University of Nottingham
UK | CHINA | MALAYSIA

Accurate Detection Methods for GAN-generated Earth Observation Images Using Expert Visual Perception

Matthew Yates

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

Supervised by Dr Michael Pound & Dr Robert Houghton

March 2023

Abstract

Image generation techniques, such as generative adversarial networks (GANs), have become sufficiently sophisticated to cause growing security concerns regarding image authenticity. Although generation and detection methods are often applied to a range of images such as objects and faces, more domain specific image types such as Earth Observation (EO) have received relatively little attention, leaving the field vulnerable to potential malicious misuse of this technology. This thesis investigates the current state of EO specific GAN generation and detection methods using an interdisciplinary approach. This work argues that further detection methods should incorporate both human and computational detection to improve current techniques. Evidence to support this conclusion is given by the following contributions:

1. A literature review of the current state of image generation and detection with respect to EO imagery.

2. A new benchmark evaluation of current GAN models in the task of the unconditional generation of synthetic EO imagery.

3. A Comparison between detection methods in both human and computer detection systems towards synthetic EO imagery that quantifies the key behavioural differences and effectiveness for each approach. The findings from two image detection studies show that these systems prioritize different image features for making accurate detections.

4. An eye-tracking image detection study between expert and novice users. The results find that experts exhibit more efficient and effective visual search strategies for detection.

5. The development of a novel framework to improve current techniques by guiding a CNN detection model using eye gaze data from self-reported high experience individuals. The results found that this approach increased detection performance over control models.

Acknowledgements

Firstly I'd like to thank my supervisors, Dr Michael Pound and Dr Robert Houghton for their guidance, assistance and feedback throughout this research project. I'd also like to thank Glen Hart and the Dstl for their project partnership and the experience with working with them. A special thanks goes to the whole Horizon CDT team for all the work they put into facilitating doctoral research, helping students like myself and making the process as engaging as possible. Thanks to all my friends and family for all their support during these long 4+ years. Finally, I would like to express my sincerest gratitude for my Mum whose patience and continuous encouragement has helped me throughout the entirety of my academic journey.

List of Publications

1. Evaluation of synthetic aerial imagery using unconditional generative adversarial networks, ISPRS Journal of Photogrammetry and Remote Sensing (2022) [Chapter 3/ Chapter 4]

   ...

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivations

The last decade has seen remarkable progress in the fields of machine learning and computer vision as more sophisticated algorithms are designed and the hardware to run them becomes more powerful and accessible. Generative algorithms for creating images and video have become much more robust with the inception of Generative Adversarial Networks (GANs).

This ability to convincingly generate and alter visual data brings a new set of challenges and threats when the technology is used with malicious intent. One recent example of this is the rise of "Deep Fakes," which use GANs and other deep learning techniques to splice one face onto another in motion, making it possible to create video of real people doing things, and with audio synthesis, saying things that never actually occurred. This is one example, which shows how this technology has serious and potentially dangerous ramifications across multiple levels, from people's personal data being altered and misused without their consent, to larger reaching consequences of national security if this technology becomes used for digital propaganda to influence large populations on a larger scale.

The implementation of generative techniques becoming more accessible and easier to use, with generative systems being able to be trained in hours using consumer level GPUs and various generative web APIs available within a web browser. The

generated content itself is also becoming more realistic and harder for the human eye to distinguish [1, 2]. Current research into the detection of generated imagery is a rapidly growing field yet research gaps persist. One shortcoming in current research is the comparatively lack of work looking at domain specific data outside faces and objects, which have already been explored much more comprehensively [3–5].

Earth observation (EO) data (e.g. satellite aerial imagery), is an example of a domain specific data type whose authenticity is relied on in a variety of technical fields from remote sensing to urban planning. The lack of specific understanding into the generation and subsequent detection of generated EO data presents a significant security threat. During the time frame of this PhD project the U.S. government publicly released military intelligence of EO images depicting Russian Forces building up on the border of Ukraine [6]. The release of these images as evidence of imminent Russian aggression drew the attention of the international community. This is one recent example of the importance of being able to trust that vital EO image data is authentic. Although there are no currently well known cases of entirely fake EO images using deep learning generative techniques, digitally altered EO data is already an issue. On 1 August 2014 the Russian Ministry of Defense released satellite images in relation to the downing of Malaysia Airlines Flight 17, yet a subsequent investigation by Bellingcat analysed the images, concluding that they had been modified using Adobe Photoshop software [7]. As generative algorithms become easier to use it is reasonable to predict that they may replace, or be used alongside manual methods of digital image manipulation, as has been found in other domains.

Despite the larger public exposure to image generation and manipulation techniques such as deep fakes, many domain specific image types such as EO data are often not considered targets and remain vulnerable to the spread of misinformation. Although there is the argument that research into novel detection methods just leads to further advancements in the generation technology in a continuous, escalating loop, it is hard to deny the importance of researching current vulnerabilities and also gaining insights into how this kind of technology functions, both technically and in human

interactions.

Aside from the defence, security, and trust implications of investigating generative EO data, it also presents a novel method for evaluating current generative models. The nature of the differences in features found in aerial imagery compared to faces and objects presents a distinct set of challenges for current generative models which may not have been extensively assessed on similar data sources.

The research for this PhD looks at novel ways to improve current detection methods for synthetic earth observation data. This thesis achieves this by evaluating the current performance of state-of-the-art (SotA) GAN models and then assessing the strengths and limitations of both automatic and visual methods of perceiving false image content. The results of these findings are then utilized to inform the direction for improving current methods of detection. In addition to new academic contributions, any findings during this research also hopes to be able to be applied towards the relevant projects and goals of the Defence, Science and Technology Laboratory (Dstl) with whom this project is partnered with.

1.2 Aims and Objectives

This project seeks to combine Computer Vision and Machine Learning with human factors research into human/algorithm interaction and visual perception. The work in this thesis explores the current state of generation and detection of image data and proposes a novel detection method for synthesised earth observation images. Earth observation (EO) data (e.g., aerial imagery) has been chosen for this project as it is a novel, domain specific image type that has not been widely researched in GAN (Generative Adversarial Networks) generation and detection, despite its importance for many applications.

The main aim of this project is to evaluate the generation and detection of GAN generated EO imagery and to improve current detection methods by using a multi-disciplinary approach from both computer vision and human factors. This will be achieved with respect to 3 main objectives:

1. Evaluate current GAN generation models on the ability to synthesis realistic EO imagery.

2. Analyse detection methods in both human and computer visual systems towards generated aerial imagery and to quantify the key behavioural differences and effectiveness for each approach.

3. Improve current synthetic image classification systems with the use of expert human detection data.

1.3 Contributions

In the areas of GAN image detection and generation for EO imagery the work in this thesis makes the following contributions:

1. (Chapter 2) A literature review of the current state of image generation and detection with respect to EO imagery. An overview of current GAN models is given as well as a review of detection methodologies for synthetic images and human evaluation studies.

