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

Betul Irmak Celebi
b.i.celebi@student.tudelft.nl
Professors: Derek Lomas and Ujwal Gadiraju
Supervisors: Willem van der Maden and Garrett Allen

01. Introduction

- The focal aim of generative image models has been to create realistic images rather than aesthetically pleasing ones [1].
- The AI and experience project Landshapes [2] utilizes a style architecture GAN to create aesthetic satellite images to create climate fascination.
- Formalizing aesthetic quality is too complex. To measure and improve the aesthetic quality of such models, it is necessary to keep human feedback in the loop.

02. Research Question

- To what extent can we improve the aesthetic quality of the model, by selecting the most pleasing outputs and retraining the GAN?
- How does the relationship of the dataset size and the number of training iterations affect the aesthetic improvements and diversity during the iterative training?

03. Methodology

Dataset setup

- Generating 6000 images from Landshapes GAN
- Curation of images as "positive" or not in collaboration with 4 students from TU Delft Industrial Design Engineering Faculty

Krippendoroff's α = 0.87

- Creating subsets of the positively annotated images
  - 500 images
  - 1000 images
  - 2985 images

Iterative training

- Transfer learning with Landshapes GAN and saving intermediate snapshots at:
  - 80 kimg
  - 200 kimg
  - 500 kimg

Evaluation

Aesthetic quality evaluation

- 4 choice behavior experiment with images shown from different GANs
  - 40 images from each GAN
  - 49 participants included

Diversity evaluation

- Precision and recall
- Pixel-wise k-means clustering of generated images
  - Inertia within clusters
  - Variance between clusters

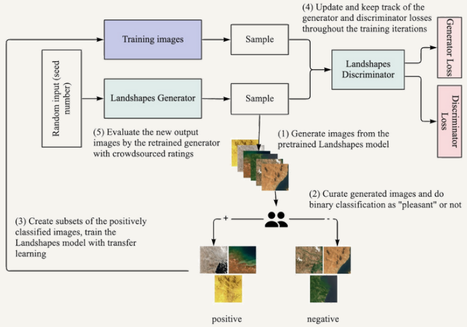


Figure 1: Image curation and training pipeline of Landshapes GAN

04. Results

Crowdsourced aesthetic results

- All produced models outperformed the baseline model
- Average human scores fluctuated over training iterations
- Clear bias towards coastal biomes in human ratings

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 1: Number of times images belonging to respective GANs with different subset level and iterations with the margin of error.

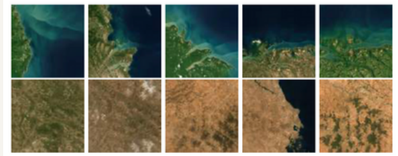
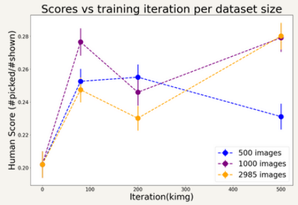


Figure 2: First row with the images that obtained the highest score and second row with the images obtaining the lowest score



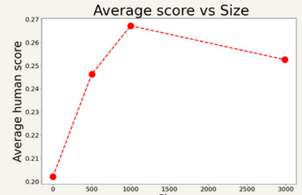
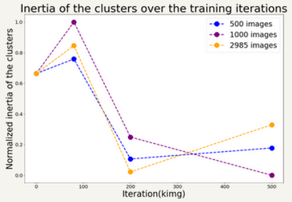


Figure 3: Human scores obtained for the models trained with different sizes of datasets and training lengths. All were calculated through the ratio of #picked/#displayed in the surveys.

Diversity and novelty results



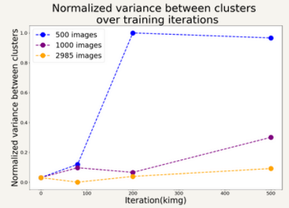


Figure 4: Plots of inter-cluster variance and inner cluster inertia of all models trained with different subsets, over training iterations

- Decrease in recall over training progress for all models
- Overrepresentation of forest and coastal images in the outputs was present
- There's a similar pattern between averaged inertia and aesthetic results over iterations with fluctuations over intermediate iterations
- More balanced representation of biomes as dataset size increases, while there is a significant imbalance with the smallest dataset

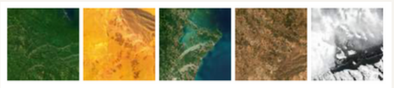
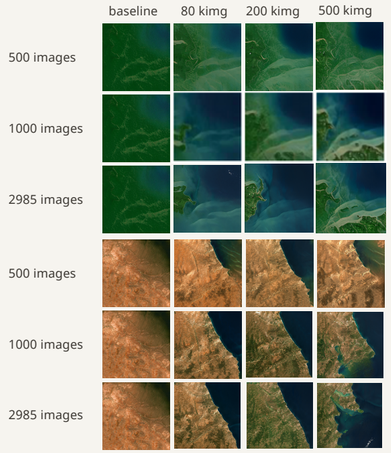


Figure 4: 5 clusters representing different biomes were created

Same seed images from each model show the changes on the baseline model over the iterations. There is an increase in color variation and more balanced water and land mass apparance over the iterations.



05. Conclusion

- the perceived aesthetic quality of GAN models can be improved significantly using this method
- Although this improvement can be achieved with relatively small datasets, problems regarding the diversity and novelty in the generated images emerge.
- 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. Aesthetic bias affects the diversity and evaluation of the models.
- Methods of correcting the bias, by categorizing and balancing biome representations could be a step to increase the land shape variations

Related literature

[1] N. Murray, "PFAGAN: An Aesthetics-Conditional GAN for Generating Photographic Fine Art," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 3333-3341, doi: 10.1109/ICCVW.2019.00415.
[2] F. Ueberschaer, "Ai for experience: Designing with generative adversarial networks to evoke climate fascination," 2021. [Online]. Available: http://resolver.tudelft.nl/uuid:731b92cc-ec9e-4543-a608-c0edbdb14aaf