Department of Computer Science

Submitted in part fulfilment for the degree of
MSc in Software Engineering.

Predicting Memorable Regions Of Images

Babar Khan

2023-April-23

Supervisor: Adrian Bors

Contents

Contents

1 Abstract

I investigate a range of deep conditional image to image generative models and their success in learning mappings of images to memorable regions. These models could hopefully be used by marketers, educators, or others to determine which parts of their images are memorable, and thus help to more easily create memorable images. This was done by conducting supervised training using image, VMS pairs in the VISCHEMA dataset. I compare these models using the L1 loss of their predicted memorability maps across the validation portion of the VISCHEMA dataset.

I have found that autoencoders, even though they share similar network structures to GANs and Diffusion models [1], are unable to learn a very accurate mapping, they tend to smooth out the blocky VMS components into circular shapes that fade towards the edges. My best autoencoder model took the least amount of time to train of the three, and was able to achieve a mean L1 loss of ~ 0.132 across the validation distribution. GAN models can be difficult to work with, lots of research has gone into creating versitile and robust GAN architectures, picking one of those and adapting it to your needs can reduce development time considerably. Using a modified pix2pix GAN model [2] I was able to generate VMS maps far more accurately than with an autoencoder model, they maintain the blocky shape of the regions and achieve a lower mean L1 loss of ~ 0.0740. This model took significantly longer to train than the autoencoder. Diffusion models take longer to train than GAN and autoencoder models, and while they can produce high quality images [3], [4], the training loop requires a much larger amount of computational resources than an autoencoder or GAN does, which makes it infeesible for most people to train diffusion models for their own image distributions. My best model was able to achieve an L1 loss of (Final Diffusion Loss). Unfortunately I was unable to complete a full hyperparameter search due to computational limits. Diffusion models lend themselves well to multiple GPU architectures which unfortunately I didn't have access to.

Current image to image models are able to, with great accuracy, predict which regions of images are memorable, and could be used in the process of creating images that are memorable. While diffusion models have seen lots of mainstream attention recently, until home computing power is significantly greater, it is infeesible for most to train their own from home in

any reasonable amount of time.

2 Executive Summary

I found a deep interest working on this project on as I'd never created an image to image model, considered the memorability of images, or even read a research paper before starting this project. Luckily my project supervisor, Adrian Bors, suggested a range of great research papers on image memorability, and that allowed me to start with a good footing.

My dissertation is written using LaTeX, and bar a few small hiccups, it was very easy to learn to use as there are lots of online resources available. I typically write notes in markdown, which are very similar to TeX files making the transition easier than going from a traditiomal WYSIWYG editor.

I find it difficult to work at home, and prefer to work in the University Library or any of the other various study spaces available. Unfortunately, my laptop is not computationally powerful enough to train most deep networks, my workflow became a cycle of writing code at the library and running it on my computer overnight. This work structure did take a while to get used to but it didn't pose much of a hassle in the end. I could have theoretically set up a jupyter remote server to run on my computer at home, and connected to that from my laptop, but I was concerned about vulnerabilities that could arrise.

There is a lot of openly available research into machine learning and computer vision, AlexNet sparked somewhat of a machine learning renaissance in 2012, and in the decade since we have seen countless incredible papers released. The industry as a whole is incredibly open, not only do academics produce research but large companies with hugh funding typically post their research online for anyone to read. This abundance of research makes getting up to speed on the state of the art very easy. Not only is there a lot of free research available, but there are also countless articles and videos produced with the aim of making learning concepts easier. I've taken 3 courses at the University of York on inteligent systems, and in one of them I learned to implement my own unconditional GANs, because of that the transition to creating my own conditional image to image networks wasn't that hard.

Unfortunately diffusion models take a long time to train and even longer to generate images for, to evaluate one of my diffusion models with 2000

noise steps over the entire validation dataset would take roughly 4 hours, compared to the seconds for an autoencoder or GAN. This

3 Introduction

3.1 Background

We are constantly surrounded by imagery, on the way to work, on the internet, on TV, in stores, etc. Some images stick out more than others, we see hundreds, if not thousands, of images a day, and yet culturally and individually we all remember similar ones.

Due to cultural significance an image can become memorable. The 2015 dress [5], a country's flag, or the 1932 image "lunch atop a skyscraper" [6], come to mind. I argue that the memorability of such images is tied to the culture surrounding them, not necessarily due to intrinsic properties within. The focus of this research is on the intrinsic properties within an image that make it memorable to an individual, like seeing an advert on a bus and then later recognising the same advert online.

It has been shown that the memorability of an image is an intrinsic property, independent from the viewer [7]–[10]. This was achieved by performing a memorability game where participants are presented with a stream of images, and on an interval shown an image that they had already seen. Most participants were able, or unable, to remember the same images. How often an image was recognised is proportional to its memorability. This is taken further by Akagunduz et al. [11] where, in a similar experiment, they measure which regions of images are memorable.

3.2 Motivation

If we can accurately predict which regions of images are memorable then the process of creating memorable images becomes much easier. This would be useful to people in many fields, marketing and education being the most prominent. Businesses want to create memorable images so that when you see their products you have a sense of familiarity, that familiarity causes customers to be more likely to purchase your products. In education, creating memorable images could help with creating memorable diagrams,

tools, or resources that make learning easier for students.

3.3 Structure

The next chapter is my literature review, I cover research into memorable images and do a brief overview of image to image models.

The following chapter is the methodology, I describe three experiments where I test the ability of different models to predict the memorable regions of images

After that is the experimental results, where I discuss in detail my findings

Then the conclusion, where I discuss the implications of my findings

4 Literature Review

4.1 Memorability

In R. Breners work [12] subject's abilities to remember a series of units was tested, a unit was different depending on the test,

> In the digit test, for example, each digit was a unit; in the sentence test each sentence was a unit, etc. ... Each nonsense syllable constituted a unit ... Each consonant constituted a unit ... Each [geometric] design constituded a unit. [12, p.468]

It was found that people are not very good at remembering nonsense sylables or sentences but are much better at remembering digits, consonants, and colours. Remembering geometric designs is placed somewhere in the middle. This is of interest because it's very similar to our investigation into what aspects make images memorable. Brener was interested in consonants, colours, geometric designs, etc but the finding that different units can be significantly harder or easier to remember is useful to us. If we swap out units for image properties such as texture, contrast, saturation etc, or even more abstract features that a CNN may recognise, then it should be possible to find how memorabilty is impacted by these.