2. (Chapter 3) A new benchmark evaluation of current GAN models in the task of the unconditional generation of synthetic EO imagery. Different GAN evaluation metrics for this task are also reviewed including Frechet Inception Distance and Kernel Inception Distance. During the model evaluation a synthetic EO image dataset is produced for use in further image detection studies.

3. (Chapters 4-5) A Comparison between detection methods in both human and computer detection systems towards synthetic EO imagery that quantifies the key behavioural differences and effectiveness for each approach. The findings from two image detection studies (one online and one in person) show that these systems prioritize different image features for making accurate detections.

4. (Chapter 5) An eye-tracking image detection study between expert and novice users where participants are asked to accurately identify the synthetic image in a series of real/synthetic image pairs. The results are analysed with gaze entropy metrics and visual inspection of eye fixation heatmaps. It was found that experts exhibit more efficient and effective visual search strategies for detection than participants with low experience, even when the images are novel to both groups of participants. A dataset of 3200 images consisting of real and synthetic EO images. This dataset can be found at https://www.kaggle.com/datasets/matty0512/expert-gaze-maps-for-realfake-eo-images.

5. (Chapter 6) The development of a novel detection framework that improves on current techniques by guiding a CNN detection model using eye gaze data from self-reported high experience individuals. The results found that this approach increased detection performance over control models. The implementation of this model provides evidence that expert visual detection data can be used to improve existing computational models of synthetic image detection for domain specific image types.

Chapter 2

Background

2.1 Deep Learning based Generative models

Digital image generation has become an increasingly important area of Computer Vision since Generative Adversarial Networks (GANs) were first outlined in 2014 [8]. This field covers a wide range of tasks such as unconditional image synthesis [9, 10], super resolution [11], anomaly detection [12] and more recently text-to-image [13]. As the field has grown and matured there now exist several different core architectures that are used for achieving the different generative tasks and a further variety of models within each respective category [14]. The performance of these models has also improved over time as models are updated in both complexity and size [2].

2.1.1 Generative Adversarial Networks

Until the explosion of Diffusion models in 2022 the most popular class of generative model was GANs with hundreds of different variants applied to different problems and data types [15] often with state-of-the-art performance. GANs have been used as the basis for many different generative tasks [11, 16] but are suited to unconditional image synthesis starting from a random latent vector [2, 17]. Their spread has led to substantial large media coverage [18] of NVIDIAs StyleGAN series of models in particular for their ability to quickly synthesis photo-realistic faces.

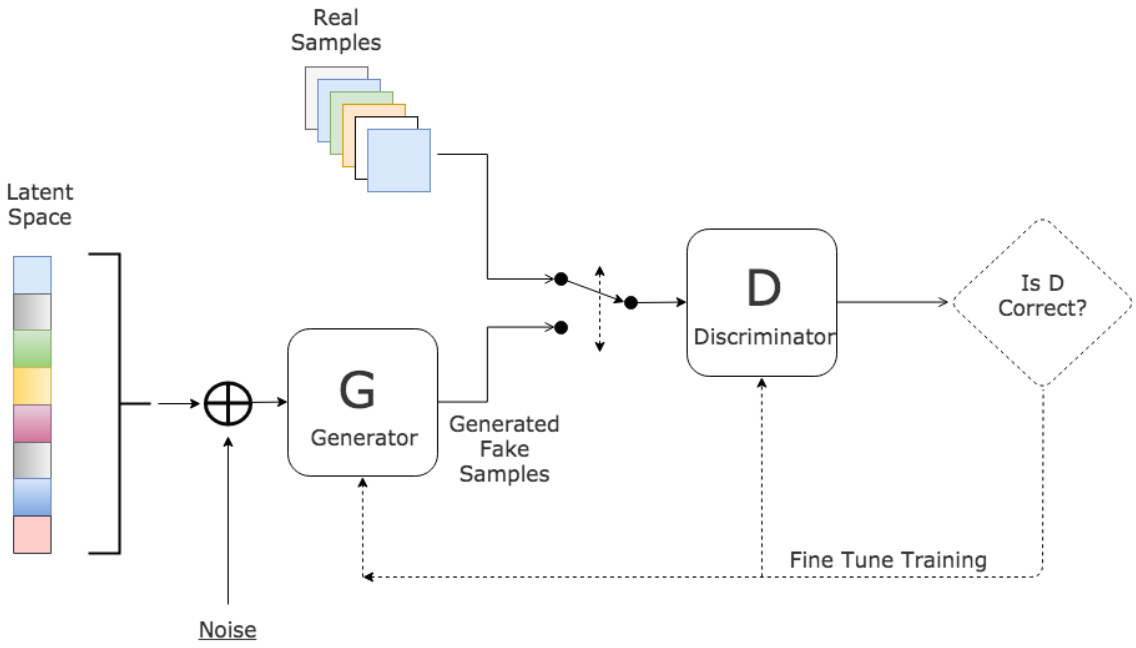

Figure 2.1: *Basic structure for Generative Adversarial Networks (GANs). GANs work by taking input as a latent vector and learning to generate an image based on the distribution of features from a training dataset. The model consists of a generator G and a discriminator D. D is tasked with learning the training dataset, whilst G is tasked with generating images. G will generate an image which is passed to D for prediction, if the image doesn't fit the dataset, the loss from the prediction is backpropagated through G. This continues until G generates an image that D predicts came from the training dataset.*

While the exact details can vary wildly between specific models, the underlying concept behind these models is that two opposing neural networks, a generator (G) and a discriminator (D) that are pitted against each other in a zero-sum game [8]. Network G creates an image which is given to the D network in an attempt to fool it into a false classification. The resulting loss from this classification is then used to improve both networks via gradient descent until D cannot distinguish the synthesised image from the real image. This process can be summarised as the equation (2.1):

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim \text{pdata(x)}}[\log D(x)] + \mathbb{E}_{z \sim \text{pdata(z)}}[\log(1 - D(G(z)))] \quad (2.1)$$

The inception of GANs saw a huge increase of interest in the field of data generation [15]. Although the original network demonstrated the ability to use neural networks

to generate data, rather than simply output a prediction like for classification tasks, there are still challenges in both the research and application of the models. One of the main problems that GANs can encounter is that of mode collapse [19]. Mode collapse is a problem that can occur during training when the GAN becomes stuck in generating limited and repetitive outputs, failing to capture the entire diversity of the target generation. One of the main causes for this in earlier models is the training balance between the generator and the discriminator. If the discriminator does not adequately learn the distribution of the training data, the generator is able to fool the network with inaccurate images. This results in the generator receiving no feedback loss from the discriminator to adjust its weights. Mode collapse can also occur from an imbalanced dataset, if some features (or modes) are more frequent in the data, the GAN will not produce images that capture the full distribution of the data. Conversely, if the data is too complex the GAN may also struggle to fully capture the diversity of the full dataset.

