Measuring Surprisal for Style Transfer and GAN Generated Images

Mona Abdelrahman
Massachusetts Institute of Technology
monaabd@mit.edu

Holly Rieping
Massachusetts Institute of Technology
hrieping@mit.edu

Abstract

Generative adversarial networks (GANs) learn to generate new data based on the data they were trained on. In this work, we train a DCGAN to generate images from scratch. We also use style transfer methods to modify existing images to generate a new image. However, there is no one true way to evaluate a generated image. As such, we use an Inception network to give each generated image an Inception score (IS). The inception score calculates a score for the network based on image quality and diversity. Once we have the inception scores, we survey humans to give a score for each image based on the same criteria. We then compare the computed score to the human responses across our generation methods and image classes and identify correlation between human satisfaction with images and their inception score.

1. Introduction

Currently, there is no objective way to measure or evaluate human surprisal, or how "surprised" someone would be by looking at something, for generative images. Many different strategies exist to score the performance of generative images, one of the most popular being the Inception Score [5, 7]. A non-automated popular strategy is to have humans annotate images with aspects that surprised them or made the most sense to them.

For our project, we wanted to see if any correlation existed between automated, computed scores like the Inception Score or the conditional probabilities assigned to images when they are classified by the Inception model and human-given scores found from surveying humans on generated images. Within this question, we wanted to see how such a correlation would vary between different image generation techniques like Generative Adversarial Networks (GANs) [3] or applying a style transfer to an image [2]. We also wanted to see how the correlation may vary between indoor generated scenes and outdoor generated scenes.

To explore these questions, we created two GANs: one that generates bedroom images, and one that generates mountain images. We also applied a style transfer to images of bedrooms and images of mountains. We then found the Inception Scores for various combinations of these sets of images, as well as the conditional probabilities for each image of being either a bedroom or a mountain. Once we had all of our computer-given data, we surveyed 235 humans to collect our human-data on our indoor vs. outdoor and style transfer vs. GAN image comparisons.

With our approach, we hope to find correlations that will allow us to predict some form of human surprisal for our various image generation techniques, or find that no such correlations exist with the computer-provided metrics we have chosen to explore.

2. Related Work

2.1. Places365 Dataset

We chose to train our GANs on classes from the Places365 dataset [8]. We chose this dataset because we wanted to generate scenes rather than just objects so we could compare scene recognition between humans and computer scores. The dataset also has a wide variety of both indoor and outdoor classes we could choose from. This dataset has a wide variety of image types in every image class, including a mix of natural everyday scenes and more professionally done/standard stock photo images.

2.2. GANs and DCGAN

A Generative Adversarial Network (GAN) is a type of neural network designed by Goodfellow et al. [3] where there are effectively two networks "competing" in a zero sum game (where one network must win and another must lose) to generate sets of objects that have the same statistics and features as the original training set provided. The two networks in a GAN are a generator and a discriminator. The discriminator is a network that tries to distinguish between the training data and the images created by the generator. The generator effectively learns by trying to have the discriminator label it's output as a real image. GANs are widely used to be able to generate more data for additional training, simulating scenarios, or even in animation. In this

project we will use two GAN to first create images of an indoor scene and then of an outdoor scene. These images will then be used to measure how surprised humans see them as and how "unnatural" they are considered to be.

In our paper, we are using a specific type of GAN called a Deep Convolutional GAN (DCGAN). The main difference here is that a DCGAN is specifically made to generate images and is composed of mostly convolutional layers.

2.3. Style Transfer

We will be adapting our style transfer code from pset 5 of this class to generate a set of images from existing images, rather than from scratch like our GAN. We chose to include style transfer as a way to compare the Inception scores and human responses for images generated from scratch and images generated from existing images. This can also provide additional insight into what is considering to be "unnatural" in the view of human subjects. For example, it may be worth exploring if something with an artistic style but a scene that makes sense is considered to be more natural than a picture created by a GAN that maybe looks normal but has furniture placement that doesn't make sense to a human.

Our pset 5 code follows the Neural-style Algorithm described by Gatys et al. [2]. The Neural-Style algorithm takes in a style image, such as a painting or a texture image, and an input image that will have the style applied to it, and outputs the style-transferred image. The input image is first reconstructed from the `conv1_1`, `conv2_1`, `conv3_1`, `conv4_1`, and `conv5_1` layers from the VGG-Network. The style image is then reconstructed from subsets of the `conv1_1` through `conv5_1` layers of the CNN. The output image is then synthesized by matching the reconstructed input image (content representation) with the reconstructed style image (style representation).

2.4. Inception Model and Inception Score

We will be using the Inception v3 model described by Szedgedy et al. [5] to calculate the conditional label distribution $P(y|x)$ [7]. The pretrained inception_v3 model included with pytorch is trained on ImageNet classes, so Inception scores calculated from this model would have a range [1,1000] since the model supports 1,000 classes. A higher score indicates that the generated set of images has a diverse set of images that distinctly look like the classes. Once we have the conditional label distribution for a set of generated images, given to us by the Inception v3 model, we can calculate the Inception score for the set using the equations outlined by Salimans et al. [7].

We get the conditional label distribution $P(y|x)$ by running a set of images through the Inception model. We then pull the value $P(y)$ from the distribution so we can calculate the Kullback-Leibler Divergence (KL-Divergence). The KL-Divergence is found as:

$$KLD = P(y|x) * [log(P(y|x)) - log(P(y))]$$

for each class for each image in the set. We then sum all of the KL-Divergence values for the classes, and average that over the number of images in the set:

$$average\_KLD = sum(KLD)/number\_of\_images$$

Then, we undo the logs we did in the KL-Divergence by setting the average to an exponential to get the inception score:

$$inception\_score = exp(average\_KLD)$$

2.5. Human Perception of Generated Images

Since one of our goals was to measure human surprisal on the different generated images, we looked into previous work on gauging human perception of various images. As mentioned in our introduction, one of the most popular techniques for evaluating generative images is by having humans annotate images in something like an Amazon Mechanical Turk study [5]. This technique allows participants to mark what aspects of the image look the most or least realistic to them, and this information could be used to fine tune the generative technique to best improve the realism of the images it generates. However, we wanted to explore correlations between human surprisal and automatic computer-given scores, so we decided to not survey for image annotations. Rather, we chose to have humans mimic the conditional probability classification of the Inception model as well as give their own ranking of how natural each image was.