The mean score of 5.31 found for geometric designs seems to indicate that people are not great at recalling images, however in R. Nickerson's work [13] we can see the opposite, subjects are found to be extremely good at recognising images. Nickerson found that subjects shown a series of images are, with great accuracy, able to recall if the photo is one they have seen before or not. An item is referred to as 'new on its first occurrence and old on its second occurrence'[13, p.156], it was found that subjects shown one image at a time, each with equal chance of being old or new, are able to correctly distinguish them with 95% accuracy.

Like in [13], R. Shephards work [14] also found that subjects are able to distinguish with great accuracy new and old images, the mean percentage of images correctly identified was 99.7% after a delay 2 hours, 92.0% after a delay of 3 days. 87.0% after a delay of 7 days, and 57.7% after a delay of 120 days. The introduction of a delay into Nickerson's experiment allows

us to see that regardless of whether the image is in our short-term memory (2 hour delay) or our long-term memory (3-120 days) subjects are able to correctly identify an incredible amount of images.

L. Standing built on the work of Nickerson, Shepard, and Brener in [15]. Four experiments were ran which tested memory capacity and retreival speed for pictures and words, I am interested in the performance related to pictures specifically. He found that the capacity for image recognition from memory is almost limitless, when measured under testing conditions. In Standing's first experiment he tested 'normal' and 'vivid' images. He describes 'normal' images as those that 'may be characterized as resembling a highly variegated collection of competent snapshots'[15, p.208], and 'vivid' images are described as 'striking pictures ... with definitely interesting subject matter'[15, p.208]. Much like how in Brener's work [12] the memorability of an object was variable based on what classification of unit was tested, this work shows that on an even more specific level, the classification of the image into normal or vivid, has an impact on subjects memorabilty, Standing found that the vivid images were more likely to be remembered by subjects.

Lots of work has been done to further build on that of Standing, Nickerson, Brener et al. T. Brady et al. [16] performed research into how much information is retained in long term memory and found that subjects are able to remember not just the gist of an image, but also fine grain information such as the state of objects within an image, and that they are able to distinguish between variants of objects shown in images. An example shown is an abacus in two different states, 13/14 subjects were able to distinguish which one they had seen before.

T. Konkle et al. build further on the work in [15], [16] by studying the impact that categorical distinctness has on memorability. This was tested by creating a dataset of images where composed of categories such as tables, cameras, and bread etc. Each category had between 1 and 16 images, a memory test like those performed in [15], [16] was performed and the percentage correctly identified was found to decrease as the number of images within a category increased. From this we can see that categorically distinct images are more likely to be remembered.

Studies by Isola et al. [7], [8] were performed with the goal of identifying a collection of visual attributes that make images memorable, and to use those to predict the memorability of an image. It was found that properties such as mean saturation, and the log number of objects, has less impact on the memorability score than object statistics. Categories such as: person, person sitting, and floor, were most helpful image memorability. The categories: ceilings, buildings, and mountains, were least helpful. Their approach was limited by the fact that the object statistics were annotated by

hand, this would both make automating the process of determining image memorability impossible, and limit them from finding any abstract properties that helped/hindered memorability.

Khosla et al. [17] built on the work in [7], [8] by, instead of determining memorability of an entire image, creating a model that discovers memorability maps of individual images without human annotation. These memorability maps are able to distinguish which areas of an image are remembered, forgotten, or hallucinated. Their approach, similar to Isola et al. in [7], [8], is limited by the arbitrarily picked list of features that define memorability.

4.2 Using Deep Learning to Predict Memorability

4.2.1 Memorability Datasets

Deep learning models are composed of a large number of simple functions, these functions have adjustable weights that can be tuned to produce a more optimal output. A loss function is used to calculate how correct the output of the model is, and then each weight, W, is updated by calculating $\frac{\partial L}{\partial W}$ and performing gradient descent.

Deep learning models can be trained through supervised or self-supervised learning. In supervised learning the model is trained across a pair of input and output pairs, this requires a pre-annotated dataset to train on. This process can be costly and it's important to train on a dataset without any biases, as these can appear in your final models [18], [19].

There are two datasets of interest. LaMem [9], created by Khosla et al., and VISCHEMA [11] created by Akagunduz et al.

LaMem contains 60,000 annotated image and memorability pairs. Memorability is measured as a continuous value in the range [0,1] where a higher value indicates that the image is more memorable. They then train a CNN to predict the memorability of an image, resize their images, and pass overlapping regions of the image into their CNN. The generated memorability maps are tested for accuracy by adjusting the original images to de-emphasise the regions indicated as memorable, these adjusted images are then tested using the original memorability game used to create the dataset. They find that their memorability maps do reliably identify memorable regions.

VISCHEMA is a significantly smaller dataset, containing around 1600

images, memorability map pairs in the VISCHEMA Plus version. Their experiment differs from that of Khosla et al., they ask participants to memorize a set of images, and then during a test phase rate how well they remember each images and to select the regions that made them remember it. In order to predict memorability the network output utilises fully connected layers which I believe is unnecessary, and may even hurt performance. More modern computer vision network structures, such as those in [1], [2], [20]–[22], use fully convolutional networks. These maintain spatial locality which allows them to generalise better.

4.2.2 Image Synthesis Models

Autoencoders

The autoencoder model is composed of 3 parts, an encoder, a bottleneck, and a decoder, these are typically used for compression. The U-Net [1] is a variation of an autoencoder which has also found great success in image segmentation within the medical field. It makes use of residual connections between corresponding layers in the encoder and decoder blocks, these allow the model to preserve data for use in the decoding stage. This architecture is the backbone of many image to image models [2], [22], [23], and has seen many extensions [24], [25]

Generative Adversarial Networks

GAN architectures[20] use competitive co-evolutionary algorithms where a Generator and a Discriminator compete. The Generator is typically given a latent space vector and uses that to produce an image, the Discriminator has to determine if that image belongs to a given distribution. These networks have seen great success [2], [26], [27] but are typically unstable and require a lot of parameter fine tuning. They can also suffer from: non-convergence, mode collapse, and diminished gradients [28].

Diffusion Models

Diffusion based generative models work by learning to iteratively remove Gaussian noise from a sample T times until it produces an image from the training domain. The model was first proposed by Sohl-Dickstein et al. [29], and has been further developed by Ho et al. [21], Dhariwal and Nichol [23], and Saharia et al. [22].