Aside from mode collapse another issue with GANs is that they are resource heavy to train. This is a challenge that that many deep learning architectures face, but particularly applies to GANs. As training the model requires optimising two separate networks, this can quickly become computationally intensive as deeper networks are used and the models become more complex. GANs also require large amounts of training data to produce high quality generated samples, as more recent models seek to generate higher resolution and more realistic images, the datasets required to train them also increase in size, adding the computational costs.

Other challenges that GANs face are interpretability and control. One criticism of deep learning models in general is their inherent black box nature [20], that is, the reasons for producing a specific output are often opaque to researchers. It is often hard to interpret why a GAN produces specific output patterns and researchers often have to speculate on what features a GAN prioritises when learning a given data type. This question has led researchers to investigate methods to control what output a GAN produces and to generate data with specific features. As GANs learn the overall distribution of a dataset, rare or uncommon features for a given data type

may be harder to generate. This also produces the problem of bias and fairness, if there is a bias in the training dataset, this will be reflected in the output images. Since Goodfellow's original 2014 implementation there have been numerous deviations and updates to the original structure with several models bringing noteworthy changes that have led to increased, training speed, stability, image quality and overall performance. Unconditional GAN models for image synthesis will be looked at in more detail in Chapter 3, as these are the primary GAN variants that this research investigates. Aside from unconditional image synthesis GANs, there have been many other GAN architectures for other generative tasks. The next section provides a high level overview of the different GAN types as well as some of the more impactful variations on the original architecture. As there are 1000s of different types of GAN models, the few models that were selected to be discussed here each brought either new innovations to the architecture or significant steps in generative image quality.

CycleGAN

CycleGAN [21] is an image-to-image translation model which attempts to learn the identifiable features of one image and then translate those features into a target image. This could be a texture or "style" of one set of images which is then translated to another set of images whilst keeping the overall structure and features of the target set. The significance of this model is that it was on of the earlier GAN models that showed uses for the architecture in novel tasks outside of unconditional image generation. This model has become one of the most popular models for style transfer and image to image translation tasks.

Wasserstein GAN

The Wasserstein GAN [22] improves upon the performance capabilities of conventional GAN architectures by making a few important improvements. Instead of using the discriminator to evaluate the generated samples as real or fake as in a conventional DCGAN, this model instead uses a "critic" that scores each sample image

based upon perceived realness and fakeness. This is calculated by measuring the distance between the distributions of the real and fake datasets. The distance metric used is Wasserstein distance [23], also known as the Earth Mover Distance, which calculates the optimal way to get from one distribution to another. The advantage of this over a conventional DCGAN is that it can optimize the GAN without having to balance the training of both the discriminator and generator, reducing the likelihood of mode collapse from an imbalance in the strength of either network. This happens as the critic is trained to its optimal state and then provides a loss for the generator to continue to optimize.

$$W_p(P,Q) = (\inf_{J \int (P,Q)} \int ||x-y||^p dJ(X,Y))^{\frac{1}{p}} \tag{2.2}$$

Due to the training stability and robustness provided from using Wasserstein critic, this method has been widely adopted amongst other successful GAN variations such as the popular models PGGAN and StyleGAN [17].

PGGAN & STYLEGAN

By utilizing several different techniques for increasing GAN performance such as Wasserstein Loss, PGGAN set a new benchmark in performance and image quality for unconditional image generation [24]. In addition to creating almost photorealistic faces PGGAN can generate images to a resolution of 1024X1024p, whereas most previous GANs had been only capable of working to image resolutions of 512X512p or lower. The core feature that is proposed by PGGAN is its "progressive growing" architecture. The model first learns to generate an image at a low resolution (e.g. 4X4) and when performance converges at that resolution it then up samples the image to a higher resolution, repeating this process until the maximum resolution of 1024X1024. This method means that the network is forced to first embed large scale structures at the lower resolutions before progressively adding more fine features and details towards the end of training. This approach speeds up the training process and makes it less prone to mode collapse, as compared to earlier architectures that

process only a single resolution size.

StyleGAN [1] is another progressive growing GAN which exceeded the performance of PGGAN in terms of image quality and visual control of features. The model architecture used the fundamental training progress of PGGAN, the progressive growing of image size, but added a few additional changes to the generator. These changes included Bilinear sampling for the upsampling between layers instead of PGGAN's nearest neighbour layers, the addition of Gaussian noise added prior to each activation map and also a standalone mapping network. These changes were implemented to allow StyleGAN to conditional generate images of different conditionally learnt styles. This also led to StyleGAN being able to outperform PGGAN in unconditional image synthesis tasks at various resolutions. StyleGAN2 [25] included further improvements to the StyleGAN architecture, including the replacement of the progressive growing feature. Both PGGAN and StyleGAN2 will be explored in further detail and benchmarked for generation performance in Chapter 3. Both PGGAN and StyleGAN have been chosen to be looked at due to their significant advancements in image quality and training stability.

2.1.2 Other Generative Models

GANs are currently the most popular and researched class of generative models and for these reasons are the primary focus in the work within this Thesis. Despite this it is also important to note that there are several other generative models that are widely used and rival the top GAN models for performance in specific tasks.

Transformer Networks

Transformer networks [26] have been found to either match or surpass GANs in generated image quality with potentially better scalability and faster and more robust training. The transformer network architecture is a deep learning, self-attention model that is commonly used for natural language processing (NLP) tasks where it is the leading class of model in terms of performance and usage [27]. Transformers

have more recently been applied to computer tasks such as image generation [28]. The theoretical advantage of transformers is that their central self-attention mechanism overcomes the poor global spatial understanding of image features inherent to convolution-based models [29] and can track more long-range dependencies in the data. One of the current issues with Transformers is that they require large amounts of data to train which also requires substantial computational power. Despite the potential limitation of the necessary computational power needed to train, transformers have seen a huge recent increase in use for image generation, particular in combination with Diffusion models [30] for text to image tasks.

Variational Autoencoders

Variational autoencoders (VAEs) are another class of model that is often seen in image generation task [31]. These models learn to first encode a given input such as an image and then to reconstruct the original input based on these learnt encodings. Although VAEs have not shared the same state-of-the-art (SotA) results of the top GAN and Transformers models they are usually much quicker and less computationally intensive to train [32], making them a more viable option in cases of limited resources. Some of these shortcomings have been overcome with VAE-GANs [33]. A combination of both VAE (Variational autoencoders) and GAN features, providing the advantages from each respective class of model. OpenAI's original DALL-E 1 model [13] is one notable model that partially relies on a VQ-VAE component to learn the encodings for given input images.