Previous studies into human perception [1, 6] found that the gist of an image is more prominent in human perception than specific objects and that for images of a lower resolution, holistic processing of the image becomes more prominent. As such, we decided to generate images of scenes rather than objects so that our human survey participants would be asked to recognize and score the images more holistically as the scene rather than a single, specific object.

3. Methods

3.1. Dataset

We chose to train one GAN on the Bedroom class from the Places365 dataset [8] as our indoor scene and another GAN on the Mountain class as our outdoor scene. Each class had 5000 total images in the training set and 100 in the validation set, so we divided those 5,100 images into 75% training images (3,825 images), 15% test images (765 images), and 10% validation images (510 images). We used this split to ensure that there were enough images for each class to be trained on for all of our models, but also enough to get an accurate representation during the validation and testing phases.

3.2. DCGAN Image Generation

We created two GANs, one that was trained on the mountain class from the Places365 dataset and generated mountain images, and one that was trained on the bedroom class from the Places365 dataset and generated bedroom images.

For our generator, we used 5 convolutional 2D transpose layers. Layers 1-4 were each followed by a batch normalization layer (with the size of the previous layer's output) and a ReLU layer. All of the convolutional layers used no bias. We then had one final convolutional 2D transpose layer followed by a tanh as our output layer.

- The first convolutional 2D transpose layer has 100 input channels, 512 output channels, a kernel size of 4, stride of 1, and no padding.
- The second convolutional 2D transpose layer has 512 input channels, 256 output channels, kernel size of 4, stride of 2, and zero padding size of 1.
- The third convolutional 2D transpose layer has 256 input channels, 128 output channels, kernel size of 4, stride of 2, and zero padding size of 1.
- The fourth convolutional 2D transpose layer has 128 input channels, 64 output channels, kernel size of 4, stride of 2, and zero padding size of 1.
- The final convolutional 2D transpose layer has 64 input channels, 3 output channels (one for each RGB channel), kernel size of 4, stride of 2, and zero padding size of 1. This was followed by a Tanh activation function layer.

For our discriminator, we had 5 convolution 2D layers with no bias. Layers 2-4 were followed by a batch normalization 2d layer (with the size of previous layer's output) and a leaky ReLU layer with a negative slope value of 0.2. Each convolution 2D layer is as follows:

- The first layer has 3 input channels (one for each RGB channel), 64 output channels, a kernel of size 4, a stride of 2, and a zero padding size of 1.
- The second layer has 64 input channels, 128 output channels, a kernel of size 4, a stride of 2, and a zero padding size of 1.
- The third layer has 128 input channels, 256 output channels, a kernel of size 4, a stride of 2, and a zero padding size of 1.
- The fourth layer has 256 input channels, 512 output channels, a kernel of size 4, a stride of 2, and a zero padding size of 1.
- The fifth layer has 512 input channels, 1 output channels, a kernel of size 4, a stride of 1, and no padding.

For each GAN, we trained it for 700 epochs with a learning rate of 0.0002. We used an Adam optimizer [4] with a beta1 value of 0.5 and a beta2 value of 0.9999. The loss that was used was a Binary Cross Entropy Loss. We stopped training each GAN once the generator was not able to decrease in loss in a way comparable to the discriminator. We used a low learning rate to ensure that none of the errors in an epoch were too influential, and this also caused us to have a high number of epochs. We used Binary Cross Entropy loss since the discriminator is effectively classifying between two classes (real vs generated).

Each GAN generated 64, 64x64 images, as right now GANs can typically only make up to 128x128 sized images, and we wanted to balance between image quality/size and the ability for the GAN to produce images that looked similar to their class with limited resources.

3.3. Style Transfer Images

For our style transfer images, we implemented the Neural-Style algorithm as described by Gatys et al. in "A Neural Algorithm of Artistic Style" [2]. For our code, we altered the PyTorch tutorial "Neural Transfer Using PyTorch" written by Alexis Jacq that we followed in this class' pset 5.

We input a "style" image and a set of images we want to transfer the style to. We then output the set of images with the given style integrated into them.

For our style image, we chose Vincent Van Gogh's *The Starry Night* (1889) as his distinct brush stroke style and the color scheme of this particular piece are very familiar and recognizable to most people. We tested other famous art pieces like Pablo Picasso's *The Weeping Woman* (1937), Leonardo da Vinci's *Vitruvian Man* (1490), and a piece from Claude Monet's *Water Lilies* series (1920-1926), but found that Starry Night gave us the best range of conditional probabilities for our images while maintaining both the image appearance and artistic style.

We wanted to have a consistent image source throughout the project, so we chose our sets of images to apply the style transfer to randomly from the test set of our dataset we trained our GANs on. We chose 15 images from the bedroom class test set and 15 images from the mountain class test set.

As mentioned in section 2.4, the Neural-Style algorithm reconstructs the input image from the different conv layers of the VGG-Network. We chose to reconstruct our images using the conv5 layer of the network to preserve more of the image content in the style transfer rather than using a higher conv layer and losing image content in favor of adding artistic style. We found that the conv1 layer kept the best balance of image content and image style for our intended purpose.

3.4. Inception Model and Score

Initially, we planned on using the pretrained Inception v3 model to evaluate our bedroom and mountain GAN and style transfer images, but we decided that the [1,1000] inception score range the pretrained model gave us would not be representative of our images since we only have two classes. So, we retrained the Inception v3 model on our bedroom and mountain classes using the same training dataset that the GANs were trained on.

