How can crowdsourced workers effectively rate artwork images produced by Generative Adversarial Network transformers?

Moshiur Rahman
Responsible Professor(s): Derek Lomas, Ujwal Gadiraju
Supervisor(s): Willem van der Maden, Garrett Allen
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

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

Abstract

Generative Adversarial Networks (GANs) can create artwork images and we need effective ways of rating their aesthetic values. This could help us determine the most aesthetic artwork images (and identify the GANs that created them) and train GANs to produce more aesthetic artwork images in the future. In this research, we analyzed the effectiveness of using two different survey formats (binary-choice and four-choice) for displaying GAN-produced artwork images to crowdsourced workers and gathering their ratings. The artwork images were of different landscapes like the desert, arctic, coastal regions, etc. Additionally, we investigated how the choice of showing different images together (image groupings) per question affects the final rating results. Results demonstrate that the four-choice format is superior to the binary-choice format in producing more consistent, reliable, and accurate results. The effects of the different image groupings were insignificant for the results of the four-choice format. In contrast, different image groupings displayed statistically significant changes in the results for the binary-choice format. However, it was found that crowdsourced workers preferred the binary-choice format more as they found it to be less strenuous and more effective in allowing them to express their rating choices.

1 Introduction

In today's world, Artificial Intelligent (AI) agents can create artworks. This is done using an AI framework called Generative Adversarial Networks (GANs) [1]. The generators of GANs learn features of training images and create additional fake instances of it [1, 10]. Unlike conventional GANs (where neural network discriminators are used to determining the quality of the data instances created by the generators [1, 10]), the created images are then rated by human discriminators to determine their aesthetic values. This method is called HumanGAN [7]. In essence, humans are shown different images created by different GANs and they have to give some form of rating feedback on the images. This in-turn is processed and used as training data for the generators to further produce more aesthetically pleasing images. The benefit of this method is that the human ratings capture constructive feedback about the types of artworks humans find aesthetically pleasing. During the backpropagation phase to the generators, the rating feedback is then transferred effectively which allows the generators to further methodically learn from the training images and improve in creating more aesthetically pleasing images [7]. Figure 1 shows two images produced by two different GANs (GAN250 and TCDNE) [1]. The images look different and have different aesthetic values. Figure 1a is more colorful, has more complex shapes, and could be considered to be more aesthetically beautiful than the image in figure 1b. If human discriminators rate these images, this type of feedback would go back to the GAN generators. This would result in the generators creating artwork images more similar to figure 1a than figure 1b.

[1] https://drive.google.com/drive/folders/1F-4lBU69J6BemX-t2jDp6xorWHBgNgRy?usp=sharing

(a) Image produced by a GAN named GAN250 images

(b) Image produced by a GAN named TCDNE images

Figure 1: These artwork images were produced by two GANs. An example way of acquiring rating feedback would be to ask humans to choose the image they find more aesthetically pleasing from the two images

Collecting rating data about the beauty, aesthetics, and the overall "preferability" of artwork images can be done in several ways like crowdsourcing, traditional surveys, interviews, etc. However, in this research, we will only consider crowdsourcing as an option. The reasons why crowdsourcing is more effective (for this research) than the other aforementioned methods is because firstly, it will expose us to a more diverse set of audiences, secondly, the data will be easier to collect and process, and lastly, it is faster and more economical [9].

Extensive research has been done on designing surveys with a binary-choice format (i.e. choosing the more aesthetic image from the two presented options) and the scale rating format (e.g. rating an image from 1 to 10, where 1 is the very unaesthetic and 10 is very aesthetic) [2, 5]. In general, it has been concluded that in a crowdsourcing setting, the binary-choice format is more effective than the scale rating format because it is less difficult and cumbersome for the participants [2].

In this research, we will compare a binary-choice format with a four-choice format (i.e. choosing the more aesthetic image from four presented options) along. Even though the binary choice format is used and commended in much research [2, 5], we believe that the four-choice method is a worthwhile study. Some studies suggest that having four choices in a text-based multiple-choice is optimal [16, 6] and therefore we want to investigate whether this is also the case or not in an image-based multiple-choice survey. Additionally, there has been little to no research conducted investigating the four-choice format and see its effectiveness. We want to understand how it will compare against an established method in crowdsourcing.

The artwork images[2] that we will use are of landscape sceneries from a bird's-eye point of view. Two different survey formats will be used: the binary-choice format and the four-choice format. We will create a 128-question survey for the binary-choice format and a 64-question survey for the four-choice format. In addition to comparing the two different formats, we will also investigate the effects of image grouping (the groups of images shown together on each page of the survey) on each format. Image grouping is an essential factor because the participants will be shown two or four images and will have to choose one that they find the most aesthetic. We want to investigate whether this factor ultimately affects the final results of which images are considered to be the best (and therefore which GAN produces the best images). Altogether, 256 images from four different GANs will be used for each survey. The images will remain the same and will be a control variable. We will host the survey on Prolific which is a crowdsourcing platform. After acquiring the results, will analyze the

[2] https://drive.google.com/drive/folders/1F-4lBU69J6BemX-t2jDp6xorWHBgNgRy?usp=sharing

data using different statistical tools like the Chi-squared test, Jensen Shannon Divergence, and the Pearson correlation coefficient. Firstly, we will investigate the reliability of the results. Secondly, we will conduct goodness of fit analysis. Thirdly, we will conduct a test of independence analysis. Fourthly, we will examine the effects of image grouping on the results for the two formats. Finally, we will consider the attitudes of participants towards the two survey formats.

The primary research question of this investigation is: how can crowdsourced workers effectively rate artwork images produced by Generative Adversarial Network transformers? There are two sub-questions. Firstly, which of the two, the binary-choice or the four-choice, survey formats is better for gathering crowdsourced ratings of GAN-produced artwork images? Secondly, to what extent does image grouping affect the final results for the binary-choice and the four-choice survey formats?

We hypothesize that the four-choice format will be better than the binary-choice format. This is because the image groupings of the four-choice are larger which will allow participants to choose from a larger set of images giving them options and flexibility. This can be quite challenging in the binary-choice format as there are only two images and if the participant finds the aesthetic values of the images to be the same, it will be harder to choose. Secondly, the effects of the image grouping in the surveys could be more problematic for the binary-choice than the four-choice. This is because if we are rating images from more than two GANs, a fewer number of images from each GAN will be shown simultaneously. Since participants are forced to choose an image for every question, there is a higher chance that an image from a worse GAN to be selected compared to an image from a better GAN in the binary-choice format than in the four-choice format. In addition, because the group size is smaller in the binary-choice format compared to the four-choice format, the overall final ratings for the images and the GANs could vary more between one binary-choice survey to another binary-choice survey compared to the different versions of the four-choice surveys.

2 Methodology

The methodology section contains details about designing and creating the surveys. This includes details of the software pipeline that was built to speed up the survey building process, and extra survey features that were added to the surveys to check for consistency and the reliability of the results.