Diffusion Models

Diffusion models [34] are an example of a more recent model type to be applied to image generation. These are likelihood-based models that learn the encodings of an image distribution while applying an increasing amount of Gaussian noise to the image at each time step. The network then learns to "denoise" the image back to the original noise. Diffusion networks have been found to beat GAN performance [35] in terms of standard evaluation metrics (e.g. FID) but also in terms of diversity

of generated samples. A large drawback with current Diffusion models is the long training and generation times, making them a lot slower than GANs. This is seen to a greater extent in generating images, a trained GAN model can very quickly generate large amounts of samples for a set of input vectors. As Diffusion networks work by denoising an image of random noise through a series of linear time steps it is much slower as this process is required for each generated sample.

Recently diffusion models have risen sharply in popularity due to forming the backbone of many large text to image models such as OpenAI's DALL-E [13] and Stable Diffusion [36]. These are large models trained on vast quantities of data making them capable of generating high quality images from highly specific text prompts. Although these models achieve very impressive results, this thesis will not include any detailed focus on them as the resources needed to train these models to SotA are simply not realistic for many researchers, due to the cost and access of the computational power required for training. DALLE 2 for example was trained on a dataset of 250M images, contained a decoder of 2.3B parameters and was trained on 256 V100 GPUs for 2-4 weeks [37]. Smaller models with similar architectures are able to be trained with less computational power and data but are not able to achieve the same high resolution and image quality as current GAN models under these circumstances.

2.1.3 Image Generation Tasks

The types of generative models discussed have been applied to a wide range of image generation tasks with different aims and objectives. The most fundamental generative challenge is unconditional image synthesis [17]. As the name suggests, this is referring to the generation of image samples from a given dataset with no other conditions (e.g. labels). Examples of GAN models that have achieved exceptional performance in this category include PGGAN [17] and BigGAN [2], the evaluation of these models is usually concerned with image quality and image variation.

As an extension of the objectives of unconditional image generation, conditional

generation models are designed to synthesise images based on additional task specific constraints, offering more control over the types of samples produced. This is achieved using class labelled datasets [1, 38], StyleGAN3 [39] is an example of a SotA conditional GAN, being the latest iteration of NVIDIAs original unconditional PGGAN [17] architecture. One disadvantage with using conditional generation is that it requires more training images than unconditional generation as there needs to be sufficient samples for each target class. Unconditional and Conditional image generation models have been applied to a wide range of image types, in particular the synthesis of photorealistic faces [1] and large object datasets such as ImageNet [2]. Currently the top implementations of these models are very computationally expensive to train, requiring long training times and multiple GPUs to produce large resolution (e.g. 1024 x 1024p) and high-quality images. Despite the current limitations, the advancements for both conditional and unconditional generation have seen massive leaps in terms of photorealism and image quality when comparing examples from the original GAN implementation in 2014 [8] to those generated by the 2021 model StyleGAN3 [39]. Recent work [40] has found that the face generation abilities of the most advance models (e.g. StyleGAN3) are now indistinguishable from real face photos in the context of human visual perception 2.2.

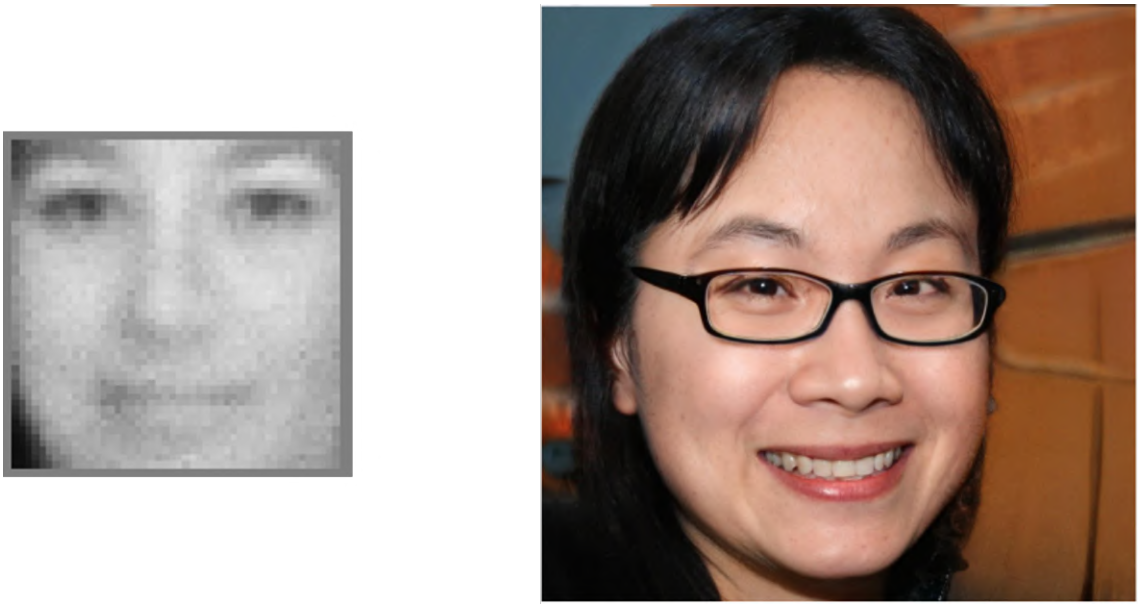

Figure 2.2: *Examples of the progress between the original GAN implementation (left) [8] in 2014 and more recent models such as StyleGAN3 (right) [39]*

Style transfer is another powerful application of generative models [41]. This involves using an image-to-image translation algorithm (e.g., CycleGAN [21]) to transfer the visual "style" and textures of dataset A onto dataset B while retaining the defining features such as objects present in dataset B. One of the core models used for variations of this task is CycleGAN.

The CycleGAN architecture consists of two complete GAN models and works by generator A taking an input image from dataset A which it then presents to discriminator B which attempts to distinguish it as being from dataset A or B. Using the loss from this decision generator A learns to generate an image from dataset B using dataset A. Another system which is in place in the model is the generated image is then used as input into generator B which tries to reverse engineer the generated image to create a fake version of the original image which is then compared with the original using a cycle consistency loss function. Examples often used to demonstrate the abilities of these models often rely on benchmark datasets for tasks such as changing hair types in human face images [42] or transferring one artistic style to a contrasting image of artwork [41]. These methods have been shown to be applicable to other domain specific tasks such as creating road map style images from overhead RGB satellite images [43].

2.2 Synthetic Image Detection

As generative models become increasingly popular, easier to implement and more applicable to a wider range of tasks, the research into detection methods becomes a higher priority. This is evidenced by the broadening field of research dedicated to deep learning based image detection [43]. Currently there are a wide range of proposed detection methods which can vary in terms of performance and technique used depending on the specifics of the detection target. Models have been created specifically to detect manually created fakes (e.g. Photoshopped images) [44] or the diverse types of generative model-based fakes. There are unique challenges for each domain specific image type making it a difficult research goal to create a universal

detection model, although there are some papers which claim to have developed such an approach [45], although these results have been disputed [46]. As this Thesis is primarily concerned with CNN based GAN models, the detection literature presented in this section will reflect this focus and look at detection techniques for CNN synthesised images.