We retrained the pretrained model on our two classes using the following parameters: two epochs, the SGD Optimizer, Cross Entropy loss, batch size of 64, learning rate of 0.001, and momentum of 0.9. We only used two epochs since we only had two classes and wanted to avoid overtraining the pretrained model. The learning rate was 0.001 because we only needed the model to recognize two new classes as the model was already pretrained.

After retraining the Inception model, our inception score range became [1,2] since it was retrained on two classes and each image could only be classified as a bedroom or a mountain. An Inception score closer to 2 means the set of images being scored had a near-equal representation of bedroom and mountain images and that the model had high probabilities of each image being either a mountain or a bedroom. Conversely, an Inception score closer to 1 means that the set of images being scored could lack diversity (e.g. the set was only of bedrooms or the set was only of mountains) or the images could have low probabilities of being either a mountain or a bedroom (e.g. the model could think an image has a near equal probability of being a mountain or a bedroom).

We created a variety of sets to have Inception scores that are representative of different aspects of our project. We created a set of all 128 GAN generated images (64 bedroom images and 64 mountain images) to score our GANs. We created a set of all 30 style transfer generated images (15 bedroom images and 15 mountain images) to score our style transfer generation.

For our human survey, we created three sets of images in four categories, for a total of twelve sets of images. The four categories were: only style transfer images, only GAN images, style transfer bedroom images and GAN mountain images, and GAN bedroom images and style transfer images. Each set had two bedroom images and two mountain images. The sets were created using images with high (above 0.95), medium (roughly 0.8), and near-equal (near 0.5) probabilities of being a mountain or a bedroom such that we could have a set with a high Inception score, a medium Inception score, and a low Inception score.

3.5. Human Subject Surveys

We wanted to gather as much comparative human data as possible, so we designed four surveys that correlated with our four image categories and sets as listed in section 3.4:

1. only style transfer images
2. only GAN images
3. style transfer bedroom images and GAN mountain images
4. GAN bedroom images and style transfer mountain images

Each survey had three sections with four images each, for a total of twelve images per survey. Two of the images in each section were of bedrooms, and the other two images were of mountains. The three sections correlated to the set in that category that had a high Inception score, medium Inception score, and a low Inception score.

For example, the third survey (style transfer bedroom images and GAN mountain images) had three sections that had two style transfer bedroom images and two GAN mountain images each. Appendix A further explains how the images were distributed among the four surveys.

For each individual image, we asked:

1. How natural does the image look? On a scale of 1-10, 1 being "very unnatural" and 10 being "very natural"
2. How much does the image look like a bedroom or a mountain? On a scale of 1-10, 1 being "bedroom" and 10 being "mountain"

For each section of four images, we asked participants to rank the four images from most to least natural.

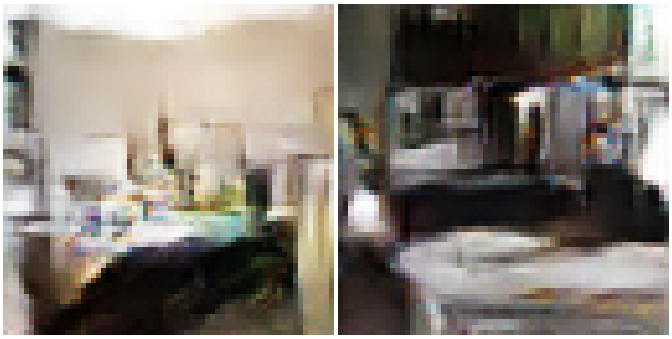
Since we wanted to measure surprisal, we defined "natural" at the beginning of each survey to be "something you easily recognize without objects or sections that you would consider to be odd, surprising, or inconsistent with the image." We asked participants to measure on a scale how much each image looked like a bedroom or mountain to mimic the classification our retrained Inception v3 model did on each image. We asked participants to rank the images from most to least natural in addition to their natural scores for each image in order to see how comparison affected their idea of "natural".

3.6. Data Analysis

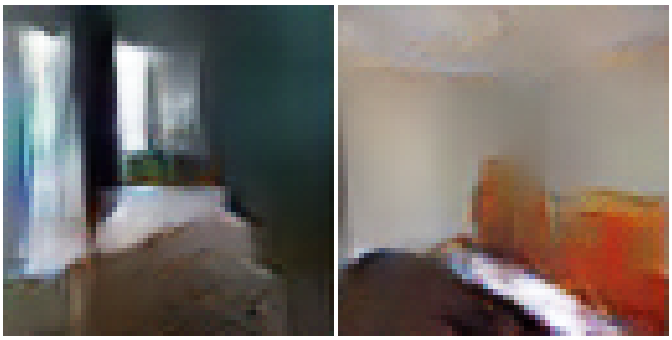
Overall, we had a few key categories we wanted to compare data in. First, we wanted to compare the retrained Inception v3 model conditional probabilities with human scores of how much each image looked like either a bedroom or a mountain in order to see which most accurately classified the images. Second, we wanted to compare data for indoor images against outdoor images to see which class performed better for humans and the retrained Inception v3 model. Third, we wanted to compare GAN generated

images against our style transfer images to see which generation technique performed better for humans and the retrained Inception v3 model. Finally, we wanted to measure human surprisal for each type of image to see which images were the most and least natural as well as the most accurate.