2.1 Tools For The Experiment

There were four different tools used for this experiment. Firstly, we used Qualtrics [3] platform to design, format, and create the surveys. Secondly, after creating the survey, we hosted it in Prolific [4], which is a platform where people can host surveys and questionnaires for crowdsourcing research. We only exposed it to crowdsourced workers in the United Kingdom and the United States of America to minimize potential language barriers when completing our survey. Thirdly, we used the Python Programming language [5] while creating the survey and while processing and analyzing the data. The Pandas library [6] was used for the second part. Lastly, the images were generated from four different GANs, "This city does not exist" (TCDNE), "Satellite", "Baseline", and "GAN250" [7] [8]. TCDNE was

[3] https://www.qualtrics.com/

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

[5] https://www.python.org/

[6] https://pandas.pydata.org/

[7] Image data set: https://drive.google.com/drive/folders/1F-4lBU69J6BemX-t2jDp6xorWHBgNgRy?usp=sharing

[8] Example images from the four GANs can be seen in Figure 10 - 13 in the Appendix section

created by the StyleGAN-2 [9]. The Baseline and GAN250 were created by StyleGAN2-ADA [10]. The Satellite images were used as training data to generate the Baseline image and the Baseline images were used to generate GAN250 images. The StyleGAN2-ADA and the Baseline images were acquired from Frederik Ueberschär's AI FOR EXPERIENCE: Designing with Generative Adversarial Networks to evoke climate fascination projects [15].

2.2 Creating The Survey

Each survey contained a total of 256 landscape images from four different GANs and altogether four separate surveys were created. The surveys were named Binary-Choice-1 (BC1), Binary-Choice-2 (BC2), Four-Choice-1 (FC1) and Four-Choice-2 (FC2). The BC1 and BC2 surveys had 128 questions long and had a binary-choice design format per page. Each of the choices was an image from one of the four GANs. While completing the survey, the participants had to choose the image they found to be "the most pleasing the eye". The FC1 and FC2 have a four-choice design format per page and there were 64 questions altogether. The participants had to choose of the four images that they found "the most pleasing to the eye". The reason why we chose to have a maximum survey length of 128 questions is because research [13] shows that the ideal survey should be below 15 minutes and according to Qualtrics, the estimated time for 128 questions was below the limit. Note that the surveys with 64 questions were also estimated to be below 15 minutes according to Qualtrics.

2.2.1 Automating Selecting The Images For the Survey

Since manually organizing and selecting the images from the different GANs and creating the survey was a laborious process, we decided to automate it. In the beginning, we tried to manually create a survey of a similar format for testing purposes and it took roughly 1 hour and 30 minutes, and after the automation, the whole process took around 10 minutes. The repository containing the code can be found on GitHub [11].

There were five steps to the automation.

1. Put the images of the different GANs into different folders and rename all of them into the correct format by running the script in correctlynameimage.py file.
2. Create four different graphic folders in Qualtrics and upload the images there.
3. From the "My Library" section in Qualtrics, copy the GAN image IDs into four different text files.
4. Run the survey creation algorithm by from the formatqsffile.py file. This will create the edit the Template.qsf file and put the GAN images in the survey. The algorithm should output a new file named survey.
5. Upload the survey.qsf file to Qualtrics.

2.2.2 Extra Survey Features

Along with the research survey questions, we have added four other key features to make the survey more reliable and participant friendly. The other sections are research description and consent form,

[9] https://thiscitydoesnotexist.com/
[10] https://github.com/NVlabs/stylegan2-ada
[11] https://github.com/moshiur112/RPprojectcode

example questions, attention questions, and questions about the attitude towards the two survey formats.

Firstly, for the description and consent form section, we explained the background and the purpose of the research, necessary contact information, and a consent form to which the participant had to agree to agree that they are voluntarily participating.

Secondly, in the example questions section, the participants had to fill in three example questions that were similar to the research survey questions. This was to help them to become acquainted with the research questions and how they should answer them. The responses were not recorded for this question.

Thirdly, throughout the research survey question, attention-checking images were added in random locations. There were these questions added to the four-choice surveys and six added to the binary-choice surveys. The reason why the binary-choice surveys had twice as many were because the surveys were twice as long. In the attention checking questions, all the images of the question were the same and the participants were asked to select a specific one of them (e.g. "Select the image on the top-right" for a four-choice survey). This was done to check whether the participants were paying attention to the images while answering the questions or were hey clicking randomly [12]. If the participant got any of the questions wrong, we would discard their whole response data. This section guaranteed that the participants were attentive and serious about the answers.

Lastly, for BC1, BC2, and FC2 the surveys, after the research question section, we added 10 additional questions in a format that is different from the research question section. So if the survey had a binary-choice format for the research questions, this section contained 10 additional questions with the four-choice format. In this section, the participants also had to choose the image that they found to be the most "pleasing to the eye". The purpose of this section was to get them acquainted and experienced with answering questions in a different format than the research questions. Afterward, three questions were asked to assess the attitude of the participants towards the two survey methods. The first question was about the format preference, the second question was asked about which format allowed them to express their choice more effectively, and accurately and lastly, which choice was less laborious in their opinion. Overall, this section's purpose was to understand the attitude of the participants towards the two survey formats. The whole pipeline of creating the survey can be seen in Figure 2.

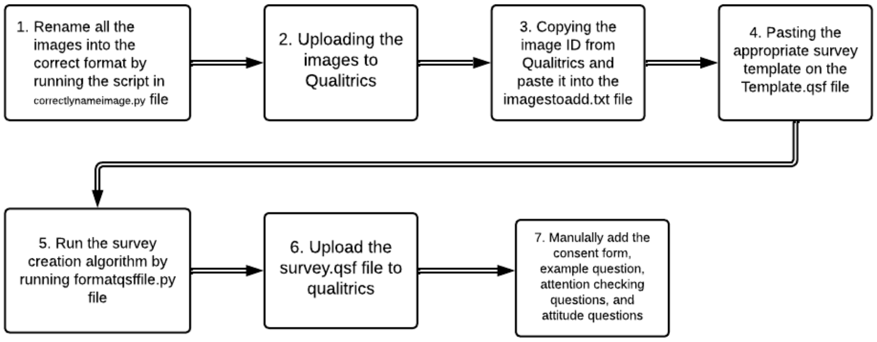

1. Rename all the images into the correct format by running the script in correctlynameimage.py file
2. Uploading the images to Qualitrics
3. Copying the image ID from Qualitrics and paste it into the imagestoadd.txt file
4. Pasting the appropriate survey template on the Template.qsf file
5. Run the survey creation algorithm by running formatqsffile.py file
6. Upload the survey.qsf file to qualitrics
7. Manulally add the consent form, example question, attention checking questions, and attitude questions

Figure 2: Pipeline of creating the survey

3 Results and Discussion

The put surveys were put on Prolific and we got 16 responses on the BC1 survey, 20 responses on the BC2 survey, 48 responses on the FC1 survey, and 18 responses on the FC2 survey [12].