For the detection of CNN synthesised images (e.g. GAN samples) many of the successful techniques [45, 47] involve the use of CNNs as image classifiers [43] as the core method in combination with various pre-processing or post processing techniques. A 2018 survey of detection techniques [43] ran a benchmark test on a selection of the highest performing detection methods on a variety of different Tensorflow datasets [48] all generated from CycleGAN and used the performance of its discriminator network as a baseline. Multiple tests were conducted for each dataset with various levels of Gaussian blur added. This was implemented to simulate the kinds of image artefacts caused from jpeg compression, which is often seen on social media images, giving the study a higher level of ecological validity. They found the best performing model in this comparison was a CNN based model [49] followed by the large CNN model, XceptionNet [50] and a third method using Steganalysis based method [51]. The hardest datasets were found to be win2sum (a dataset which transfers the season of image A to that of image B) and map2sat (transfers the roadmap style of image A to the topography shown in image B), both datasets had few obvious artefacts when viewed with the human eye. All the models saw various drops in detection performance when tested on the jpeg blurred images which questions the robustness of these techniques when deployed in the real world. Out of the models tested, XceptionNet showed the lowest drop in performance between the control dataset and the compression datasets.

Following on from the results of the benchmark survey the researcher proposed a new GAN image detection method [52] based on techniques from the field of Steganalysis in combination with a deep CNN classifier. The model first computes the co-occurrence matrices directly on the image pixels for each colour channel (RGB) then passes them though a standard CNN classifier. This method works

as images generated by GAN image to image translation have differences in their pixel statistics compared to real images. These differences can then be discovered by image forensics methods made to bring out latent information in image data. The results found that this new model vastly improves performance compared to the models tested in the previous survey across most of the datasets, particularly in the win2sum and map2sat datasets which were found to be the most challenging to classify. However, it should be noted that this new method performed relatively poorly against the other models on the Facades dataset [53], again highlighting the difficulty in making a universal detection method. This method was also only tested on style transfer datasets.

The use of co-occurrence matrices as a detection tool is also found in another paper concerning GAN image identification [54]. The detection method outlined uses a feature set composed of co-occurrence matrices extracted from the residual images of different colour components of both the real and generated samples. The paper concludes that there are identifiable markers present in generated images when looking at the HSV and YCbCr colour spaces. This is because of the inherent differences in how colour is processed in creating images naturally (via a camera then converted to digital) and during generation (up sampling from a small latent vector until then transformed into a tensor with three channels at the end). The method is tested on generated faces and bedroom furniture images. One potential limitation of this method is it only targets models that leave these specific fingerprints and may not work in the presence of additional post processing of the images such as added filters (e.g. jpeg blur).

For specific images types such as faces, more specialized detection techniques have been proposed. Most notably are the various detection methods for synthetic face images. As face generation has been extensively researched in comparison to other more domain specific imagery, the visual fingerprints of what to look out for in distinguishing between real and fake are well established. Generated images can be exposed by looking for common image artefacts like unnatural asymmetries in the facial features, different lighting conditions between the eyes and malformed

accessories such as incoherent necklaces or ear rings [55]. On initial inspection of these images they may appear to be photorealistic but when the individual features of the faces are isolated, these image artefacts become easier to identify. Facial landmark locations have also been examined as a flaw in GAN photorealism [56]. Similar to the issues of asymmetry, GAN faces may often exhibit an unnatural configuration of the facial features, even if the individual features are all rendered themselves to a high degree of photorealism.

Contrary to much of the current detection literature Wang et al. [45] published a CNN detection paper with the claim that GAN image detection is a solved problem, presenting a pre-trained CNN model which is capable of classifying GAN images with high accuracy across diverse synthetic image datasets generated by a variety of the most popular SotA GAN models.

This model uses a ResNet architecture [57] with a customised training procedure. Residual Networks, or ResNets, are a set of deep convolutional neural network structures. They consist of a deep architecture with many convolutional layers (e.g. ResNet-50 contains 50 convolutional layers), and use residual blocks to facilitate efficient learning despite their depth. Residual blocks contain regular convolutional layers with the addition of skip connections. Skip connections take the input to the residual block and add it directly to the output of the block, bypassing the convolutional computations of the block itself. Skip connections $H(x)$ are therefore defined as $H(x) = F(x) + x$, where $F(x)$ is the output of the residual block and x is the input to the block. The advantage of skip connections in CNNs is that they prevent the degradation problem that can occur in very deep models where the ability to propagate information throughout the network is lost. Skip connections allow learnt feature embeddings from lower layers in the network to flow through to the higher layers. The use of skip connections and residual blocks has seen ResNets become a common choice of CNN architectures where very deep models are needed.

The proposed GAN detection model from Wang et al. uses a ResNet-50 trained on several real image datasets and corresponding synthesised images created using PGGAN [17]. To improve the model's ability at generalising, the training images

undergo a series of augmentations including adding various levels of Gaussian blurs to add robustness in the presence of common jpeg compression artefacts. In addition to achieving high accuracy (90%) on images from PGGAN, the paper also presents the results for testing the model's classification ability for other popular GAN architectures, including StyleGAN2. The paper concludes that this relatively simple method of CNN detection works well across most current GAN generation models and that GAN detection is not currently the security issue that other research alludes to it being.

These claims have however been questioned through work covered in a more recent paper ([46]). This work found that the CNN detection model proposed by Wang et al. (2019) struggled to perform at an adequate level needed if it was deployed in a real-life scenario. The detection performance from the original paper was not able to be replicated in these new conditions. Like the tests in the detection models original paper, this study tested the model using a dataset of StlyeGAN2 generated faces. Novel detection models are an important part of the image detection field but research into the specific properties and features present in generated images is equally important. A more analytical approach to fake image detection is taken in the paper "What makes fake images detectable? Understanding properties that generalize" [58]. The authors explore the general properties of fake images using a fully convolutional patch-based classifier with a limited receptor field to focus on parts on the image that appear fake. The model is trained to identify local patches which appear contain GAN image artefacts rather than focusing on the global semantics of the entire image. A technique to exaggerate these "fake patches" so that they can be used to further improve detection methods is also demonstrated. This is a novel approach as it uses a detection technique that looks at local image artefacts for predicting if an image is fake or not rather than global semantics such as colour spectral analysis or analysis of the image in the Fourier domain. They also include a method for displaying the results of the classifier as a heat map which gives us insight into what features are used for detection. This is particularly relevant at a time where there is a push against "black box" methodologies [59] and a need for

more explainable machine learning models.