Diffusion models, first introduced by J. Sohl-Dickstein et al. [29], are a machine learning model that work through a forward diffusion process systematically destroying the structure in a data distribution, and the learning of a backwards process to restore the structure. The method uses a Markov chain to convert x_t into x_{t-1}. Starting with x_T, a sample of Gaussian noise, a generative Markov chain converts this into x_0, which is a sample from the target data distribution. Because the model only estimates small pertubations of noise, x_t given x_{t-1} , rather than an entire transformation from x_0 to x_T, it is tractable to train.

The DDPM, a UNet based diffusion model,[21] is capable of producing high quality images and achieves state of the art FID scores across the CIFAR10 dataset. Dhariwal and Nichol [23] show tweaks that allow for diffusion models to achieve state of the art FID scores across the ImageNet dataset and when used in combination with upsampling diffusion further improve FID scores. They do this by using improvements proposed in [27], [30]–[33]. These improvements also reduce the number of noise steps required from thousands to (in some cases) 50. Through the decrease in noise steps they are able to reduce the amount of time that it takes to generate an image. Chen discusses in [34] how changing the resolution of an image has an impact on the noise scheduling required, he finds that the optimal scheduler at a smaller resolution may cause under training for higher resolution images. Multiple strategies are proposed to adjust noise scheduling. Firstly, changing the noise schedule functions to those based on cosine or sigmoid, with temparature scaling. Secondly, reducing the input scaling factor from 1 increases the noise levels which destroys more information at the same noise level. They then combine these into a compound noise scheduling strategy.

5 Methodology

5.1 Requirements Capture

The two dataset options I have considered are LaMem [9] and VISCHEMA [11]. I have chosen to use the latter to train my image to image network because Akagunduz et al. successfully measure the memorable regions of images, rather than a score per image, while Khosla et al. are able to convert these scores into memorability maps, that is inherently less robust than the experiment presented in [11].

The VISCHEMA Plus dataset contains 1600 image to VMS mappings, 1280 of these compose the training portion and the remaining 320 the validation portion. I aim to use a deep convolutional image to image model to learn a mapping of these images to their corresponding VMS labels. Our model should learn a general understanding of the mapping such that when it is provided with an image that matches our distribution, it can accurately create a VMS label for it. Our model will need to learn to create accurate mappings and we can test that through a loss function, such as L1, and through qualitative analysis.

5.2 Motivation

As autoencoders, GANs, and diffusion all produce images in different ways I think it would be interesting to compare how the 3 of them perform when asked the same task. I will run three experiments where I train a UNet autoencoder, a GAN, and a diffusion model to produce a VMS label given an image from the dataset. I hope to be able to find the strengths and weaknesses of each of these systems in this application. For each model I will experiment with network parameters and training hyperparameters to fine tune models that produce the best output.

In each experiment I will automate a system that: tests the effectiveness of different numbers of layers and channels in each network, varies the normalisation method, optimisation function, learning rate, and batch size to find

the best training environment and model parameters. It is not be feesible to test every combination of these variables, thus I will employ a training strategy to test every variation of a single parameter/hyperparameter while keeping the rest constant.

I will start each model with an estimation of good parameters, then iterate through each individual parameter, varying only that one and measuring the impact on performance. After testing all variations of a single parameter, I will pick the one that gave the highest score and move onto the next parameter. This will bring the size our search space down by an order of magnitude, however we won't be exploring the entire search space and will potentially miss out on good options.

This training strategy will be especially necessary in experiment 2 as training within a GAN is typically unstable and good parameter choices are necessary. In experiment 3 I will also vary the noise scheduler and the beta values.

5.3 Experiment 1

Model: This network is a UNet autoencoder that takes as input a 64x64x3 tensor of floating point values in the range [-1,1], these are the images in our dataset. It outputs a 64x64x3 tensor of floating point values in the range [-1,1], these are our label estimates. This model can be described as:

$$L_{pred} = model(I)$$

Model Loss: This is simply the L1 loss between L_{pred} and L_{real}, describing how closely the output of our network matches the corresponding label. At each epoch we will calculate this over the training dataset, use that for backpropagation, and then calculate it across the validaton dataset to test how well we have generalised.

$$L = L1(model(I), L_{real})$$

Training Strategy: Please see algorithm 1 for the training loop.

5.4 Experiment 2

I will train a conditional generative adversarial network to generate images of VMS maps given images from the VISCHEMA dataset. I will use

Algorithm 1 UNet Autoencoder Training Strategy

for every epoch **do**
  for I, L_{real} in training dataloader **do**

    $L_{pred} = model(I)$
    $loss = L1(L_{pred}, L_{real})$
    Update weights of model with backpropagation

  end for
end for

an adapted Pix2Pix network [2], making tweaks that I think will increase performance. Pix2Pix by Isola et al.[2], is an image-to-image conditional generative adversarial model based on the UNet [1]. Designed to translate images from one style into another. In [35] M. Arjovsky et al. introduce a GAN variant based on the Wasserstein distance between the output distribution and the image distribution, the benefit of this is that the Wasserstein distance is continuous and differentiable almost everywhere, reducing the risk of dimminishing gradients. In [36] N. Makow investigated the use of Wasserstein Distance in the Pix2Pix model but unfortunately found that it does not perform much better, I would still like to experiment with it as VISCHEMA plus is different from any dataset used in [2] and as Makow states they were unable to perform a complete hyperparameter search, meaning that its possible we could achieve greater results than vanilla Pix2Pix. I will be adapting the following Pytorch implementation [37].

Lots of work has gone into making GAN models more stable and I will use these findings in my own models. In [33] A. Brock et al. found that when 'increasing the batch size by a factor of 8 ... models reach better final performance in fewer iterations, but become unstable and undergo complete training collapse'. Because of this I will experiment with early stopping and low batch sizes. In [38] Z. Shengyu et al. show how using differentiable augmentation on your images increases the quality of the outputted images. With a dataset with as few as 100 images they are able to produce high quality outputs, this is useful to us because the VISCHEMA plus dataset only has 1280 training images, which typically wouldn't train very well on a GAN.

Generator: This network takes as input a 64x64x3 tensor of floating point values and outputs a 64x64x3 tensor of floating point values in the range [-1,1]. This model can be described as:

$$L_f = G(I)$$

Discriminator: This network takes as input a 64x64x6 tensor of floating

point values. Channels 1,2, and 3 store the image and channels 4,5, and 6 are its corresponding label, either real or generated. It outputs a 4x4x1 tensor of boolean values. Each value in this output represents a 16x16x6 region of the input.

The loss for each network is computed as described in [pix2pix paper], across an entire batch of images and then the weights are adjusted with backpropagation. The optimiser used is one of the hyperparameters that we will search for.