We completed a series of steps to process the data, rank the different GANs and evaluate which survey format is better. Firstly, we removed all the poor responses. This includes responses where the participants did not give their consent. In addition, we checked all the attention-checking questions and removed the responses of the participants if they got more than one wrong. Afterward, performed chi-squared analysis between the responses with zero wrongly answered the attention-checking questions and responses with exactly one wrongly answered attention-checking question. If there was no statistically significant difference, we conducted another chi-squared analysis between the responses with zero wrongly answered attention-checking questions and responses with zero and one wrongly answered attention-checking questions, and if there was no statistically significantly different, then we used the responses with zero and one wrongly answered attention-checking questions for further analysis. Secondly, we conducted goodness of fit test for all four surveys to check whether to evaluate whether there is a statistically significant difference between the images chosen for each GAN [3]. This was done by calculating the total images chosen for each GAN and comparing it with the expected value (total images divided by four) for each GAN. Thirdly, we evaluated the effects of the different groupings on the results of the two survey formats. For this section, we did three types of analysis. Initially, a test of independence [4] was conducted using a Chi-squared analysis between the two versions of the surveys for the two survey formats (i.e. BC1 vs BC2 and FC1 vs FC2). This was done to check whether image groupings in a survey have a statistical effect on the final results. This was followed by conducting a Jensen-Shannon divergence distance measure and Pearson correlation coefficient test to verify the results of the Chi-squared analysis and further understand the results. Lastly, the attitude questions of the surveys were examined to further understand the attitude of the participants towards the two survey formats and to assess which format is better for gathering ratings.

3.1 Removing Poor Responses

After acquiring the data, the first step was to remove poor responses. We used the attention-checking questions to determine whether the questions were thoughtfully answered or not. Firstly, we removed all the responses from all the surveys where more than one attention-checking question was answered wrong. This left us with all the questions in the BC1, BC2, and FC2 surveys, 37 responses in FC1 survey.

Response Count Details		
	FC1	BC2
Zero or one mistake(s)	37	20
Zero mistakes	29	18
One mistake	8	2

Table 1: Response count details about how many made zero or one mistake(s) in the attention questions

Secondly, we separated all the responses where one of the attention-checking questions were answered wrongly. For the BC1 and FC2 surveys, all the questions were answered correctly so this step did not apply. Two questions were separated from the BC2 survey, leaving it with 18 responses where

[12] A visual example of the differently ranked images from different GANs can be seen in Figure 14 of the Appendix section

all questions were answered correctly. Eight questions were separated from the FC1 survey, leaving it with 29 responses with no attention-checking questions that were wrongly answered. The details are summarized in Table 1.

GAN name	BC2 - zero or one mistake(s)	BC2 - zero mistakes	BC2 - one mistake	FC1 - zero or one mistake(s)	FC1 - zero mistakes	FC1 - one mistake
TCDNE	239	15	224	155	24	131
Satellite	631	64	567	598	135	463
Baseline	796	81	715	723	148	575
GAN250	894	96	798	892	205	687

Table 2: Image count distribution from the BC2 and FC1 survey responses with zero, one, and zero or one mistake(s) in the attention-checking questions

Thirdly, we looked at the distribution of images chosen from the different GANs. This can be seen in Table 2. Afterward, we first did a test of independence between the responses with zero mistakes and one mistake using the Chi-squared test [4]. The significance level was set to 5% (0.05). In this case, our null hypothesis is that there is no difference between the two distributions and they are similar. If there is a statistical difference (where the p-value will be less than the significance level), that means that the test of independence fails and the responses are unreliable and we have to discard them [4]. If there is no statistical difference, then we do another chi-squared between the zero or one mistake and the zero mistake distribution. If there is no statistical difference, then we can infer that the distribution of the responses between all the responses with zero or one mistake(s) and all the responses with zero mistake is the same and the additional data points are reliable and we can use it for the final analysis.

Responses with number of mistakes in the attention-checking questions	BC2	FC1
One mistake	0.955	0.151
Zero or one mistake(s)	0.698	0.890

Table 3: The p-values after completing the Chi-squared test of independence between the responses with zero vs one attention-checking question and zero vs zero or one attention-checking question mistakes. This was done for the BC2 and FC1 surveys

Table 3 shows that the p-value (0.151) is greater than 0.05 when doing the chi-squared test between responses with zero mistakes and responses with one mistake for survey FC1. This indicates that the null hypothesis is not rejected and we cannot determine whether there is a difference between the distributions. Since we cannot determine a difference, we include the responses with one mistake in the attention question along with the responses with no mistake and do another chi-squared test against the responses with no zero mistakes. Since the p-value (0.890) is greater than 0.05 and there is no statistical difference according to the chi-squared test, we will use the data set with zero or one mistake(s) for the final analysis. Similar tests are applied to the BC2 survey responses. Since the p-value for both the responses with only one mistake and zero or one mistake(s) is more than 0.05, we will use the responses with zero or one mistake(s) in the final analysis.

After this section, we used the attention checking questions to determine whether the survey responses were thoughtfully answered or not, and all the poor responses were removed from the four different surveys. There are 16 responses for BC1, 20 for BC2, 37 for FC1, and 18 for FC2.

3.2 Goodness of fit test

With the selected survey responses, for each survey, we can check whether there is a statistical difference between the different rated GANs for each survey. Essentially, for each survey, this test will indicate whether the different ratings it got for each GAN are statistically different than they would be if all the images produced by the GANs are of the same aesthetic value. This test is important because if there is not a statistical difference, then we could not effectively compare the different ratings between the different GANs.

We can do this test by performing a Chi-squared test analysis between the cumulative chosen image counts of the different GANs and their expected value if all the images of the different GANs were chosen evenly. This type of test is known as the goodness of fit test [3]. Our significance level will be set to 5%. Table 4 contains the details of the surveys and their expected value for all four GANs.

GAN name	FC1	FC1 Expected	FC2	FC2 Expected	BC1	BC1 Expected	BC2	BC2 Expected
TCDNE	155	592	63	288	310	512	239	640
Satellite	598	592	278	288	472	512	631	640
Baseline	723	592	358	288	649	512	796	640
GAN250	892	592	453	288	617	512	894	640

Table 4: Image count for different GANs and their expected values for the FC1, FC2, BC1, and BC2 surveys

After completing the Chi-squared analysis, we found that all the p-values for all the different surveys were below 0.001 (Table 5) and since this is below the significance level, we can infer that there is statistical difference between the different rated GANs and their expected values for all the surveys. We can also infer that all the quality of the images produced by the four different GANs and not the same.

Survey Name	p-value	Are they independent
FC1	< 0.001	True
FC2	< 0.001	True
BC1	< 0.001	True
BC2	< 0.001	True

Table 5: The calculated p-values for the goodness of fit Chi-squared test

From this test, we conclude that for each survey, there is a statistical difference between the images produced by the different GANs. This test was crucial because if there was not a statistical difference, then we could not effectively compare the different ratings between the different GANs.

3.3 The effects of the different groupings on the results of the surveys

Since the images in the surveys are grouped randomly, it is important to check whether the image grouping is a factor in determining which GAN ratings. In this context, image grouping means which images are shown together in a survey question. Since we are using four different GANs, for the binary-choice surveys, there could be six different types of image pairings (e.g. TCDNE and Satellite, TCDNE and Baseline, TCDNE and GAN250, ect) that could occur if we do not allow multiple images of the same GAN per question. However, for the four-choice survey, four images from four of the GANs are shown for all questions.