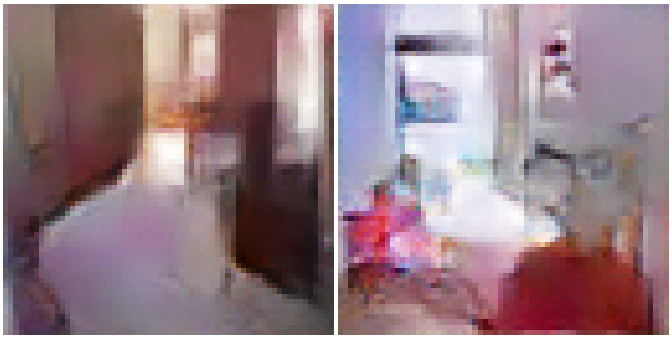
4. Results



(a) gan_bedroom_4 (b) gan_bedroom_15



(c) gan_bedroom_25 (d) gan_bedroom_43

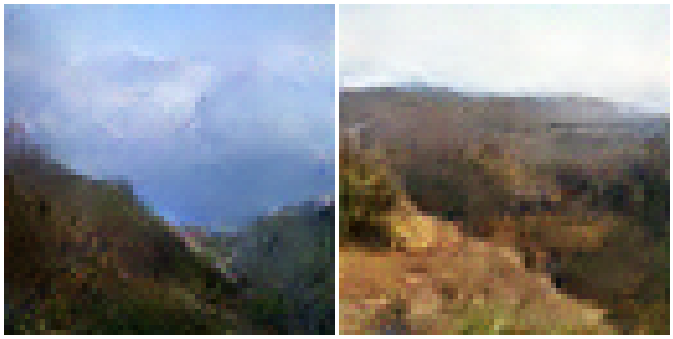


(e) gan_bedroom_50 (f) gan_bedroom_62

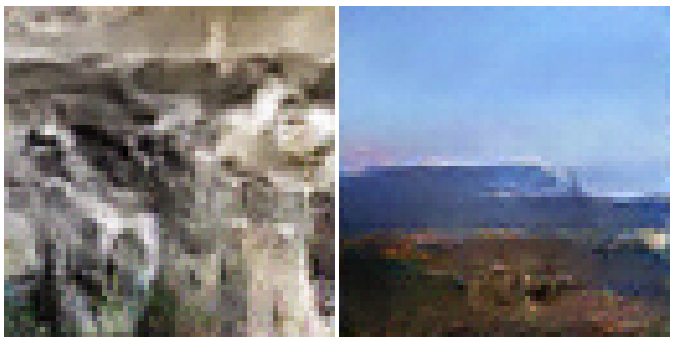
Figure 1: GAN Bedroom Images

4.1. GAN Image Reception

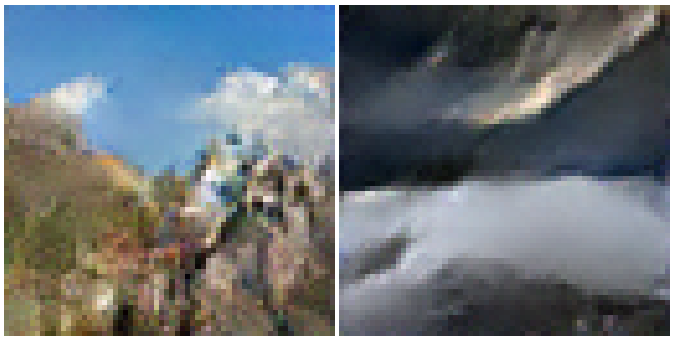
We generated 64 images from our GAN trained on the bedroom class, and 64 images from our GAN trained on the mountain class. Six of these bedroom images that we



(a) gan_mountain_11 (b) gan_mountain_21



(c) gan_mountain_22 (d) gan_mountain_37



(e) gan_mountain_41 (f) gan_mountain_61

Figure 2: GAN Mountain Images

used in our survey are in Figure 1, and six of these mountain images that we also used in our survey are in Figure 2. Running our Inception score code on all 128 GAN images we generated, we got an Inception score of 1.3716205, in the range [1,2].

For our second survey of all GAN images, our high Inception score set of images had an Inception score of 1.8755617, our medium Inception score set had a score of 1.33302, and our low Inception score set had a score of 1.0380573. Appendix A lists which images from Figures 1 and 2 were used for survey 2.

The graph "Section 2: Human vs. Inception P(correct

class)" [Figure 3a] graphs the probability the retrained Inception model gave each image for the correct class the image is against the average human probability computed from the survey results. Comparing these results to the $y = x$ line of equality, we see that about half of the images perform better with humans than the retrained Inception model, while the other half underperform with humans. Specifically, four of the six bedroom images had a higher probability of being classified as a bedroom with humans than with the Inception model, and five of the six mountain images had a higher probability of being classified as a mountain with the Inception model than with humans.

The "Section 2: 'Natural' Scores Histogram" [Figure 3b] is a frequency histogram of the average "natural" score given to each image by humans in the survey. The bins are left-closed, so the '4' bin contains the continuous numbers [4,5). This distribution shows that mountain images generated by the GAN appeared more "natural" to survey participants than bedroom images as mountain images have a bin range [4,7] while bedroom images have a bin range [3,4].

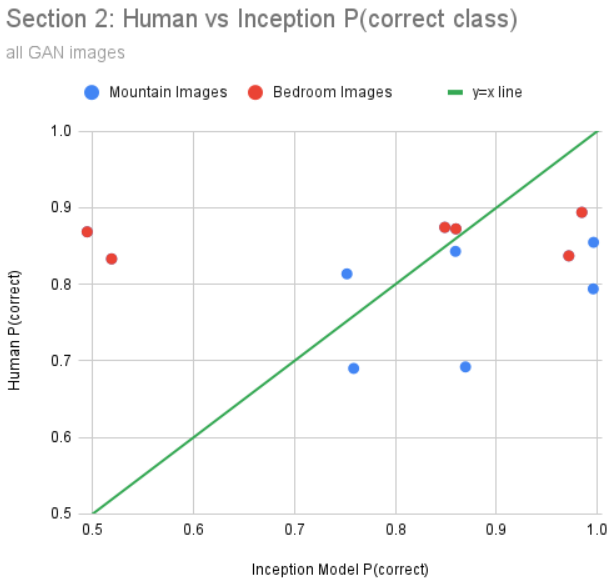
4.2. Style Transfer Image Reception

We generated 15 bedroom images and 15 mountain images using our style transfer code. Six of these bedroom images that we used in our survey are in Figure 4, and six of these mountain images that we also used in our survey are in Figure 5. Running our Inception score code on all 30 of our style transfer images, we got an Inception score of 1.3192803 in the range [1,2].