Generator Loss: The generator loss describes how well it can trick the discriminator, and how closely its output matches the real label for the given image.

$$L = MSE(D(L_f, I), 1) + L1(L_f, L_r)$$

Discriminator Loss: The discriminator loss describes how accurately it is able to predict, given a label and an image, if the label is real or not.

$$L = 0.5 \times (MSE(D(L_f, I), 0) + MSE(D(L_r, I), 1))$$

Because these two loss values are relative to the performance of eachother they can't be used to see if our generator has converged on a solution. Therefore it is necessary, at each epoch, to also compute the L1 loss of the fake labels and real labels across the training and the validation dataset, as these values are independent of the discriminator. This will inform allow us to see if the generator is overfitting, underfitting, or training well.

Training Strategy: Please see algorithm 2 for the training loop.

Algorithm 2 GAN Training Strategy

for `every epoch` **do**
  for I, L_r `in training dataloader` **do**

    $L_f = G(I)$

    $pred_fake = D(L_f, I)$
    $loss_G = MSE(pred_fake, 1) + L1(L_f, L_r)$
    `Update weights of G with backpropagation`

    $pred_real = D(L_r, I)$
    $loss_D = 0.5 \times (MSE(pred_fake, 0) + MSE(pred_real, 1))$
    `Update weights of D with backpropagation`

  end for
end for

5.5 Experiment 3

I will adapt a PyTorch implementation of Palette [39] and train it to predict labels from the VISCHEMA dataset. Palette is a versitile conditional image to image diffusion model proposed by Saharia et al. [22] that is able to uncrop, inpaint, colorize, and remove JPEG artifacts from images. Because of this versitility I think it could learn to predict VMS maps from images. The conditional image passed to the model will be the image from the dataset, the ground truth will be the label. This network takes as input a 128x128x6 tensor of floating point values, this is the concatenation of an image and the noise added to the ground truth after t steps. It outputs a 128x128x3 tensor of floating point values, this is the models estimate of the noise added at time $t-1$. This model can be described as

$$noise_{pred} = model(I, t)$$

Model Loss: The model loss describes how well it can estimate the noise added to an image between time steps $t-1$ and t, it is computed across an entire batch of images and then the model weights are adjusted with backpropagation.

$$L = MSE(model(I, t), noise_{real})$$

The loss calculated for backpropagation is relative to the performance in estimating noise added, not for the performance when calculating VMS labels. Because of this I will also have to calculate the L1 loss of the generated labels against the real labels. This will take a lot of computational time so I will do it at the end of training. I will be able to see if our model has converged by observing the backpropagation loss.

Training Strategy: Please see algorithm 3 for the training loop.

5.6 Results Analysis/Testing

After performing all 3 experiments I will compare the results across the validation dataset qualitatively and using the L1 loss. The L1 loss will tell me how statistically close the outputs are but through qualitative analysis I can see if the outputs have cheated. I can also compare the L1 loss across the training and validation sets to see if any models have generalised well, if the loss across the training dataset is much smaller/larger than the validation dataset then it will imply that the model has become overfit/underfit respectively. Ideally they should be similar.

Algorithm 3 Diffusion Model Training Strategy

for `every epoch` **do**
  for I, L `in training dataloader` **do**

    $t = random(0, noise_steps)$
    $noise_{real}, L_{noisy} = noise\_image(L, t)$ ▷ Apply t steps of noise and return noise added at step t

    $input = concatenation(I, L_{noisy})$
    $noise_{pred} = model(I, t)$
    $loss = MSE(noise_{pred}, Noise_{real})$
    `Update weights of model with backpropagation`
  end for
end for

6 Experimental Results

6.1 Experimental Environment

I have implement the experiments in Python using the PyTorch machine learning framework, however the methodology that I describe should produce the same results in any programming language or framework.

This model was trained on a computer using an RTX 3070 with 8GB of VRAM and an AMD Ryzen 3600 with 32GB of system RAM. Testing the hyperparameter options for experiment 1 took approximately X days and I trained the final model over the course of 40 minutes. Testing the hyperparameter options for experiment 2 took approximately 3 days and I trained my final model over the course of 5 hours. Testing the hyperparameter options for experiment 1 took approximately X days and I trained the final model over the course of Y hours

All experiments were performed with differentiable augmentation as described in [38]. I performed translation and cutout, each with a 90% chance.

6.2 Hyperparameter tuning

For experiments 1 and 2 I automated the process of exploring the parameter and hyperparameter search space. In my search I varied the following parameters: The normalisation layer used, the channel layouts used, the optimiser used, the learning rates for the optimiser, and, if the Adam optimiser was used, the beta values. Because training a diffusion model is far more computationally expensive than training an Autoencoder or a GAN I was unable to automate the process of tuning the hyperparameters in experiment 3. Training a GAN model on the VISCHEMA dataset would typically take around 100 minutes, but training a diffusion model would take around 20 hours due to the greater number of epochs required.

In experiment 2 I allowed for the generator and discriminator models to use different normalisation functions, optimisers, and learning rates.

Exploring this search space exhaustively is unfortunately not feesible, there are over 5000 different combinations possible, and if I tested each combination once for 100 epochs then it would take approximately 300 days to test per experiment. Instead I will have to explore a subset of this search space. For each experiment I estimated some good default parameter and hyperparameter values, by varying these values I can lower the scope of the search space to approximately 100 combinations, which took $\sim$3 days per experiment to test. Unfortunately this does mean that not every combination has been tested, however we should achieve a good approximation of the best parameters and hyperparameters.

In experiment 3 I was only able to explore a small range of possible hyperaparameters and as such my results do not emulate the recent success found in diffusion based image generation. However, with a greater number of computing resources it may be possible to do so.

6.3 Experiment 1

In this experiment I trained a UNet to predict the VMS maps of images, I trained the model using backpropagation of the L1 loss of the model output given an image, and the corresponding label.

Diagram of my model:

The best parameters and hyperparameters that I found for my model were the following:

After 80 training epochs I was able to achieve an L1 loss across my validation dataset of 0.132.

A sample of best output images:

A sample of the worst output images:

When performed without differentiable augmentation the results looked like this:

FID score of Z, L1 Loss of Y.