Despite this interesting approach in trying to understand the visual fingerprints of fake images, the proposed patch-based classifier does not achieve the same level of performance as reported by other models [45, 54]. Although detection performance was not state-of-the-art, the authors presented the results of which features in the images were most used for prediction. They found that background textures, mouths and eyes were the most used features for detection. In particular, the generated models seemed to struggle when it came to replicating sharp boundaries between features such as background and head or hair and face. This understanding of what to look for in trying to distinguish between real and generated images is useful for improving detection models and in training humans in detection, an often-overlooked factor of this area of research.

Spectral analysis of GAN generated images is a detection method that, until recently, has shown to be one of the most reliable methods for differentiating between real and generated images [60]. When images are analysed using Fourier transformation to produce a frequency spectrum, the differences between real and generated images become much easier to spot for many of the most commonly cited GAN models. Although this is a relatively reliable method of GAN image detection, at least compared to some of the other methods discussed, a more recent paper claims to be able to counter such spectral analysis techniques[61]. By introducing a novel preprocessing pipeline for GAN generated images, this method is able to mitigate the occurrence of spectral artefacts which act as GAN fingerprints when using spectral analysis methods. Additionally this method was found to also decrease the reliability of spatial domain based detection methods such as the aforementioned CNN detection model by Wang et al (2019), although the impact on accuracy was less than that found for the spatial domain models. A similar method of attacking GAN detectors by removing the key GAN fingerprints was presented in Wesselkamp et al. (2022) [62]. By using targeted and untargeted fingerprint removal, this method was also able to mislead Fourier spectrum based detection methods. This recent set back to GAN detection highlights the rapid back and forth arms race between

detection and generation. The nature of this dilemma presents the need to be constantly researching novel methods of detection to expand the array of available tools for countering increasingly sophisticated fake image generation.

2.3 Human detection

The images produced by GANs are usually created for the purpose of being viewed by humans, with or without the intent to deceive the target. For this reason, it is important to investigate human visual perception towards generated imagery. This is accomplished through measuring the extent to which humans can reliably detect fake images and evaluating the factors which impact the behavioural mechanisms of image detection such as image visual realism (IVR).

Image visual realism (IVR) [63] describes the perception of photorealism in an image and has been a topic of interest in the wider field of digital forensics and security outside of the field of Generative modelling. IVR and human perception has been investigated using visual search experiments from cognitive psychology. One example of this is a 2015 study [64] on visual perception towards manually altered digital images. The researchers conducted an online user study were participants were given a visual search task consisting of viewing a series of images, with the goal of detecting whether the images had been altered or not. As well as answering whether the image contained any alterations, the participants were also asked which region of the image had been altered and what their confidence was in their answer. The results from the experiment found that humans have difficulty detecting alterations in images even when they are actively primed to look for them. Participants were able to correctly identify whether an image had been edited 58% of the time but could only actually identify the forgery in the image in 46.5% of cases. It was also found that that younger participants and those with relevant experience to the experiment had greater levels of performance than other groups. The behaviour of participants (time, hints, confidence levels etc.) also had an impact on the success rate. As some of this behaviour can be taught, it suggests that with additional

training it is possible for people to vastly increase their performance in similar tasks. Image attributes that contribute to IVR (Image Visual Realism) were investigated in the context of non-AI based computer generated imagery in a 2017 study [63]. An interdisciplinary approach was taken with the goal of understanding how humans perceive visual realism and to employ this understanding to computational models. A visual realism dataset was created which contained over 2000 realistic computer-generated images as well as photographs of similar scenes. The images depicted in the dataset included faces, objects, and building. The first experiment asked participants to make a binary decision on whether the images presented were real or computer generated (CG). Participant performance was analysed for expertise groups and gender, with groups of higher expertise (graphic designer, photographer, gamer) achieving higher levels of accuracy. The second experiment was a similar setup to the first but instead looked at identifying factors related to the human perception of visual realism. This experiment asked participants to use a 5-point Likert scale to rate realism, familiarity, colour and illumination, attraction, and objects.

An ensemble of computational models consisting of convolutional neural networks and decision trees were created to predict these indicators of visual realism by learning from the annotated visual realism dataset. The results found that the computational framework created was able to predict the level of visual realism for the images in the dataset in correlation with human evaluation of the images. Although this study did not look at samples from generative models, the mixed methods approach provides a useful framework for further research into human perception and fake image detection.

Visual perception studies using human participants as evaluators are being increasingly common [40, 65] in the study of GAN image detection. A 2021 study evaluating both human and algorithmic detection of GAN images [46] found that humans have difficulty identifying real from fake images across multiple face and object datasets and the current top detection methods [45] are not yet robust enough in their performance to be deployed in real world scenarios. In the case of human detection,

the authors found that participants with previous AI experience achieved greater detection accuracy than the general population. These findings suggest that there are experiential factors which contribute to human detection ability, leading to the questions of how this could be used to improve current detection methods, both human and algorithmic.

One limitation with much of the research highlighted here for both human and automated methods is that there is little crossover between the two. The methods looking at human evaluation [46] measure accuracy in differentiating between real and synthetic and the factors associated with this but don't seek to improve on current detection methods. Conversely, research into computational methodologies for detection, neglect to look at human evaluation, an important factor to consider when photo realistic GAN generated images are often produced to be viewed by humans rather than to bypass computational detection systems. This thesis addresses this by looking at both detection methods (human/automated) and how they can be integrated to form a more comprehensive detection strategy, utilizing the advantages from each.

2.4 Earth Observation Data

Earth observation (EO) data encompasses a wide range of imaging techniques for obtaining information on the different physical systems of the Earth [66] and has many different applications spanning different fields of interest. The most common EO images come from visible spectrum satellite photography, but other image types [67] include passive imagery such as multi-spectral, hyper-spectral, microwave-radiometry and active imagery like Synthetic Aperture Radar, Lidar and GNSS reflectometry. Although there are different ways of obtaining EO data, this thesis will primarily refer to satellite-based remote sensing and use the term EO data with this assumption, unless stated otherwise.

The use of EO imagery is most prominently seen in the studies of remote sensing and mapping but is also a crucial data source in many other industries such as

environmental disaster warnings and management [68], urban planning [69] and the economic assessment of regions [70]. EO data is also heavily used for security and intelligence purposes [71], and therefore a potential target of misinformation and counter-intelligence attacks[72]. This application of EO data is of particular relevance in light of the 2022 events in Ukraine, where satellite imagery has played a key role in both military planning and the responses from the global media [73, 74]. Another concern with fake EO images is that they can be challenging to easily interpret for untrained viewers [75]. This contrasts with more familiar and common images that humans regularly encounter such as face or objects that require high levels to pass as real and not fall into the "uncanny valley" of photorealism. The implications of this being that it may be much easier to produce fake EO images that fool human detection that are of lower quality than needed for generated samples containing either faces or objects. EO images data for generative modelling tasks is a relatively unexplored area of research when compared to other image types such as faces and object datasets [25, 56]. This gap in research knowledge for EO data and generative models poses a significant security risk as the current challenges and limitations of generation and detection of EO data has not yet been rigorously evaluated.