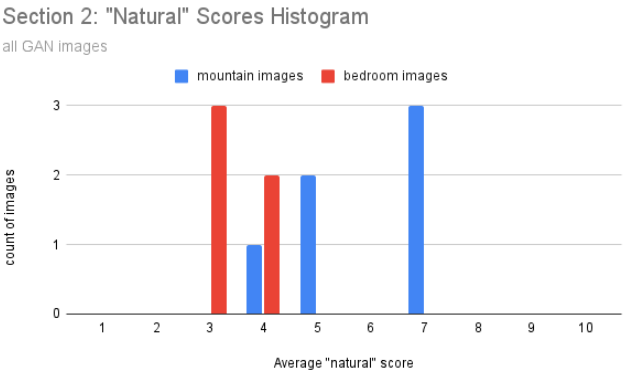
For our first survey of all style transfer images, our high Inception score set of images had an Inception score of 1.7898717, our medium Inception score set had a score of 1.2647859, and our low Inception score set had a score of 1.0060046. Appendix A lists which images from Figures 4 and 5 were used for survey 2.

The graph "Section 1: Human vs. Inception P(correct class)" [Figure 6a] graphs the probability the retrained Inception model gave each image for the correct class the image is against the average human probability computed from the survey results. Comparing these results to the $y = x$ line of equality, we see that ten out of twelve of the images were given higher probabilities of being the correct class by humans than the Inception model. Overall, for style transfer images, humans were more certain in their classifications for both classes than the retrained Inception model.

The "Section 1: 'Natural' Scores Histogram" [Figure 6b] is a frequency histogram of the average "natural" score given to each image in the section by humans in the survey. The bins are left-closed, so the '4' bin contains the continuous numbers [4,5). This distribution gives mountain images an even distribution across the bin range [5,7] and bedroom images an almost normal distribution across the bin range [6,8]. Overall, humans found style transfer bedrooms to be slightly more natural than style transfer mountains, but not by much.



(a) Section 2: Human vs. Inception P(correct class) scatter plot



(b) Section 2: "Natural" Scores Histogram

Figure 3: Section 2 Graphs

4.3. Comparison of GAN and Style Transfer Images

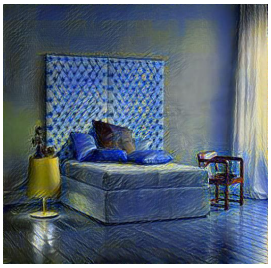
Our third survey compared style transfer bedroom images with GAN mountain images. Our high Inception score set had a score of 1.8774519, our medium Inception score set had a score of 1.3349577, and our low Inception score set had a score of 1.0455266.

The graph "Section 3: Human vs Inception P(correct class)" [Figure 7a] graphs the probability that the retrained Inception model gave each image for the correct class that the image is against the average human probability com-

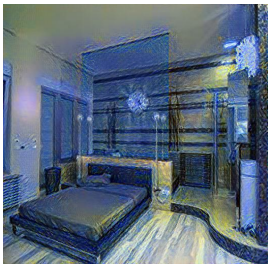
(a) starry_b1

(b) starry_b4

(c) starry_b5



(d) starry_b8

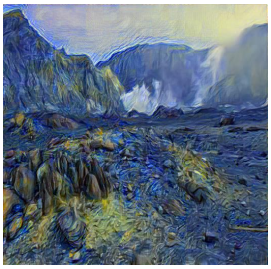


(e) starry_b12

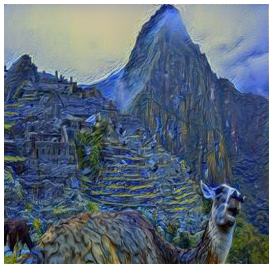
(f) starry_b13

Figure 4: Style Transfer Bedroom Images

(a) starry_m1



(b) starry_m4



(c) starry_m5

(d) starry_m9

(e) starry_m10

(f) starry_m14

Figure 5: Style Transfer Mountain Images

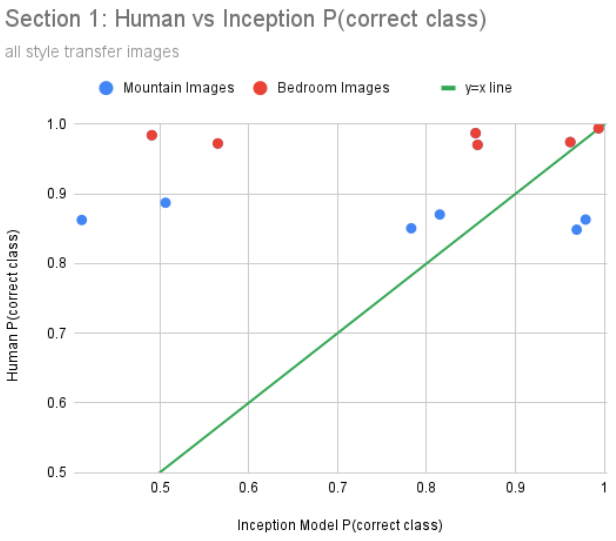
puter from the survey results. Comparing these results to the $y = x$ line of equality, we see that humans outperformed the Inception model on all of style transfer bedroom images, but the Inception model outperformed humans on all but one of the GAN mountain images.

The graph "Section 3: 'Natural' Scores Histogram" [Figure 7b] is a frequency histogram of the average "natural" score given to each image in the section by humans in the survey. The bins are left-closed, so the '4' bin contains the continuous numbers [4,5). The distribution shows that humans found the style transfer bedroom images significantly more natural than the GAN mountain images, as the style transfer bedroom images were roughly evenly spread between the bin range [7,8] and the GAN mountain images were spread between the bins [3,6] with a right skew. The two distributions have no overlap, showing that the style transfer bedroom images were consistently rated more natural than the GAN mountain images.