6.4 Experiment 2

I explored the following parameter and hyperparameter options:

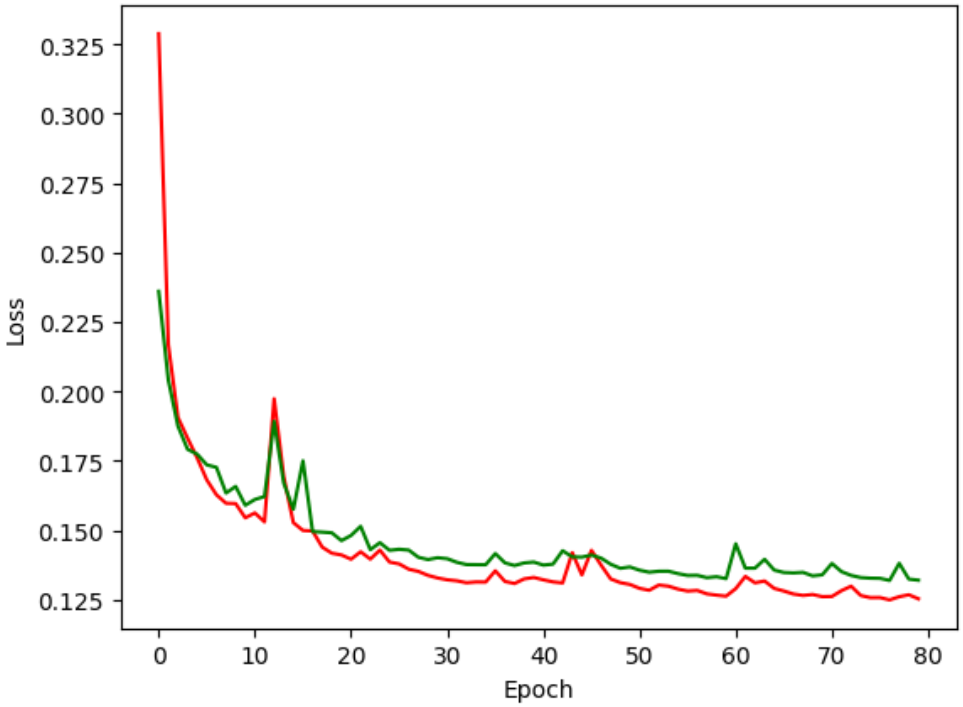

Figure 6.1: Autoencoder training graph. The red plot describes the training loss per epoch and the green plot the validation loss. The validation loss closely matches the training loss and the model does not become overfit.

Normalisation layers: Batch Normalisation, Instance Norm

Channel layouts: I iterated over generator encoder and decoder, and discriminator channel layouts of following form, with values of c from {32,64,50,100}:

- Encoder: (3, c, 2c, 4c, 8c, 16c),
- Decoder: (16c, 8c, 4c, 2c, c),
- Discriminator: (6, c, 2c, 4c, 8c, 16c).

Optimisers:

- SGD, using the following learning rates: 0.005, 0.01, 0.02,
- Adam, using the following learning rates: 0.0005, 0.001, 0.002, and using the following betas values: (0.9, 0.999), (0, 0.999), (0.5, 0.999),
- Adadelta, using the following learning rates: 0.5, 1, 2.

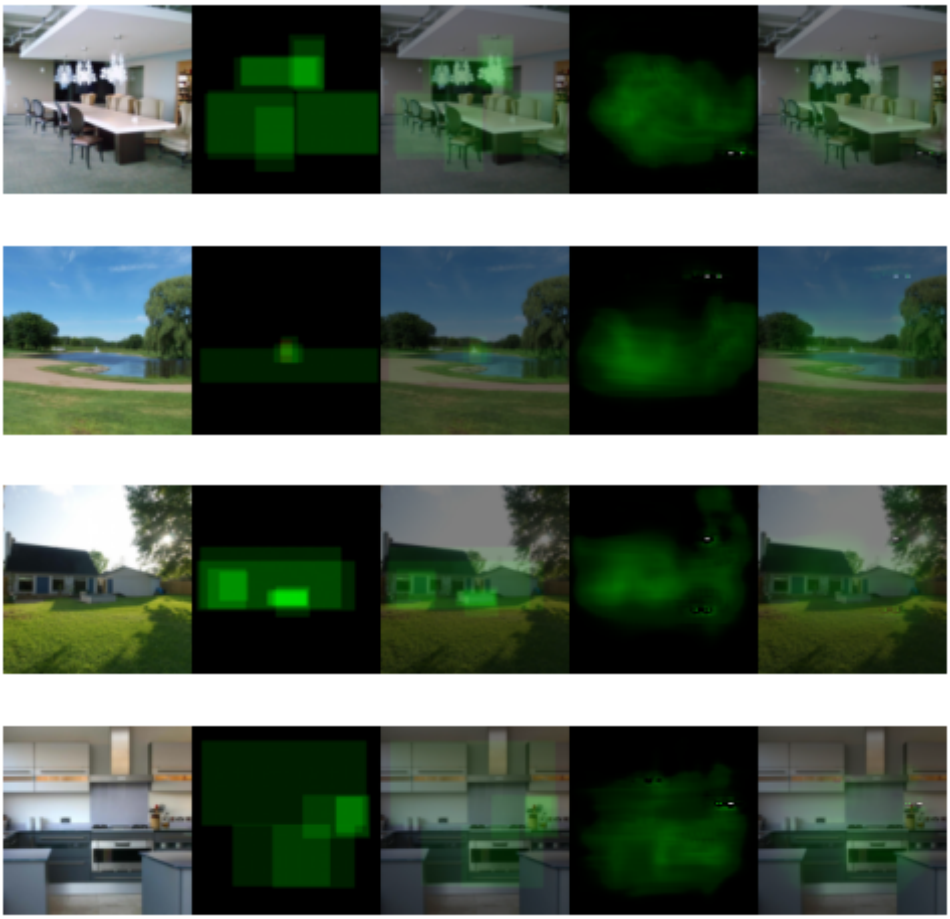

Figure 6.2: Best Autoencoder Outputs

Note: I didn't need to specify default values for the normalisation layer as these were the first variables I tested.