We will do tests of independence [4] using the Chi-squared analysis to measure for statistical difference for the GAN ratings between the two versions of the two survey formats. In addition, we will also do this test between the average of the binary-choice surveys and the four-choice surveys to check if there is a statistical difference between the results of the two formats.

Afterward, we will use measure the similarity between the two versions of the two survey formats using the Jensen-Shannon Divergence distance. In addition, we will also calculate the distance between the average of the binary-choice surveys and the four-choice surveys to check how similar are the results between the results of the two formats.

Finally, we will calculate the Pearson Correlation Coefficient to measure the association between the image scores of the two surveys for the two survey formats.

3.3.1 Test of independence using the Chi-squared analysis

We conducted tests of independence using the Chi-squared analysis to measure for statistical difference between the two versions of the two survey formats and the average of the binary-choice surveys and the four-choice surveys [4]. This was done to check if there is a statistical difference between the GAN ratings for the different versions. This test is important because if there is a statistical difference then we can conclude that image groupings between the surveys play a significant role in determining the results. Otherwise, when there is no statistical difference, then we know that image grouping is not a significant factor. The significance level was set to 5% for all three tests. Table 6 contains the resulting p-values.

	BC1 vs BC2	FC1 vs FC2	Average of BC surveys vs average of FC surveys
p-value	<0.001	0.480	<0.001

Table 6: The p-values of the different surveys for the different formats

For the binary-choice format, the p-value was less than 0.001, which signifies that the relationships between the GANs and the number of images chosen for those GANs are statistically different. This means that image grouping can play a crucial factor in determining how the images and the GANs are rated.

For the four-choice format, the p-value was 0.480, which is larger than the p-value. This indicates that the is not a statistical difference between the GANs and the number of images chosen between the two surveys. This means that image grouping does not play a vital role in determining how the images and the GANs are rated.

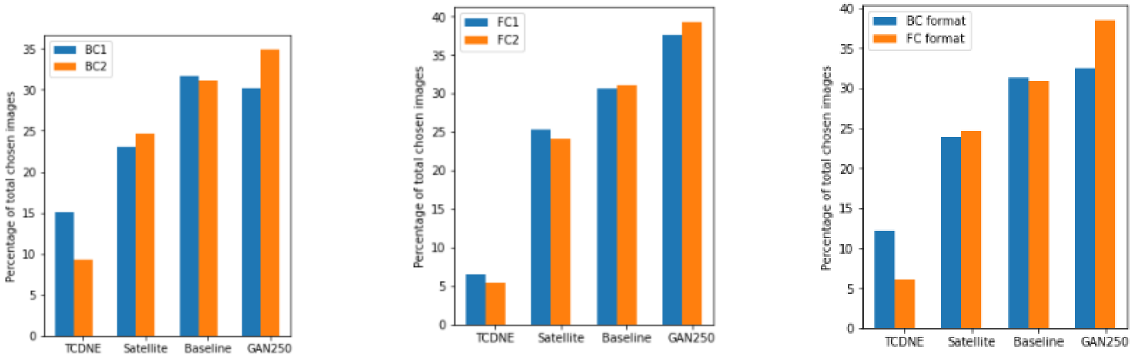
Lastly, for the test between the average of binary-choice surveys and the average of four-choice surveys, the p-value was less than 0.001. This indicates that the survey format has a statistically significant effect on the results.

By performing the Chi-squared test of independence between the different versions of the different surveys, we can infer three key points. Firstly, for the binary-choice format, image grouping plays a significant role in determining the rating results between the surveys. This is understandable because only two images from four different GANs are shown for each question. Secondly, for the four-choice format, image grouping is not a factor in determining the ratings between surveys. This is because, for each question, a single image for each is shown. Lastly, when comparing the average ratings of the binary-choice and the four-choice surveys, the is a sadistically significant difference between the overall GAN ratings. This concludes that since survey formats affect how images are grouped, it can ultimately affect the final GAN rating results.

3.3.2 Measure of similarity using the Jensen-Shannon Divergence distance

The similarity between two distributions can be measured using the Jensen-Shannon (JS) divergence distance [11, 8]. The JS divergence distance analyzes the underlying structure and how it changes when going from one distribution to another. The change is measured symmetrically which means that given two distributions A and B, the JS divergence distance measures the change in the structure of data when going from A to B and then from B to A. This is further normalized to get the final distance measure which is used to observe the similarity between A and B [11, 8]. The JS divergence distance ranges from 0 to 1 where 0 represents absolute similarity and 1 represents absolute difference [11, 8].

We measure the JS divergence distance between the two versions of the binary-choice, four-choice surveys and the average results of the two versions of the binary-choice and the four-choice surveys. This distance measure would be a good indicator of the similarities between the responses of the two versions of surveys for the two formats and the similarities between the overall results of the two formats.



(a) Percentage of images chosen for the different GANs for the binary-choice format responses

(b) Percentage of images chosen for the different GANs for the four-choice format responses

(c) Percentage of images chosen for the different GANs for the average of the binary-choice format and the average of the four-choice format responses

Figure 3: Percentage of images chosen for the different GANs for the two versions of binary-choice and four-choice survey formats and the average of the binary-choice format and the four-choice format

The input of JS divergence distance is two lists of data points ranging from 0 to 1 and we, therefore, had to normalize the data by dividing the number of image counts for each GAN by the total number of responses. We did this step for both of the surveys. Figure 3 contains a percentage representation of this data.

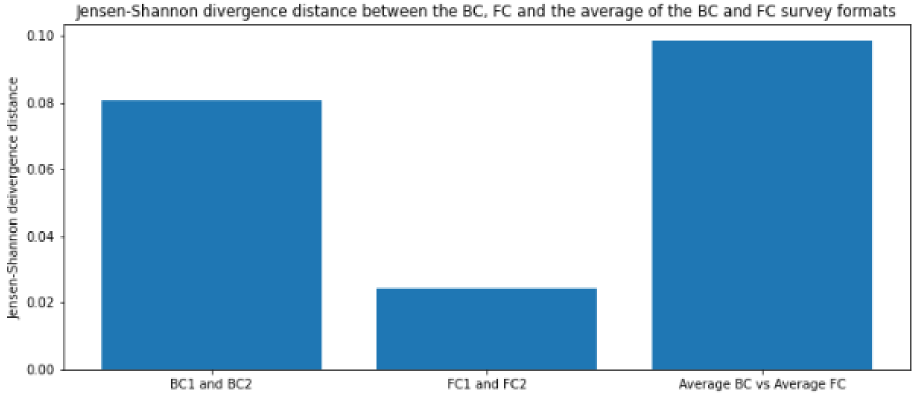


Figure 4: Jensen-Shannon divergence distance between the BC, FC and the average of the BC and FC survey formats. The distance between the binary-choice formats is: 0.081, the four-choice format is 0.024 and the average format is: 0.099