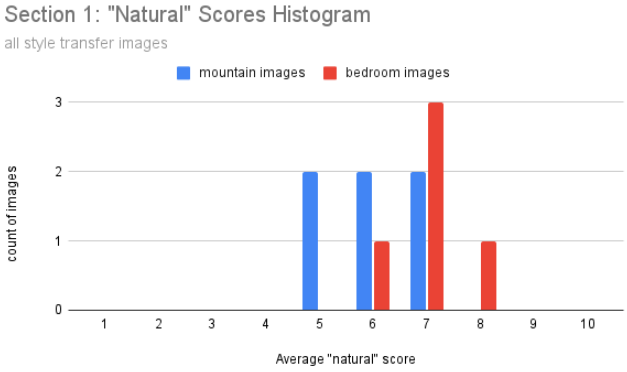
Our fourth survey compared GAN bedroom images with style transfer mountain images. Our high Inception score set had a score of 1.7880352, our medium Inception score set had a score of 1.2630899, and our low Inception score set had a score of 1.0035154.

The graph "Section 4: Human vs Inception P(correct class" [Figure 8a] graphs the probability that the retrained Inception model gave each image for the correct class that the image is against the average human probability computer from the survey results. Comparing these results to the $y = x$ line of equality, we see that humans outperformed the Inception model on four of six of the style transfer mountain images, and two of the GAN bedroom images. Humans and the Inception model performed equally on two of the GAN bedroom images. The Inception model outperformed humans on two of the GAN bedroom images and two of the style transfer mountain images. Overall, humans performed better than the Inception model, but especially performed better with the style transfer images.

The graph "Section 4: 'Natural' Scores Histogram" [Fig-

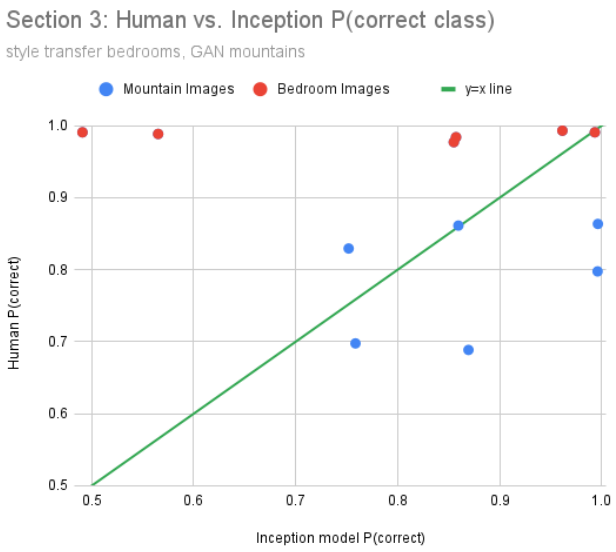


(a) Section 1: Human vs. Inception P(correct class) scatter plot

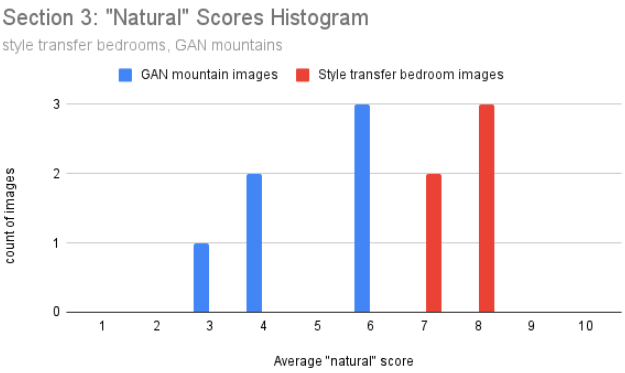


(b) Section 1: "Natural" Scores Histogram

Figure 6: Section 1 Graphs



(a) Section 3: Human vs. Inception P(correct class) scatter plot



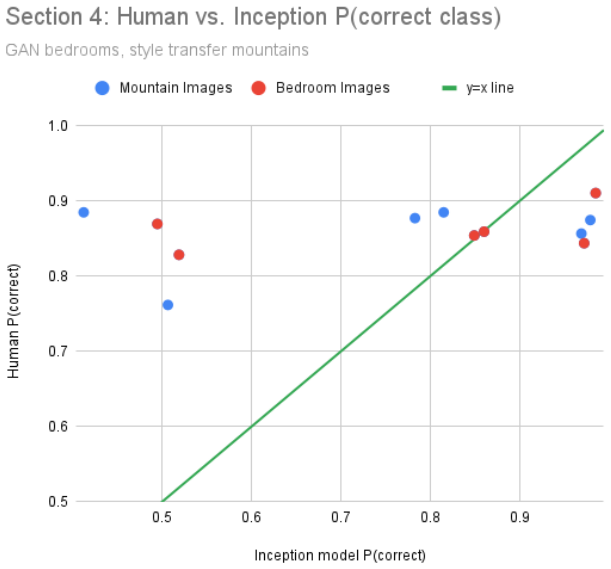
(b) Section 3: "Natural" Scores Histogram

Figure 7: Section 3 Graphs

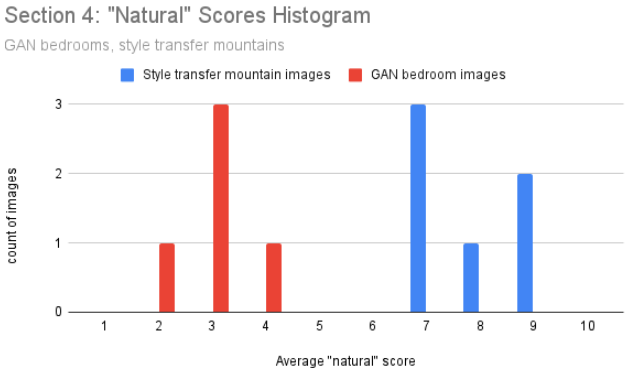
ure 8b] is a frequency histogram of the average "natural" score given to each image in the section by humans in the survey. The bins are left-closed, so the '4' bin contains the continuous numbers [4,5). The distribution once again shows style transfer images significantly and consistently outperforming GAN images as the two distributions have no overlap. The style transfer mountain images cover the bin range [7,9] and the GAN bedroom images cover the bin range [2,4], showing that style transfer mountain images were considered very natural while GAN bedroom images were considered very unnatural.