Default Generator Parameters and Hyperparameters:

- C = 64
- Optimiser: Adam, betas = (0.9, 0.999)
- Learning Rate: 0.001

Default Discriminator Parameters and Hyperparameters:

- C = 64
- Optimiser: Adam, betas = (0.9, 0.999)
- Learning Rate: 0.001

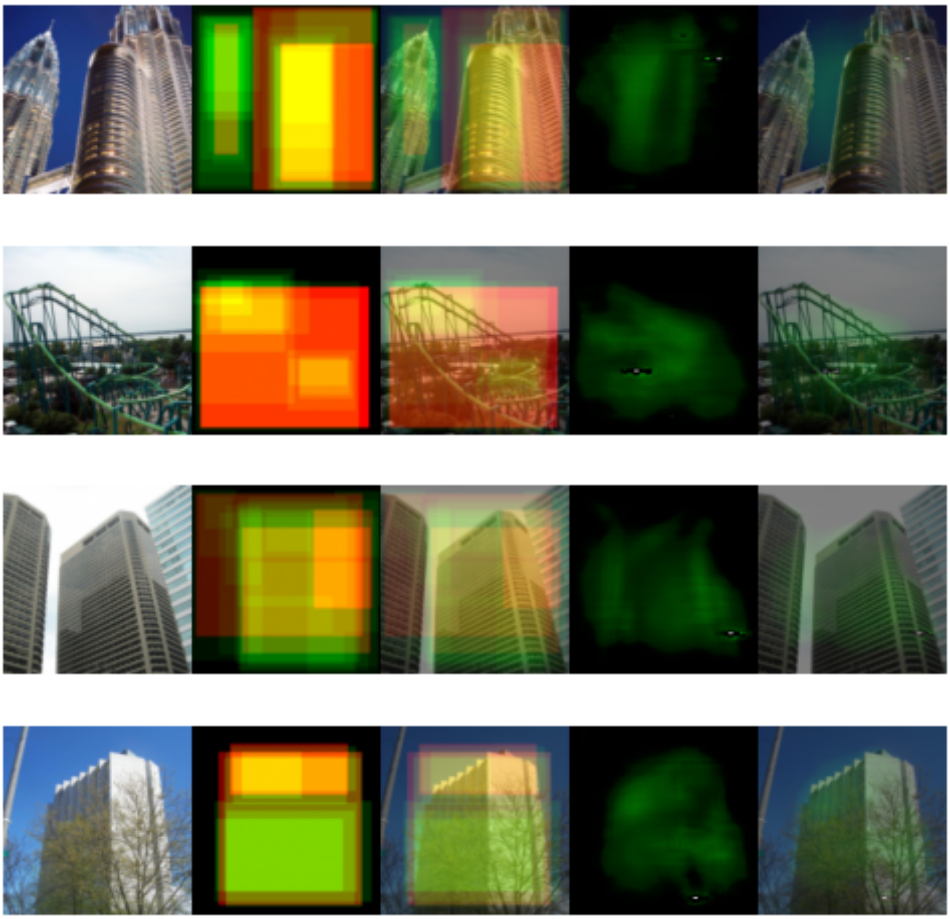

Figure 6.3: Worst Autoencoder Outputs

I iterated over each different parameter and hyperparameter and varied each one sequentially. I tested each combination for 100 epochs and used the L1 loss across the validation dataset as the score, if any new parameter options gave a better L1 score the it would become the default used going forward. This meant that I only had to iterate $\sim$100 combinations. I found the following best options.

Best Generator: Batch Normalisation, C = 32, SGD Optimiser with a learning rate of 0.01.

Best Discriminator: Batch Normalisation, c = 100, Adam optimiser with betas = (0.9, 0.999) and a learning rate of 0.001.

With this combination we achieved a loss of $\sim$1.02 across the validation dataset. I found other good results using similar combinations. Using the Adatelta optimiser for the generator and the Adam optimiser for the discriminator achieved a loss of $\sim$1.03 across the validation dataset. Using

the Adam optimiser for both the generator and discriminator achieved a loss of $\sim$1.04 across the validation dataset.

After 200 epochs of training our model is able to achieve an L1 score of 0.0740 across our validation dataset and Z across our training dataset. This is significantly better than our autoencoder model.

Figure 8.1 shows the outputs across the validation portion of our dataset with the lowest L1 Loss. They're accurately labelled with the correct regions as memorable. but frankly are quite boring, the low L1 loss across these is because the labels simply have lots of black data, the loss across these is 0 if the model also outputs black. Figure 8.2 shows what I think are far more impressive results, even if they have a higher loss than those present in figure 8.1. For these images the model is able to accurately predict large regions of the image that are memorable. The worst results are shown in 8.3, it appears that like our autoencoder model from experiment 1 that this model finds it difficult to predict regions that are detrimental to memorability.

After 200 epochs, L1 Score of 0.0740

Here are the images generated when I dont use differentiable augmentation

After X epochs L1 Score of Y

Wasserstein GAN: Using a Wasserstein image-to-image GAN as described in [36] I was able to achieve a best score of

FID score of Z, L1 Score of Y

Here are the images generated when I dont use differentiable augmentation

FID score of Z, L1 Score of Y

6.5 Experiment 3

6.6 Results, Evaluation Metrics, and Analysis

a graph of the train/val loss over time across the different experiments

a graph of the FID score at the end of the different experiments

a selection of the images with the different training methods

Some discussion about them.

7 Conclusion

Bibliography

[1] O. Ronneberger, P. Fischer and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015. arXiv: 1505.04597 [cs.CV].

[2] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, *Image-to-image translation with conditional adversarial networks*, 2018. arXiv: 1611.07004 [cs.CV].

[3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, *Hierarchical text-conditional image generation with clip latents*, 2022. arXiv: 2204.06125 [cs.CV].

[4] C. Saharia, W. Chan, S. Saxena *et al.*, *Photorealistic text-to-image diffusion models with deep language understanding*, 2022. arXiv: 2205.11487 [cs.CV].

[5] BBC. 'Optical illusion: Dress colour debate goes global.' (2015), [Online]. Available: https://www.bbc.com/news/uk-scotland-highlands-islands-31656935.

[6] M. Gambino. 'Lunch atop a skyscraper photograph: The story behind the famous shot.' (2012), [Online]. Available: https://www.smithsonianmag.com/history/lunch-atop-a-skyscraper-photograph-the-story-behind-the-famous-shot-43931148/.

[7] P. Isola, J. Xiao, A. Torralba and A. Oliva, 'What makes an image memorable?' In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 145–152.

[8] P. Isola, D. Parikh, A. Torralba and A. Oliva, 'Understanding the intrinsic memorability of images,' in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[9] A. Khosla, A. S. Raju, A. Torralba and A. Oliva, 'Understanding and predicting image memorability at a large scale,' in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2390–2398. DOI: 10.1109/ICCV.2015.275.

[10] P. Isola, J. Xiao, D. Parikh, A. Torralba and A. Oliva, 'What makes a photograph memorable?' *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 7, pp. 1469–1482, 2014.

[11] E. Akagunduz, A. G. Bors and K. K. Evans, 'Defining image memorability using the visual memory schema,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2165–2178, 2020. DOI: 10.1109/TPAMI.2019.2914392.

