrity denotes honest and valid methods in performing and evaluating a scientific study. It is the responsibility of the researcher to clearly communicate the methodology and results without any data falsification nor a conflict of interest. Due to the great data reliance of GANs, ensuring total transparency during the data collection process is vital. To make sure of the integrity of the dataset annotation, dataset setup has been thoroughly explained step by step and a high enough Krippendorff's alpha was obtained to achieve high reliability of the annotators.

With the randomness of generative models, multiple evaluation metrics were included in this study to draw conclusions, with clear explanations on why these metrics were chosen. Qualitative and manual evaluations were done on large samples of generated images to remove randomness, and get integral and holistic understandings of the model performances.

Another aspect that was carefully considered was collecting crowdsourced results for the created surveys. All survey participants that have contributed their time and data for the research have signed an informed consent, which clearly indicated the aim of the research and the way their data is processed. No personal and identifiable data was collected, as each rating was associated with a randomized unique ID per participant. To test the reliability and integrity of the participants' responses, 5 reliability checking questions were added and participants (n=1) who failed to answer these questions correctly were not included in the evaluation process. The way the survey results have been processed were clearly communicated in the study, and it was ensured that there was no data falsification in the outcomes.

7.2 Reproducibility

Due to the black box nature of artificial intelligence and neural networks, there is a lot of underlying randomness. This randomness makes reproducibility very complex and intricate especially in the realm of generative models. An important step to make sure of reproducibility, is to communicate clearly what data has been used for the study. In this study, a custom dataset was created through annotation of design student from Industrial Design Engineering faculty of TU Delft. Although annotation of images on whether they are "pleasant" or not is a highly subjective matter, it was made sure to state the Krippendorff's alpha for the curators to test and communicate the reliability of the annotators. The data and the subsets of datasets that were created through this process were clearly documented. Furthermore, every step of the experimental setup and methodology, including the environment and network properties, and how the evaluations have been carried out were provided in this study step by step. Open source resources, repositories and their related papers such as StyleGAN2-ADA were referenced. This way, the experimental steps and the results are able to be reproduced in further experiments.

8 Conclusions and future work

The aim of this study was to explore ways of incorporating human feedback in GAN models though the iterative training method to increase the aesthetic quality of the generated images. Additionally, the effects of the dataset size and training length were investigated. We have shown that the perceived aesthetic quality of GAN models can be improved significantly using this method, as all models produced have outperformed the baseline. Although this improvement can be achieved with relatively small datasets, problems regarding the diversity and novelty in the generated images emerge. The aesthetic bias in the training image curation can be reflected to the output images, represented in the reduction of diversity. We have not seen a direct correlation between training length and the aesthetic quality of the images, but we have come to the conclusion that it might have been due to instability during the training of StyleGAN.

Aesthetic improvement and evaluation comes with a lot of challenges due to the lack of ground truth. Since there is no universal standard of beauty, the feedback provider's definition and experience can affect the quality of the generated images. This also results in an aesthetic bias in the feedback, which leads to a decrease in the variety of images generated. However it becomes an interesting task to explore if human biases define the definition of aesthetic. In future works it would be advantageous to further explore ways of promoting diversity within images. Methods of correcting the bias, by categorizing and balancing biome representations could be a step to increase the land shape variations. Furthermore curation and evaluation can be done with comparison of images

of the same land shapes so more balanced datasets can be obtained to be used in training.

References

[1] Ahmed M. Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. *CoRR*, abs/2102.08921, 2021.

[2] Matthew Bisatt. Clusters, inertia, and root numbers, 2019.

[3] Ali Borji. Pros and cons of GAN evaluation measures: New developments. *CoRR*, abs/2103.09396, 2021.

[4] Qianli Feng, Chenqi Guo, Fabian Benitez-Quiroz, and Aleix Martinez. When does gan replicate? an indication on the choice of dataset size. In *International Conference on Computer Vision (ICCV)*, 2021.

[5] Qianli Feng, Chenqi Guo, Fabian Benitez-Quiroz, and Aleix Martinez. When do gans replicate? on the choice of dataset size. 2022.

[6] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[7] Ian J. Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *CoRR*, abs/1701.00160, 2017.

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

[9] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions, 2017.

[10] 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.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. 2017.

[12] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *CoRR*, abs/2006.06676, 2020.

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

[14] Krippendorff Klaus. Computing krippendorff's alpha-reliability. *PennLibraries*, 2011.

[15] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications, 2004.

[16] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models, 2019.

[17] Andrew Kyle Lampinen, David So, Douglas Eck, and Fred Bertsch. Improving image generative models with human interactions, 2017.

[18] Lars M. Mescheder. On the convergence properties of GAN training. *CoRR*, abs/1801.04406, 2018.

[19] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze discriminator: A simple baseline for fine-tuning gans. *CoRR*, abs/2002.10964, 2020.

[20] Moshiur Rahman. How can crowdsourced workers effectively rate landscape artwork images produced by generative adversarial network transformers?