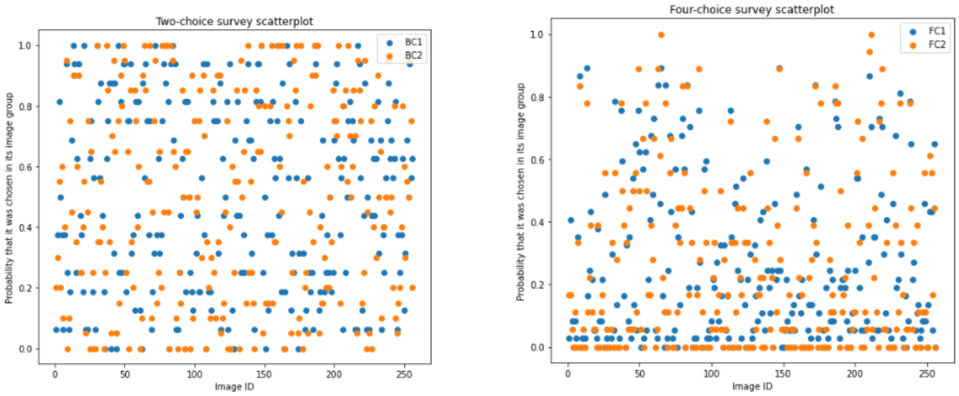
Figure 4 contains the JS divergence distance for all three groups. The distance between the binary-choice survey format was 0.081, the four-choice survey format was 0.024 and the average of the BC and FC survey formats was 0.099. This indicates that the image groupings have less effects on the four-choice formats than the binary-choice formats. It also indicates that the format we use effects can have an effect on the results of the rations of the images and the GAN.

This information corroborates the Chi-squared test of independence we conducted in section 4.3.1. The tests showed that there is not a statistically significant difference between responses of the four-choice format responses and there was a statistically significant difference in the other two groups. This pattern is seen in the results of the JS divergence distance results also. The similarity between the responses of the four-choice survey responses in higher than the other two groups.

3.3.3 Association of image scores using the Pearson Correlation Coefficient

Pearson Correlation Coefficient (r) is a statistical measure of the correlation between two data sets [14]. It ranges from -1 to 1 where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation [14]. In our case, since we are interested in the strength of the correlation rather than whether it is positive or negative, we will square the r value to keep it always positive. This will be denoted as r^2.

Using the Pearson Correlation Coefficient, we examined the correlation between image ratings of the binary-choice surveys (BC1 vs BC2) and the four-choice surveys (FC1 and FC2). This will help us understand how close the image ratings are between the two surveys of the two formats. The results can be seen in Figure 5.



(a) Binary-choice survey scatter-plot for all images

(b) Four-choice survey scatter-plot for all images

Figure 5: Binary-choice and Four-choice survey scatter-plot for all images. The Pearson's correlation coefficient squared (r^2) for the binary-choice format is: 0.342 and the four-choice format is 0.823

It can be seen that the correlation between the four-choice format (0.823) is greater than on the binary-choice format (0.342). This is another indication that image groupings between the survey formats have a less consequential effects on the results for the four-choice format than the binary-choice format. Moreover, this further support the results of Chi-squared test of independence analysis in section 4.3.1 and the results of the Jensen-Shannon Divergence distance measure in section 4.3.2.

3.4 Assessing the attitude of the participants towards the survey formats

For BC1, BC2 and FC2 the surveys, after the research question section, we added 10 additional questions with a format that is different from the research question section. So if the survey had a binary-choice format for the research questions, this section contained 10 additional questions with the four-choice format. In this section, the participants also had to chose the image that they found to be the most "pleasing to the eye". The purpose of this section was to get they acquainted and experienced with answering questions of a different format than the research questions. Afterward, three questions were asked to assess the attitude of the participants towards the two survey methods. The first question was about the format preference, the second questions was asked about which format allowed they to express their choice more effectively and accurately and lastly, which choice was less laborious in their opinion. This section's purpose was to understand the attitude of the participants towards the two survey formats. This is important as it can indicate how the crowdsourced workers feel towards the two survey versions. Furthermore, this can help improve the survey design further in the future.

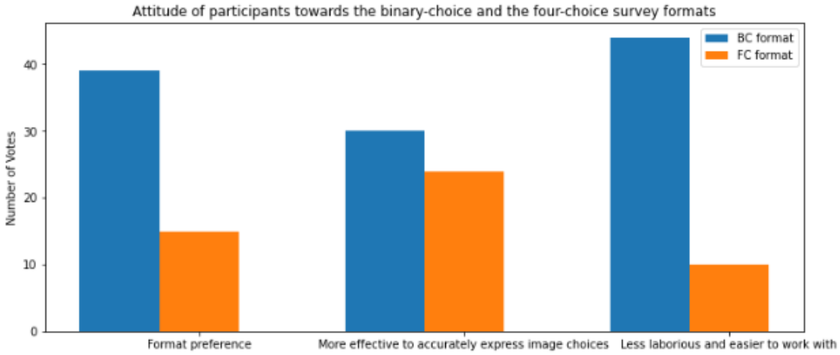


Figure 6: The attitude of participants towards the binary-choice and the four-choice survey formats

Figure 6 contains a bar chart of the attitude of participants towards the binary-choice and the four-choice survey formats. It can be seen that the binary-choice format was regarded as better than the four-choice format in all thee categories (preference, effectiveness and less laboriousness). The gap was the smallest in the effectiveness category. For the binary-choice format, the total number of votes was 30 and for the four-choice format, the total was 24. From this results, it can be inferred that the binary-choice format is slightly better than the four-choice format in being more effective to accurately express image choices. In contrast, the gap is largest in the less laborious category. This is probably because the participants have only two pictures to choose from in the binary-choice format instead of four in the four-choice format. Offering a larger amount of choice can be more strenuous for the participants. Overall, more than twice the people prefer the binary-choice format over the four-choice format. It can be concluded that the attitude of the participants towards the binary-choice is more positive than the four-choice format.

In general, it can be seen that participants prefer the binary-choice format more than the four-choice format. The four-choice format is deemed to be more laborious than the binary-choice one and this category had the largest difference between the two formats. Both of the formats had similar scores in the category for determining which format was more effective in letting participants accurately express their images choices. Lastly, in general, the binary-choice format was much more preferred than the four-choice format.

4 Responsible research

The experiments were designed in a way that allows repeatability and responsibility. The steps for creating the surveys and cleaning, processing, and analyzing the data, as well as the code, can be found on GitHub [13]. All the details are made available for other engineers to use and verify.

During the survey, no identifiable or personal information of the participants (e.g. name, age, gender, etc.) was taken. Furthermore, the survey contained a consent form informing the participants of the topic of the research and what their response data will be used for. In addition, an email was provided where they could make further inquiries about any concerning matter. Before starting the survey, the participants had to complete three example questions to become acquainted with the survey questions. This ensured that they understood the type of data we will collect. Once the real section of the survey started, they were notified.