Looking further into how comparing style transfer images with GAN images affected human perception of "natural", we computed the difference in the natural score and the probabilities given by humans for Section 1 (all style transfer) minus Section 3 (style transfer bedrooms and GAN mountains) and Section 4 (GAN bedrooms and style transfer mountains) as will as Section 2 (all GAN images) minus Section 3 and Section 4. Figure 9a computes these delta values for style transfer images, and shows that style transfer images were consistently ranked as more "natural" when given in a survey with GAN images than only with other style transfer images. The probability assigned to the correct classification has no significant change when given in a survey of all style transfer images or a combination of style transfer and GAN images.

Figure 9b computes the same delta values for GAN images, and shows that GAN images are consistently ranked



(a) Section 4: Human vs. Inception P(correct class) scatter plot



(b) Section 4: "Natural" Scores Histogram

Figure 8: Section 4 Graphs

Image Title	delta natural score (style only - style vs gan)	delta Human P(correct) (style only - style vs gan)
starry_b5	-0.6422142214	0.003150315032
starry_b12	-0.7898289829	-0.01892439244
starry_m4	-1.912922061	-0.007895404925
starry_m5	-1.812388931	-0.01135897436
starry_b8	-1.027452745	0.009855985599
starry_b13	-0.4441944194	-0.01379387939
starry_m1	-2.514851485	-0.02642802742
starry_m9	-1.502157908	-0.01431835491
starry_b1	-1.095184518	-0.006750675068
starry_b4	-0.5945094509	-0.01635913591
starry_m10	-1.369890835	0.1255902513
starry_m14	-1.952018279	-0.02223914699

(a) Delta: Style transfer alone - style transfer and GAN

Image Title	delta natural score (GAN only - GAN vs style)	delta Human P(correct) (GAN only - GAN vs style)
gan_bedroom_50	0.3680241327	-0.006334841629
gan_bedroom_62	0.5113122172	-0.0161387632
gan_mountain_11	0.7535650624	-0.008734402852
gan_mountain_41	1.23885918	-0.003609625668
gan_bedroom_4	0.5248868778	0.01357466063
gan_bedroom_25	-0.1764705882	0.02066365008
gan_mountain_21	0.9950980392	-0.01822638146
gan_mountain_22	1.043672014	0.003520499109
gan_bedroom_15	0.2941176471	0.005128205128
gan_bedroom_43	-0.5867269985	-0.0006033182504
gan_mountain_37	0.8355614973	-0.01581996435
gan_mountain_61	0.9193404635	-0.007531194296

(b) Delta: GAN alone - GAN and style transfer

Figure 9: Delta Tables

as more natural when given in a survey of only GAN images rather than a survey of both GAN and style transfer images. This is consistent with the findings in Figure 9a. Once again, the probability assigned to the correct classification has no significant change when given in a survey of all GAN images or a combination of style transfer and GAN images.

4.4. Indoor vs Outdoor Scene Recognition

In Section 1, Figure 6a shows humans generally outperforming the Inception model for both indoor and outdoor style transfer images. Figure 6b shows humans ranking indoor and outdoor style transfer images as roughly the same amount of "natural".

In Section 2, Figure 3a shows humans outperforming the Inception model for indoor GAN images, but the Inception model outperforming humans for outdoor GAN images. However, humans ranked outdoor GAN images as more natural than indoor GAN images in Figure 3b. So, in this case, we see that how much an image looks like something is not correlated to how "natural" it looks to someone in the GAN only case.

In Section 3, Figure 7a shows humans outperforming the Inception model on all indoor style transfer images, but the Inception model outperforming on almost all outdoor GAN images. In Figure 7b, humans ranked the style transfer in-

door images as more natural, so in this case we see that the images that looked more like an object were consistently seen as more "natural".

In Section 4, Figure 8a shows humans outperforming the Inception model on two-thirds of the outdoor style transfer images, and performing the same as or better than the Inception model on two-thirds of the indoor GAN images. In Figure 8b, humans ranked the style transfer outdoor images as more natural than the indoor GAN images, so we also see the case here where images that looked more like an object were consistently seen as more "natural.

As the delta tables in Figure 9 showed, the comparison between style transfer and GAN images has the most impact on whether something is considered "natural" rather than an indoor or outdoor scene. Only in Section 2 when we compared GAN indoor and outdoor images did we see a consistent result that outdoor GAN images appeared more "natural" than indoor GAN images.

5. Conclusions

In the case of probability performance, we found that humans generally performed better at giving a higher probability to the correct class than the Inception model for style transfer images, while the Inception model generally outperformed humans for GAN images.

In measuring human surprisal through our "natural" score, we found that when comparing only GAN images, outdoor scenes were considered more natural. However, when comparing GAN and style transfer images, style transfer images were always considered more natural than GAN images regardless of if the scene was indoors or outdoors.

Since we gave surveys with sets of images with varying Inception scores, we wanted to see if a correlation existed between sets with high Inception scores and high "natural" scores, but found that the biggest indicator of "natural" was when style transfer images were compared with GAN images.

We also found that for the cases in comparing style transfer with GAN images, the human probability assigned to an image is a good indicator of its "natural" ranking, but not in the case of only GAN images.

6. Future Work

One idea for future work would be to train the GAN and Inception models on more classes to get more variety of indoor and outdoor scenes to better look into that correlation. We could also apply more styles to images to see if style transfer images are always ranked as more "natural" than GAN images regardless of the style applied. Similarly, we could change our conv5 style transfer images to conv1 to see if images with less of the image content preserved and more of the artistic style integrated still outperform GAN images in the "natural" rankings.

We could also apply the same style transfer to GAN images and compare those to the original GAN images to see if style transfer GAN images would be ranked more or less "natural" than the original GAN images. Finally, after fine tuning the GAN more, we could compare the GAN and the style transfer images to real images from the training set to see if either generative technique could be considered more "natural" than a real image.