[21] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall, 2018.

[22] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

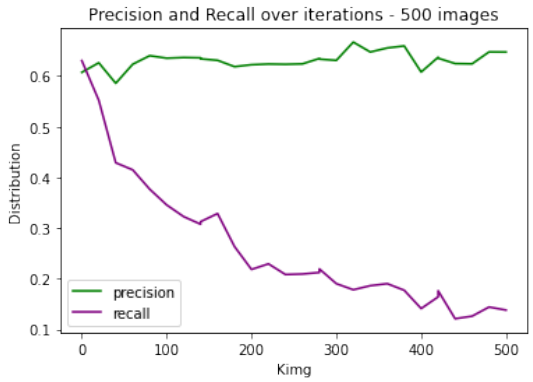
[23] Laerd Statistics. Fleiss' kappa using spss statistics, 2019.

[24] Frederik Ueberschaer. Ai for experience: Designing with generative adversarial networks to evoke climate fascination, 2021.

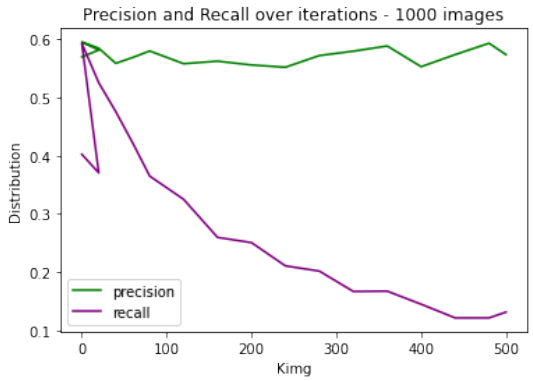
[25] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, 2015.

A Appendix

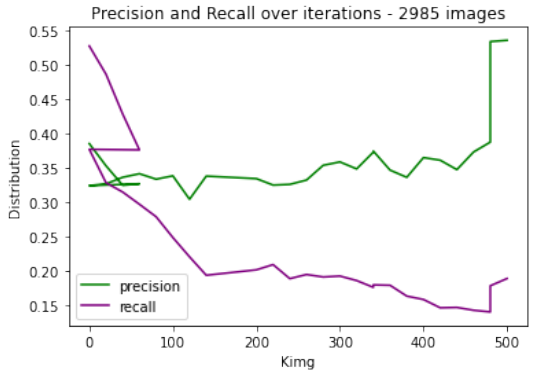
A.1 PRD results over the training



(a) 500 images



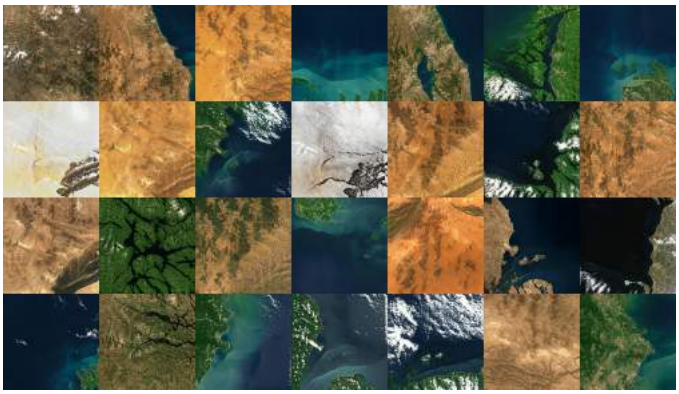
(b) 1000 images



(c) 2985 images

Figure 9: Precision and recall throughout training for different datasets.

A.2 Generated fake pictures



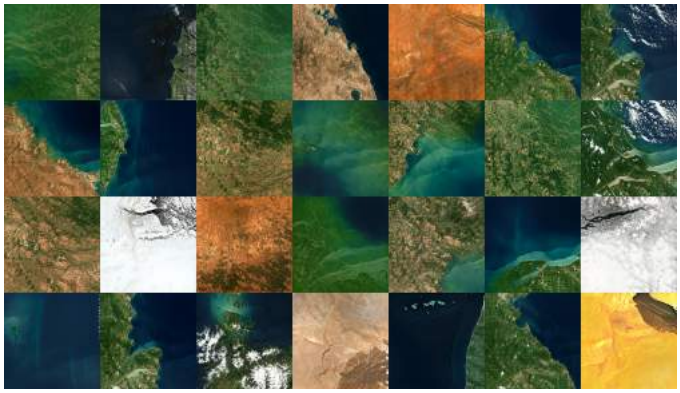
(a) 80 kimg

(b) 200 kimg

(c) 500 kimg

Figure 10: Fake images generated over training for model trained with 500 images.

(a) 80 kimg



(b) 200 kimg

(c) 500 kimg

Figure 11: Fake images generated over training for model trained with 1000 images.

(a) 80 kimg

(b) 200 kimg

(c) 500 kimg

Figure 12: Fake images generated over training for model trained with 2985 images.