[13] https://github.com/moshiur112/RPprojectcode

5 Conclusions and Future Work

5.1 Conclusion

In conclusion, despite both the survey formats having their merits and flaws, the four-choice format is superior than the binary-choice format. The effects of image grouping is significantly larger on the binary-choice format than the four-choice one and this affects the results of the image ranking. Furthermore, for the binary-choice format, if we are using images from four different GANs, there are six different possible combinations of images that can be shown per question. In contrast, for the four-choice format each question can contain different images from all different GANs. Since we can fit images from more different GANs in a single page, this means that the participants can select a clear winner in contrast to (potentially) picking potentially choosing a "less bad" image in the binary-choice format. The solution to this problem in the binary-choice format would be to more image grouping combinations and potentially show each image more than ones. This would ultimately make the data management more complex and possibly lengthen the questions in the surveys. In addition, according to out test of independence, Jensen-Shannon Divergence distance measure and the Pearson Correlation Coefficient, we found that the similarity in responses between the two versions of the four-choice format was greater than the versions of the binary-choice format and this has a statistically significant impact on the final result. This is another motivation of why the four-choice format is better than the two-choice format. Additionally, in the four-choice method, it is possible to fit more images per page which allows for a higher number of images to be rated compared to binary-choice format.

The only drawback of the four-choice format is that the attitude of participants towards the binary-choice format is more positive than the four-choice format. They find the binary-choice format to be more preferable, less strenuous, easier to work with and is a more effective format that allows them to express their choices more accurately.

5.2 Future Work

There are several ways this research could be extended in the future. Firstly, for each GAN, the images could be categorized according to the different biomes (e.g. desert, arctic, coasts, forests, etc) and compared the images of the same biomes produced by another GAN. If a GAN produces a more aesthetic picture for one biome than another, this strategy could reveal it. Secondly, it would be interesting to analyze the the concept of typicality vs novelty and to what extent people's choices would change after they are exposed to many of the same types of images and suddenly shown new ones. One example of this could be to show participants images of one biome only for many questions and then show them a mix of images of the first biome and another biome. Afterward analyze whether the choice they make is the same choice they would make if there were initially not shown many pictures of the first biome. Thirdly, we could introduce another tests for measuring consistency of the participant's choices by showing them questions that were show in a previous section of the survey and comparing whether they choose the same image. This could help us understand how consistent the participants are with their choices. Fourthly, in order to ensure the effects of the image groupings are less influential on the rating results, especially for the binary-choice survey, we could create many more versions of the surveys where the image groupings are randomized. Ideally, the best results would come if we could create a new new image grouping for every single participant, but this would be computationally expensive to create and the data would be harder to process.

References

[1] *A beginner's guide to generative adversarial networks (gans)*. 2020. URL: https://wiki.pathmind.com/generative-adversarial-network-gan.

[2] Abhishek Agrawal, Vittal Premachandran, and Ramakrishna Kakarala. *Rating image aesthetics using a crowd sourcing approach*. Springer. 2013.

[3] *Chi-square goodness of fit test*. 2022. URL: https://www.jmp.com/en_sg/statistics-knowledge-portal/chi-square-test/chi-square-goodness-of-fit-test.html#:~:text=What%20is%20the%20Chi%2Dsquare,representative%20of%20the%20full%20population..

[4] *Chi-square test of Independence*. 2022. URL: https://www.jmp.com/en_au/statistics-knowledge-portal/chi-square-test/chi-square-test-of-independence.html.

[5] Yubin Deng, Chen Change Loy, and Xiaoou Tang. "Image aesthetic assessment: An experimental survey". In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 80–106.

[6] B atool Esmaeeli et al. "The Optimal Number of Choices in Multiple-Choice Tests: A Systematic Review". In: *Int J Pediatr* 2.3 (2021).

[7] Kazuki Fujii et al. "HumanGAN: generative adversarial network with human-based discriminator and its evaluation in speech perception modeling". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 6239–6243.

[8] Renu Khandelwal. *Techniques to measure probability distribution similarity*. June 2021. URL: https://medium.com/geekculture/techniques-to-measure-probability-distribution-similarity-9145678d68a6.

[9] Jonathan Livescault. *6 great advantages of crowdsourcing you can benefit from*. May 2022. URL: https://www.braineet.com/blog/crowdsourcing-benefits.

[10] *Overview of gan structure nbsp;|nbsp; generative adversarial networks nbsp;|nbsp; google developers*. Feb. 2021. URL: https://developers.google.com/machine-learning/gan/gan_structure.

[11] Tiago Rosa dos Reis. *Measuring the statistical similarity between two samples using Jensen-Shannon and Kullback-Leibler...* Sept. 2020. URL: https://medium.com/datalab-log/measuring-the-statistical-similarity-between-two-samples-using-jensen-shannon-and-kullback-leibler-8d05af514b15.

[12] Krista Reuther. *The only 4 question types you need to know (and how to use them!)* URL: https://www.centiment.co/blog/the-only-4-question-types-you-need-to-know#:~:text=An%20attention%20check%20question%20is,your%20data%20analysis%20post%2Dcollection.

[13] Melanie Revilla and Jan Karem Höhne. "How long do respondents think online surveys should be? New evidence from two online panels in Germany". In: *International Journal of Market Research* 62.5 (2020), pp. 538–545.

[14] Vinay Singh. *Pearson correlation, a mathematical understanding!* Jan. 2019. URL: https://medium.com/@SilentFlame/pearson-correlation-a-mathematical-understanding-c9aa686113cb.

[15] Frederik Ueberschär. "AI FOR EXPERIENCE: Designing with Generative Adversarial Networks to evoke climate fascination". In: (2021).

[16] *Writing survey questions*. Oct. 2021. URL: https://www.pewresearch.org/our-methods/u-s-surveys/writing-survey-questions/.

A Appendix

GAN name	BC1	BC2
TCDNE	310	239
Satellite	472	631
Baseline	649	796
GAN250	617	894

Table 7: Image count for the binary-choice format for the different GANS

GAN name	BC1	BC2
TCDNE	155	63
Satellite	598	278
Baseline	723	358
GAN250	892	453

Table 8: Image count for the four-choice format for the different GANS

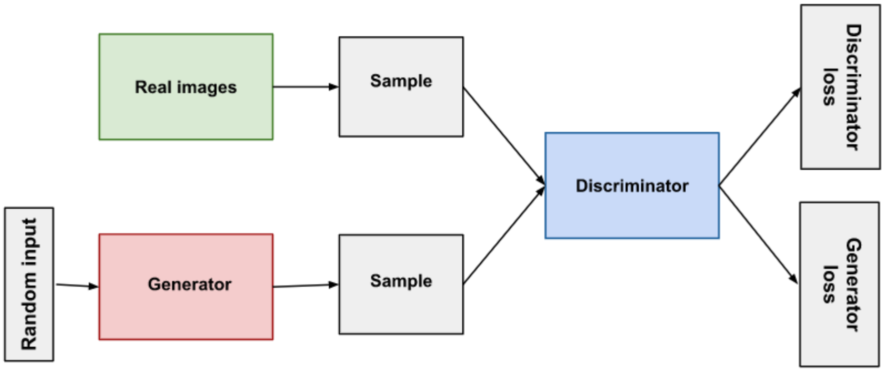


Figure 7: Regular GAN diagram with neural network discriminators [10]