Although there is not the same volume of publications for EO image generation as in other areas, the use of Satellite imagery and GANs has been explored in a few key scenarios. Aerial image data has been found to be used in work involving image to image translation for tasks such as mapping, in the case of GeoGAN[76] this was to use aerial photography to generate road maps of the same topography. Likewise, image translation methods have also been applied to other remote sensing applications such as estimating ground level views from aerial images[77] as well as a method for converting synthetic aperture radar (SAR) data to optical [78, 79]. Other work has looked at the use of satellite image training data for cloud removal for optical [80] and hyper-spectral imagery [81].

These uses of GANs and EO data are often focused on being applied to industry specific challenges, there has been less of a focus on the security and detection

aspects of synthetic aerial imagery. The first evaluation of synthetic aerial image deception comes from Zhao et al. [82] which investigated detection methods using synthetic EO samples generated from image translation GAN models. Also noting on the relative lack of research on this topic, the authors remark that while there are only a small number of case studies involving fake satellite imagery, it only takes a few impactful uses of deep fakes for all subsequent instances of geospatial data to be questioned on authenticity. The concluding remarks encouraging further research into fake EO data detection as an important pre-emptive measure.

Building up on this work regarding fake aerial imagery, a recent article (2021) has proposed a detection method specifically built for this task [83] and is currently the only peer reviewed model to be published. The model, GEO-DeFakeHop uses parallel subspace learning (PSL) where the input image space is mapped to several different feature spaces based on multiple filters. Detection between real and fake images is made by evaluating the differences of different channels for the different filters. The model is light-weight with a comparatively small number of parameters (0.8-62K) making it easy to train on limited hardware. On the UW fake Satellite Image Dataset, the authors report robust performance even under various noise conditions such as compression and resizing. The authors do note that the UW dataset only contains data from 3 different cities and generated using a single GAN and so may be more challenged by detection tasks using larger multi-city datasets generated using an ensemble of GAN models. Another issue found with the research presented is the reported performance metrics are only from a small single dataset (4K samples) which is then given a 70%, 30% split into training and testing, giving no indication at how well this model generalises to other datasets.

Despite the extensive applications of EO data and the recent flurry of work on image generation and detection, there is still a concerningly small corpus of literature investigating the intersection of these topics. The papers that have been discussed that do tackle this issue are not comprehensive enough to adequately solve this problem. The work in this thesis addresses this with further investigations into the generation and detection of GAN produced synthetic aerial imagery.

2.5 Expert Visual Detection

The literature[64] for human visual detection of real and fake digital imagery, either deep learning based or traditional methods, often looks at the role of prior expertise as a signifier of detection ability. Evidence suggests that populations experienced in looking at either forged imagery or the type of images the fakes are aiming to imitate (e.g. geospatial experts looking at fake remote sensing data) are significantly more likely to be able to distinguish between real and fake. With this established link between expertise and detection ability, further research aims to explore how detection strategies between experts and novice differ and how expert behaviour can be analysed to create better detection methodologies and more optimal methods of training.

Expert visual attention has also been extensively researched in the field of remote sensing. As previously mentioned, aerial images have been found to be a hard stimulus for humans to visually process compared to objects and faces and even when compared to ground level scenes [84]. This makes a measurable difference between experts and novices in viewing aerial images and identifying features that they contain. Studies [85] have found that this is not a quickly acquired skill either, an experiment that compared visual memory of aerial scenes between 1st year geography students and 1st year psychology students found no discernible difference in response. Both groups were found to perform much lower than a third group of experts with >15 years of experience in analysing aerial imagery.

Expert and novice differences in analysing remote sensing data was also explored in Wardlaw et al (2018) [86]. Using an online citizen science platform, participants were asked to annotate and classify surface changes on satellite images of Mars. Participants were from two groups, an expert group consisting of Planetary Science researchers, and a novice group from public visitors to the online platform. It was found that while both groups were able to spot surface changes in geological features between images, the novice group were not able to classify these features to the extent of the expert group.

In regards to synthetic EO imagery, there is a lack of research into the effect of expertise on detection. The previous literature on expertise in remote sensing and fake imagery suggests that there are cognitive behavioural differences between experts and novices. Based on this evidence, the work in this Thesis aims to explore these previous findings with synthetic EO imagery. This involves investigating the role of expertise in this context for the goal of furthering the development of novel detection techniques.

2.6 Methodology

The research in this thesis uses a mixed methods approach with a combination of experimental designs from cognitive psychology, human factors, and computer vision. The data collected comes from benchmark surveys of deep learning architectures as well as from empirical studies involving human participants, likewise, the evaluation metrics used for analysing results also encompasses automated and human measures.

This section will give a brief introduction to the research methods used in this Thesis, with specific and more detailed descriptions being found in the relevant chapters.

2.6.1 GAN Evaluation Metrics

There are a range of measures that can be used to measure aspects of GAN performance with fewer standardised procedures than seen in the evaluation of other model types such as classification models. This is because, for generative tasks, the primary aim is to produce images of a certain quality or containing specific features. Image quality may be considered by some to be an inherently subjective measure which can make it a difficult measure to quantify, leading to multiple different metrics to be applied to the task.

Automated Metrics

The aim of many GAN evaluation metrics work by attempting to evaluate the differences in data distributions between the input, real dataset and the output, generated dataset. The closer these distributions are, the more likely they will appear homogeneous, with identical distributions indicating that the two datasets are identical in features. The most commonly cited methods using this approach are Inception Score (IS) based metrics and are found in the majority of GAN research papers.

Inception Score takes in a dataset of generated samples and returns a single value score which aims to measure image variety and features. This is achieved by using the feature space of a pretrained [87] instance of the Inception V3 CNN model. By looking at the distribution of probability scores across the different classification categories from the network output, feature rendering and image variety can be estimated. If the generated images return a uniform distribution of probabilities across classes, then there is less likely to be any sort of identifiable object in the image compared to a distribution that is skewed towards certain classes, indicating that the images do contain recognisable features. Image variety can be estimated by looking at the label distribution across the entire dataset[88].