[12] R. Brener, 'An experimental investigation of memory span,' *Journal of Experimental Psychology*, vol. 26, no. 5, pp. 467–482, 1940. DOI: 10.1037/h0061096.

[13] R. S. Nickerson, 'Short-term memory for complex meaningful visual configurations: A demonstration of capacity.,' *Canadian journal of psychology*, vol. 19, pp. 155–60, Jun. 1965.

[14] R. N. Shepard, 'Recognition memory for words, sentences, and pictures,' *Journal of Verbal Learning and Verbal Behaviour*, vol. 6, pp. 156–163, 1967.

[15] L. Standing, 'Learning 10,000 pictures,' *Quarterly Journal of Experimental Psychology*, vol. 25, pp. 207–222, 1973.

[16] T. F. Brady, T. Konkle, G. A. Alvarez and A. Oliva, 'Visual long-term memory has a massive storage capacity for object details,' *Proceedings of the National Academy of Sciences*, vol. 105, no. 38, pp. 14 325–14 329, 2008.

[17] A. Khosla, J. Xiao, A. Torralba and A. Oliva, 'Memorability of image regions,' in *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, USA, Dec. 2012.

[18] J. Dastin, *Amazon scraps secret ai recruiting tool that showed bias against women*, Oct. 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[19] A. Najibi, *Racial discrimination in face recognition technology*, Oct. 2020. [Online]. Available: https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/.

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, *Generative adversarial networks*, 2014. arXiv: 1406.2661 [stat.ML].

[21] J. Ho, A. Jain and P. Abbeel, *Denoising diffusion probabilistic models*, 2020. arXiv: 2006.11239 [cs.LG].

[22] C. Saharia, W. Chan, H. Chang *et al.*, *Palette: Image-to-image diffusion models*, 2022. arXiv: 2111.05826 [cs.CV].

[23] P. Dhariwal and A. Nichol, *Diffusion models beat gans on image synthesis*, 2021. arXiv: 2105.05233 [cs.LG].

[24] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh and J. Liang, *Unet++: Redesigning skip connections to exploit multiscale features in image segmentation*, 2020. arXiv: 1912.05074 [eess.IV].

[25] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane and M. Jagersand, 'U2-net: Going deeper with nested u-structure for salient object detection,' *Pattern Recognition*, vol. 106, p. 107 404, Oct. 2020. DOI: 10.1016/j.patcog.2020.107404. [Online]. Available: https://doi.org/10.1016%2Fj.patcog.2020.107404.

[26] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, *Unpaired image-to-image translation using cycle-consistent adversarial networks*, 2020. arXiv: 1703.10593 [cs.CV].

[27] T. Karras, S. Laine and T. Aila, *A style-based generator architecture for generative adversarial networks*, 2019. arXiv: 1812.04948 [cs.NE].

[28] I. Goodfellow, *Nips 2016 tutorial: Generative adversarial networks*, 2017. arXiv: 1701.00160 [cs.LG].

[29] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan and S. Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics*, 2015. arXiv: 1503.03585 [cs.LG].

[30] J. Song, C. Meng and S. Ermon, *Denoising diffusion implicit models*, 2022. arXiv: 2010.02502 [cs.LG].

[31] A. Nichol and P. Dhariwal, *Improved denoising diffusion probabilistic models*, 2021. arXiv: 2102.09672 [cs.LG].

[32] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon and B. Poole, *Score-based generative modeling through stochastic differential equations*, 2021. arXiv: 2011.13456 [cs.LG].

[33] A. Brock, J. Donahue and K. Simonyan, *Large scale gan training for high fidelity natural image synthesis*, 2019. arXiv: 1809.11096 [cs.LG].

[34] T. Chen, *On the importance of noise scheduling for diffusion models*, 2023. arXiv: 2301.10972 [cs.CV].

[35] M. Arjovsky, S. Chintala and L. Bottou, *Wasserstein gan*, 2017. arXiv: 1701.07875 [stat.ML].

[36] N. Makow, *Wasserstein gans for image-to-image translation*, 2018.

[37] T. W. Jun-Yan Zhu Taesung Park, *Cyclegan and pix2pix in pytorch*, Commit 9f8f61e. [Online]. Available: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.

[38] S. Zhao, Z. Liu, J. Lin, J.-Y. Zhu and S. Han, *Differentiable augmentation for data-efficient gan training*, 2020. arXiv: 2006.10738 [cs.CV].

[39] L. Jiang, *Palette: Image-to-image diffusion models*, Commit 136b29f. [Online]. Available: https://github.com/Janspiry/Palette-Image-to-Image-Diffusion-Models.

8 Appendix

Best results:

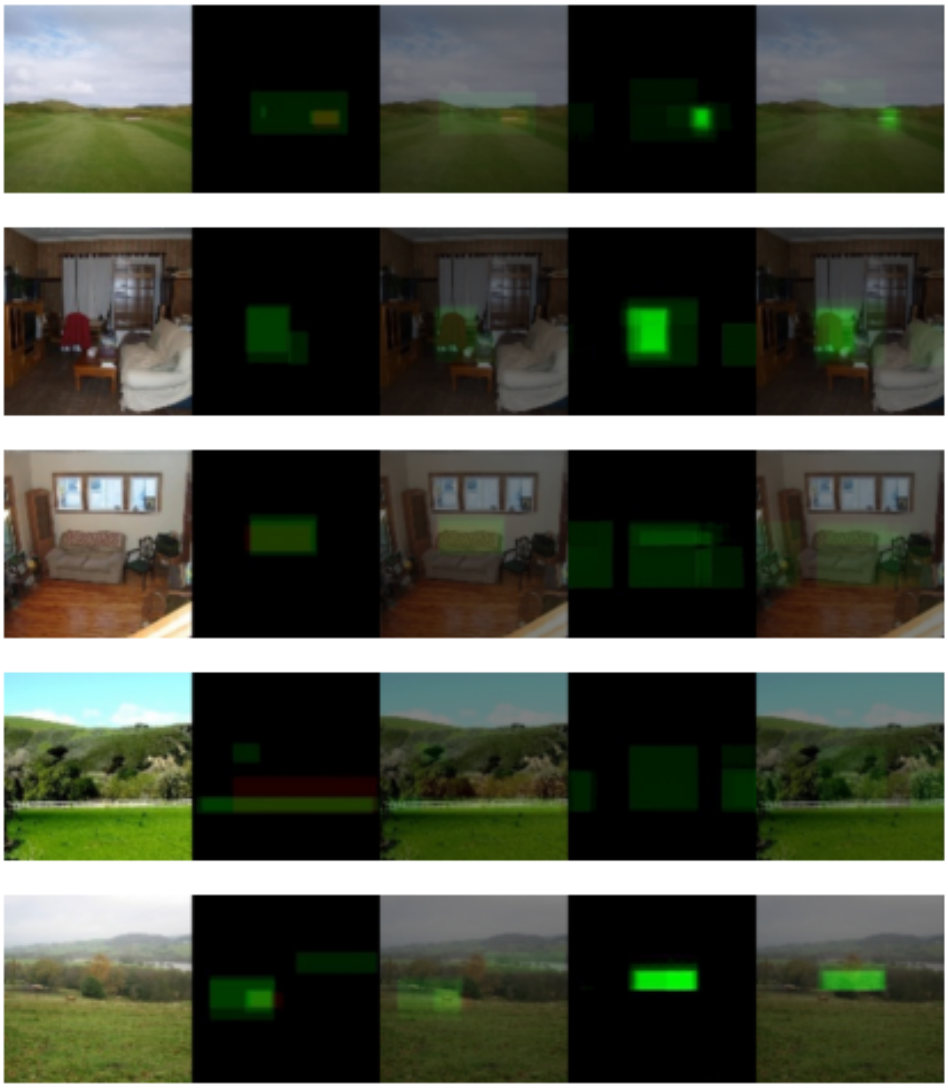

Figure 8.1: Best GAN Outputs

Favourite Results:

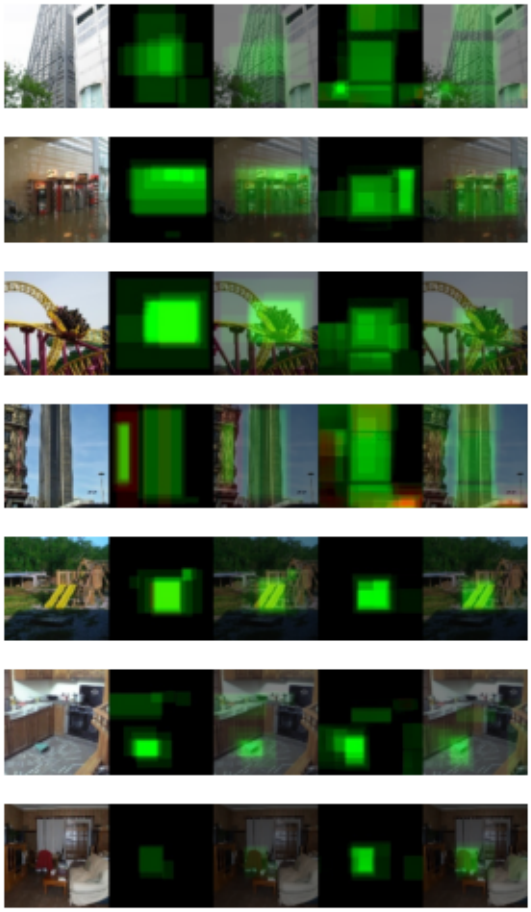

Figure 8.2: Favourite GAN Outputs

Worst results:

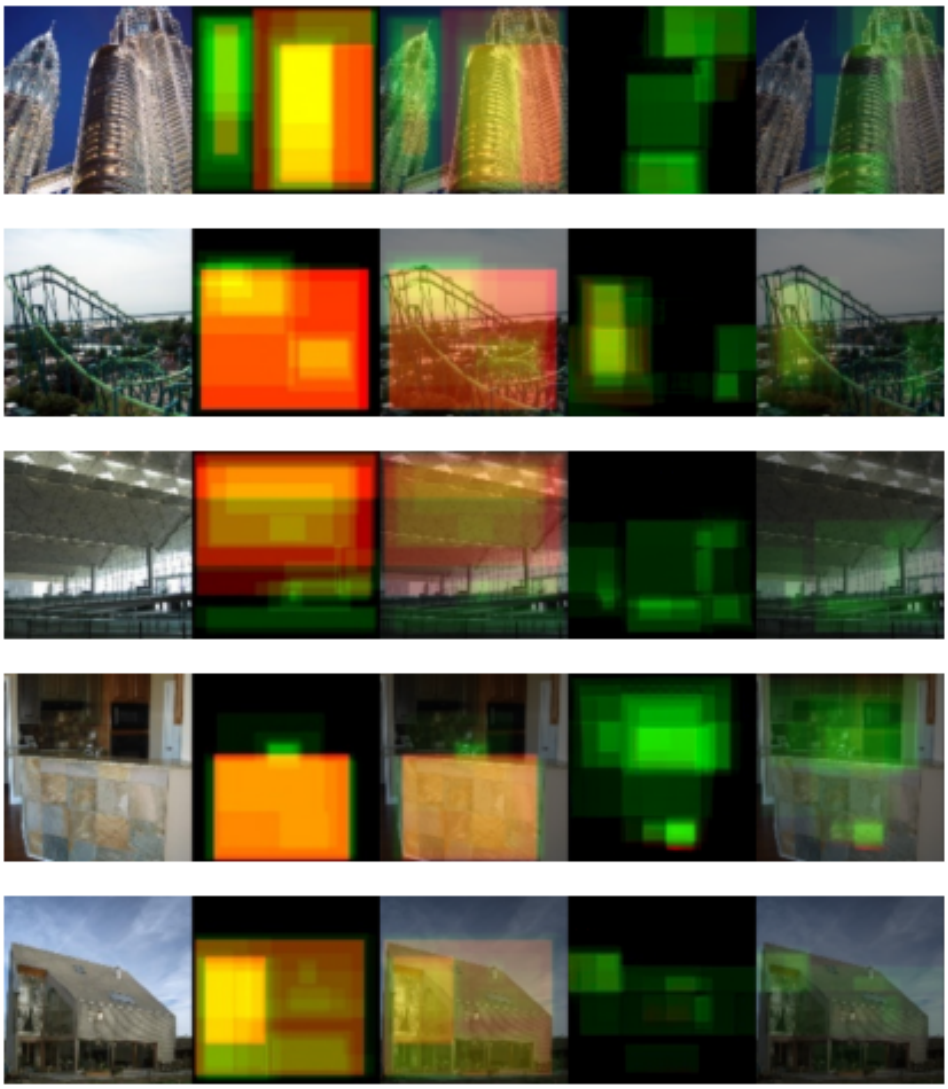

Figure 8.3: Worst GAN Outputs