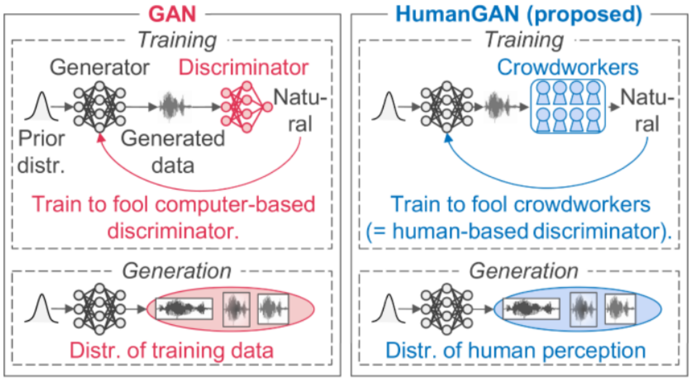


Figure 8: Difference between regular GAN and HumanGAN. The generator of a regular GAN gets feedback from a neural network based discriminator, whereas the generator of a HumanGAN gets feedback from humans [7]

GAN name	FC1	FC1 shown	FC2	FC2 shown	BC1	BC1 shown	BC2	BC2 shown
TCDNE	155	2368	63	1152	310	1024	239	1280
Satellite	598	2368	278	1152	472	1024	631	1280
Baseline	723	2368	358	1152	649	1024	796	1280
GAN250	892	2368	453	1152	617	1024	894	1280

Table 9: Image count and the number of images shown of different GANs for the FC1, FC2, BC1 and BC2 surveys

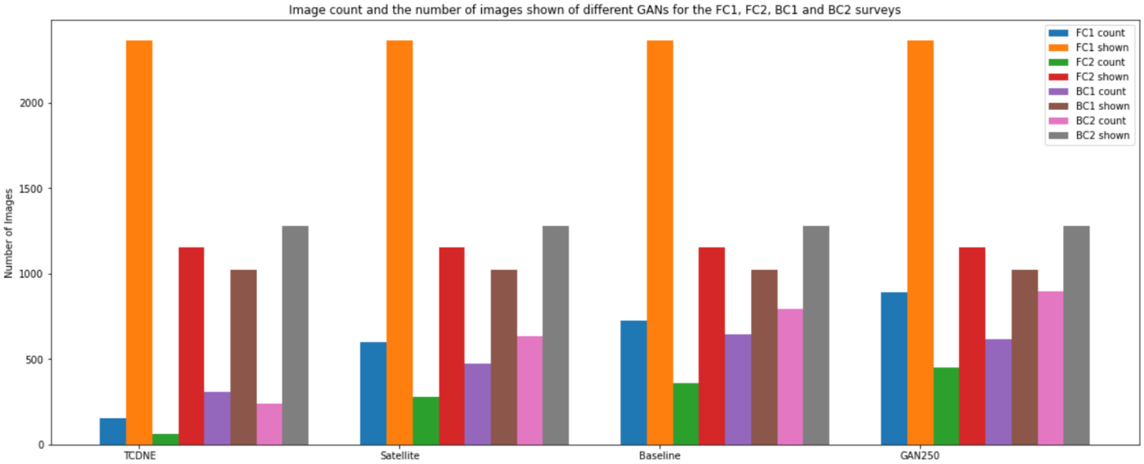


Figure 9: Bar graph of the image count and the number of images shown of different GANs for the FC1, FC2, BC1 and BC2 surveys. This shows that for all surveys, despite showing images from all GANs equal number of times, GAN250 clearly produced more aesthetic images and TCDNE produced the least aesthetic images. This can be seen by looking at the distances between the count and the shown bars for all surveys of the GANs