7. Contribution Statement

In this project my contributions consisted of helping to find the dataset, creating and training the DCGAN models, retraining the inception model, helping to design the human study, and helping to write the paper.

References

[1] Shaojing Fan, Rangding Wang, Tian-Tsong Ng, Cheston Y.C. Tan, Jonathan S. Herberg, and Bryan L. Koenig. Human perception of visual realism for photo and computer-generated face images. 2

[2] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXic preprint arXiv: 1508.06576*, 2015. 1, 2, 3

[3] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 1

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[5] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016. 1, 2

[6] Anthony Chad Sampanes, Philip Tseng, and Bruce Bridgeman. The role of gist in scene recognition. 2

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 1, 2

[8] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. 2017. 1, 2

Appendix

A. Survey Sections

	1: all style transfer	2: all GAN	3: style transfer bedroom, GAN mountain	4: GAN bedroom, style transfer mountain
High Inception Score	starry_b5, 4c starry_b12, 4e starry_m4, 5b starry_m5, 5c	gan_bedroom_50, 1e gan_bedroom_62, 1f gan_mountain_11, 2a gan_mountain_41, 2e	starry_b5, 4c starry_b12, 4e gan_mountain_11, 2a gan_mountain_41, 2e	gan_bedroom_50, 1e gan_bedroom_62, 1f starry_m4, 5b starry_m5, 5c
Medium Inception Score	starry_b8, 4d starry_b13, 4f starry_m1, 5a starry_m9, 5d	gan_bedroom_4, 1a gan_bedroom_25, 1c gan_mountain_21, 2b gan_mountain_22, 2c	starry_b8, 4d starry_b13, 4f gan_mountain_21, 2b gan_mountain_22, 2c	gan_bedroom_4, 1a gan_bedroom_25, 1c starry_m1, 5a starry_m9, 5d
Low Inception Score	starry_b1, 4a starry_b4, 4b starry_m10, 5e starry_m14, 5f	gan_bedroom_15, 1b gan_bedroom_43, 1d gan_mountain_37, 2d gan_mountain_61, 2f	starry_b1, 4a starry_b4, 4b gan_mountain_37, 2d gan_mountain_61, 2f	gan_bedroom_15, 1b gan_bedroom_43, 1d starry_m10, 5e starry_m14, 5f

B. Exact Survey Results

Section 1: all style			
Image Title	natural score	Human P(correct)	Inception P(correct)
starry_b5	7.03960396	0.9940594059	0.9934
starry_b12	6.346534653	0.9742574257	0.9616
starry_m4	5.574257426	0.8485148515	0.9688
starry_m5	6.059405941	0.863	0.9788
starry_b8	7.108910891	0.9871287129	0.8551
starry_b13	8.237623762	0.9702970297	0.8573
starry_m1	5.485148515	0.8504950495	0.7827
starry_m9	7.702970297	0.8702970297	0.8148
starry_b1	7.336633663	0.9841584158	0.491
starry_b4	7.178217822	0.9722772277	0.5652
starry_m10	7.732673267	0.8871287129	0.5064
starry_m14	6.663366337	0.8623762376	0.4122

Figure 10: Section 1 Survey Data

Section 2: all GAN			
Image Title	natural score	Human P(correct)	Inception P(correct)
gan_bedroom_50	3.470588235	0.837254902	0.9719
gan_bedroom_62	3.588235294	0.8941176471	0.9848
gan_mountain_11	7.549019608	0.8549019608	0.9964
gan_mountain_41	5.784313725	0.7941176471	0.9961
gan_bedroom_4	4.294117647	0.8725490196	0.86
gan_bedroom_25	4.490196078	0.8745098039	0.849
gan_mountain_21	7.745098039	0.8431372549	0.8596
gan_mountain_22	4.725490196	0.6921568627	0.8695
gan_bedroom_15	3.294117647	0.8333333333	0.5188
gan_bedroom_43	3.823529412	0.868627451	0.4946
gan_mountain_37	7.176470588	0.8137254902	0.752
gan_mountain_61	5.078431373	0.6901960784	0.7587

Figure 11: Section 2 Survey Data

Section 3: style bedroom, gan mountain			
Image Title	natural score	Human P(correct)	Inception P(correct)
starry_b5	7.681818182	0.9909090909	0.9934
starry_b12	7.136363636	0.9931818182	0.9616
gan_mountain_11	6.795454545	0.8636363636	0.9964
gan_mountain_41	4.545454545	0.7977272727	0.9961
starry_b8	8.136363636	0.9772727273	0.8551
starry_b13	8.681818182	0.9840909091	0.8573
gan_mountain_21	6.75	0.8613636364	0.8596
gan_mountain_22	3.681818182	0.6886363636	0.8695
starry_b1	8.431818182	0.9909090909	0.491
starry_b4	7.772727273	0.9886363636	0.5652
gan_mountain_37	6.340909091	0.8295454545	0.752
gan_mountain_61	4.159090909	0.6977272727	0.7587

Figure 12: Section 3 Survey Data

Section 4: GAN bed, style mountain			
Image Title	natural score	Human P(correct)	Inception P(correct)
gan_bedroom_50	3.102564103	0.8435897436	0.9719
gan_bedroom_62	3.076923077	0.9102564103	0.9848
starry_m4	7.487179487	0.8564102564	0.9688
starry_m5	7.871794872	0.8743589744	0.9788
gan_bedroom_4	3.769230769	0.858974359	0.86
gan_bedroom_25	4.666666667	0.8538461538	0.849
starry_m1	8	0.8769230769	0.7827
starry_m9	9.205128205	0.8846153846	0.8148
gan_bedroom_15	3	0.8282051282	0.5188
gan_bedroom_43	4.41025641	0.8692307692	0.4946
starry_m10	9.102564103	0.7615384615	0.5064
starry_m14	8.615384615	0.8846153846	0.4122

Figure 13: Section 4 Survey Data