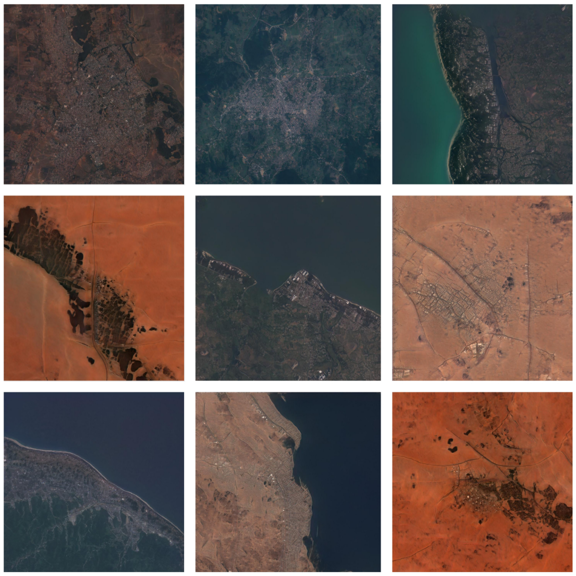


Figure 10: Examples images from the TCDNE GAN

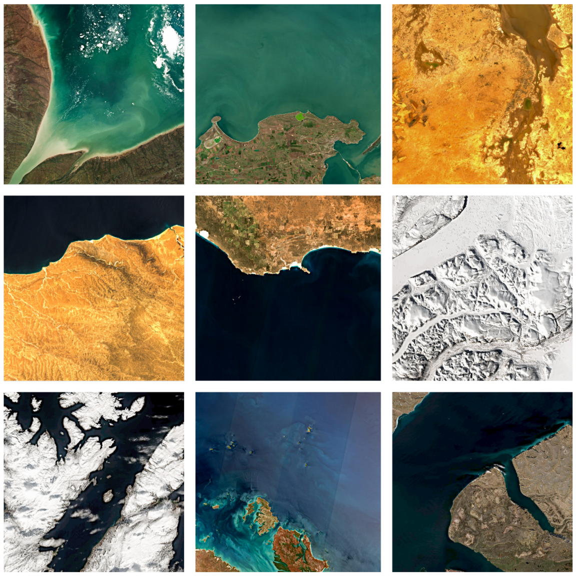


Figure 11: Examples images from the Satellite GAN

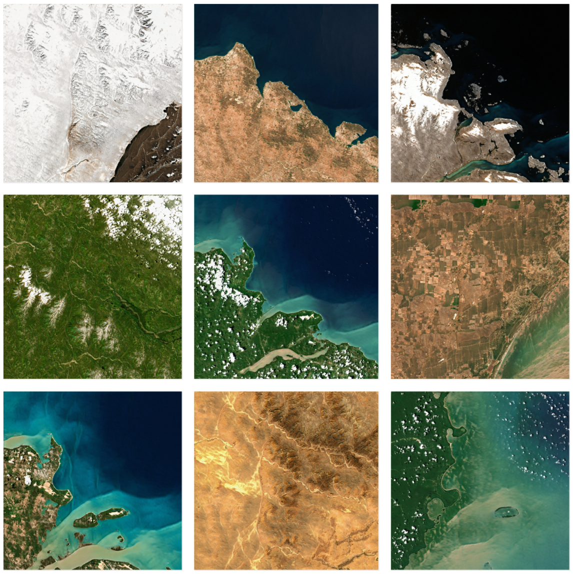


Figure 12: Examples images from the Baseline GAN

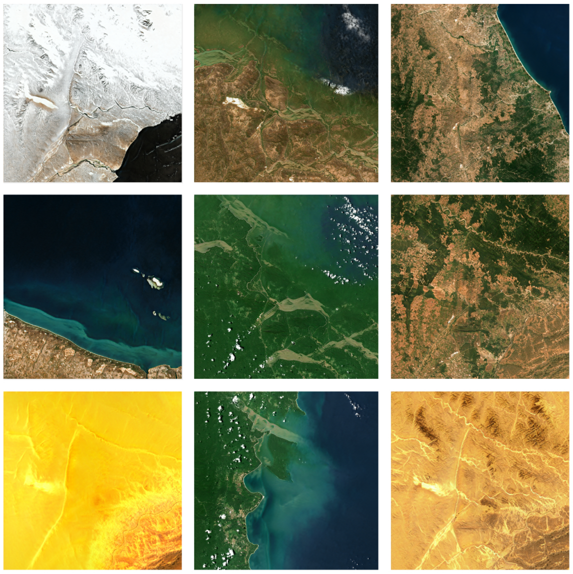
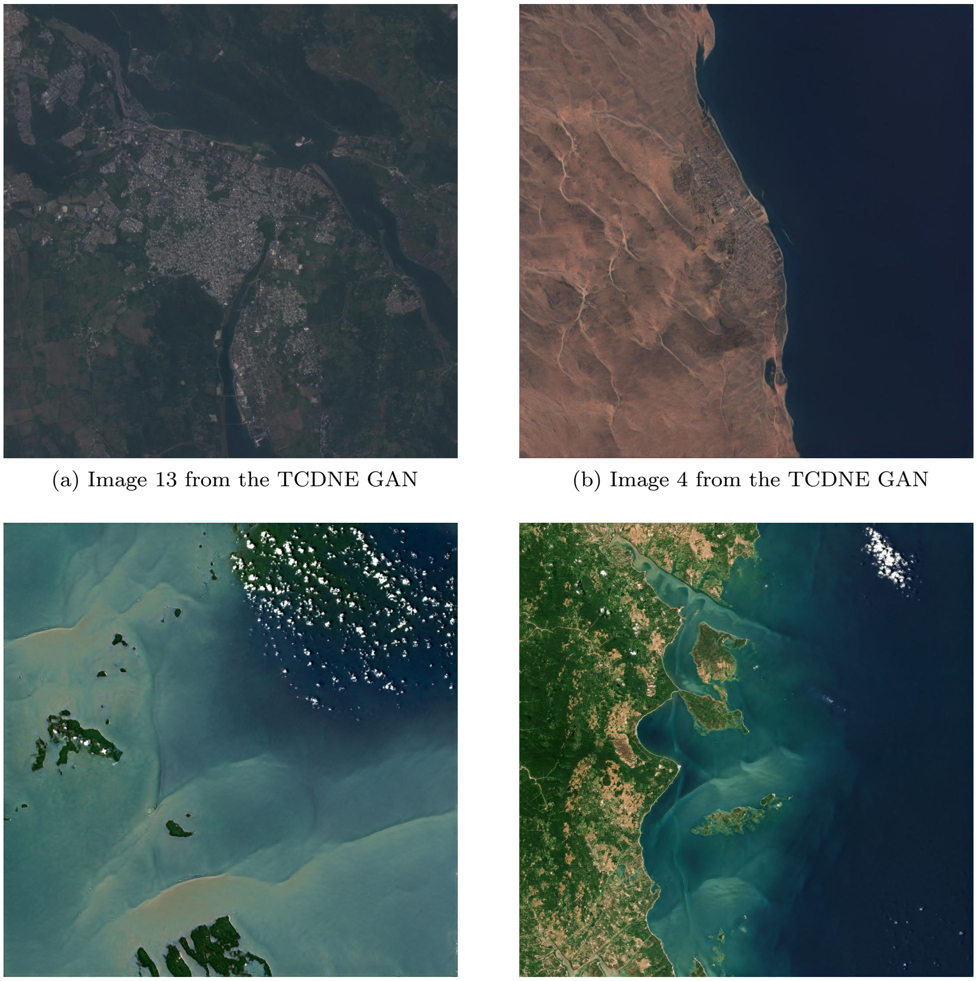


Figure 13: Examples images from the GAN250 GAN



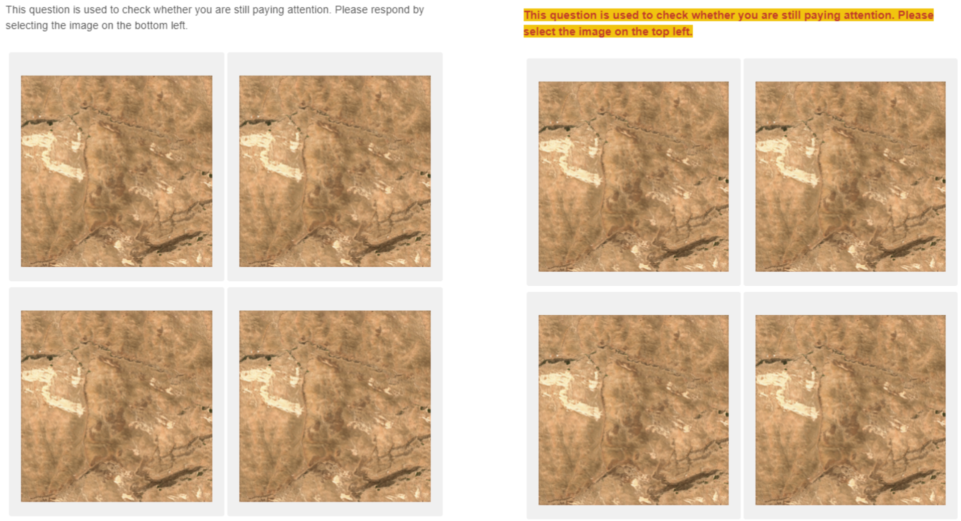
(a) Image 13 from the TCDNE GAN

(b) Image 4 from the TCDNE GAN

(c) Image 13 from the Baseline GAN

(d) Image 13 from 20 from the GAN250 GAN

Figure 14: Display of the differently ranked images from the different GANs. In all four surveys, a was ranked the lowest, the b and c, and finally d was ranked the highest



This question is used to check whether you are still paying attention. Please respond by selecting the image on the bottom left.

This question is used to check whether you are still paying attention. Please select the image on the top left.

(a) Old attention checking question format that was used for FC1 only

(b) New attention checking question format that was used for FC2, BC1 and BC2

Figure 15: In the old attention checking question format, only 37 out of 48 responses had one or zero mistakes (78%). In contrast, with the newer format, 100% of the participants had one or zero mistakes