Inception score has been found to be limited for several reasons, the first being that it only measures whether an image has been formed rather than how close it is compared to the target dataset. As the metric is also based on a pretrained classifier it is only useful when the target data contains features that have been learned by the classifier network. This presents an issue if it is applied to a domain specific image type, not covered by the dataset used to the train the classifier and therefore will not have learnt the correct image features. Additionally, IS cannot detect cases of generated data overfitting to the training data. Although still sometimes used in GAN evaluation papers for measuring sample diversity and quality, IS has been largely superseded by Fréchet Inception Distance (FID)[89]. FID is recognised as being a more reliable[90] and useful measure of GAN image quality in that it evaluates the

generated images relative to the target dataset, unlike IS which only measures the generated image distribution compared to the Inception network class probabilities. Like IS, FID still uses the Inception Network in calculating an image quality score but instead of using the class probabilities, it uses the feature maps from the final convolutional layer of the network and measures the output activations of this layer for both the generated images and the real target images. These activations are then summarised as two multivariate Gaussians and the distance is calculated between them using Fréchet distance 3.1 [89]. Like IS, this is then calculated into a score where a lower number indicates that the two datasets are closer in image quality and distribution, with the score of 0 indicating that all the images could come from the same dataset.

$$FID(r, g) = \|\mu_r - \mu_g\|_2^2 + Tr\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}}\right) \tag{2.3}$$

The calculation for FID can be seen in equation 3.1. Where r and g are two multivariate Gaussian distributions with means μ_r, μ_g and covariances Σ_r, Σ_g. The distributions in FID are the feature vectors of the real and generated images, $\|\mu_r - \mu_g\|^2$ is the Euclidean distance between their means and Tr represents the trace operator to sum the diagonal elements of the matrix

Although FID is considered[91] a more useful tool for evaluating GAN image quality than IS it is not without its own pitfalls and caveats. FID assumes that features present in the samples have a Gaussian distribution which may not be the case depending on the specific datasets used. Another issue to consider with FID is that it requires a large sample size to accurately assess quality ($> 50K$ samples) and low sample sizes can lead to an over-estimation of the true value and a lack of consideration for overfitting. Studies in correlating FID with human visual evaluation have also returned mixed results, with some suggesting that FID is consistent with human perception and judgement as an indicator of visual quality [89] and others finding that it does not highly correlate with humans [65].

Kernel Inception Distance (KID) [88] is a similar Inception based method which

has been cited more frequently in more recent GAN papers [39]. KID calculates the squared maximum mean discrepancy (MMD) for a given kernel function. MMD acts as a two-sample hypothesis test and measures the dissimilarity between two probability distributions using independent samples. KID measures the MMD between the extracted Inception features from the real and fake images to produce a single value score. Like IS and FID (Fréchet Inception Distance), a low score indicates closer distributions between the target and test data, and thus compared to FID, KID has been found to be computationally faster, needs fewer samples to provide an accurate score. KID is also more robust against instances of overfitting to the training data, where the model simply generates copies of samples from the input data [88].

Both FID and KID are calculated using the feature space from a pretrained Inception network, often using the specific weights from a 2015 model trained on ImageNet and is included in the most popular deep learning libraries[48, 92]. This inclusion makes model evaluation much easier and faster than training an Inception Network from scratch and also allows for more consistent comparisons across the literature, however, they also raise several concerns. Recent work into FID has criticised the reliance on using a model trained on ImageNet for evaluating datasets which may not be covered by ImageNet's class label. The authors of one paper [93], investigated the role that ImageNet classes can effect FID scores for generated faces, an image type not covered by the ImageNet dataset. They found that they were able to iteratively optimise the dataset of generated face images to produce a much lower FID score without making any changes to the generative model itself. In addition to highlighting how FID can be easily manipulated to produce better, cherry-picked scores these experiments show how using an evaluation model based on one specific dataset (ImageNet) may not always be appropriate as a universal evaluation metric for generated data.

FID and other inception metrics may be the most widely used GAN specific evaluation metrics but the more traditional information theoretic metrics, Precision, and Recall, can also be adapted to be used to access image generation[94]. For image

generation, the precision value is related to the proportion of the generated distribution Q that matches the true data P, and the recall value is the distribution P that can be reconstructed back from the generated distribution Q [95]. These metrics were proposed as an alternative to FID based metrics as they better evaluate the extent to which the entirety of the true data distribution has been captured by the generated data distribution. An example which is given by Sajjafi et al., 2018[94] is that a model trained on a mixed gender face dataset but only produces high quality male faces may have the same FID score as a model which produces blurry faces but of both genders. In the case of the high quality model, applying precision and recall metrics would show that the reason for having the same FID as the lower quality model is that despite producing high quality images it has a low recall as it does not fully capture the distribution of the true data.

2.6.2 Human Evaluation Metrics

Human evaluation remains a valid method of assessing image quality from generated models, if only because image quality can be subjective when it is defined by human visual perception. Despite being hard to define, several different approaches have been proposed to investigate the human evaluation of image quality using empirical methods with psychophysics design paradigms from cognitive psychology [96–98]. Human eye Perceptual Evaluation (HYPE) [65] is a human-in-the-loop evaluation method which approximates image realism for a generated sample based on human accuracy in a real/fake detection task. The work in the paper state that HYPE achieves a 66% accuracy rate for predicting human judgment on the same generated images. Human evaluation methods such as these are useful for more accurately evaluating image realism and image quality than automated metrics but are much more time and resource intensive, requiring large human participant studies and are unsuitable for tasks such as monitoring model performance during training.

Other techniques draw from cognitive neuroscience, such as the evaluation method Neuroscore [99]. This evaluation method uses brain signals of participants to mea-

sure their visual perception of generated images that are presented to them in rapid succession. Compared to other evaluation metrics, the authors state that this method is more consistent with human judgement and needs a smaller number of samples. Additionally, a CNN based brain-inspired network is proposed that can approximate a given images Neuroscore. The use of cognitive neuroscience techniques makes this an interesting and novel evaluation metric. However, one drawback from using such methods is the significant cost and time required to conduct brain imaging studies.

The human evaluation methods chosen in this thesis are most similar to the cognitive psychology approach of using psychophysics designs such as real/fake image detection tasks. The reason for picking this set of methods is that they are simple to implement and analyse. The data can then be evaluated with metrics such as detection accuracy for experiments where participants are tasked with distinguishing between a set of real and fake images. This method also allows more flexibility in the data collection stage as studies can be adapted to be completed online in cases were it is not possible to conduct laboratory based work (e.g. COVID-19 restrictions).

An additional research method applied in this thesis is eye tracking. Eye tracking metrics are a set of human evaluation tools that can be used in conjunction with real/fake image detection studies, although much harder to reliably implement remotely. Eye tracking has been used in previous work[100, 101] to explore the regions of interest (ROI) of human gaze towards visual stimuli in order to infer differences in visual attention between different groups of participants and stimuli. In the context of the work presented in this thesis, the use of eye tracking can give insight into what ROIs are important for detection for generated EO imagery and how that may change between different experience groups.

2.6.3 Datasets

The datasets used throughout the research for this thesis contain RGB satellite aerial imagery from various urban and rural areas around the globe that have been

created for remote sensing applications. The dataset used most extensively is the INRIA Aerial Image Labelling dataset[102]. This dataset consists of 180 high resolution (5000x5000p) images of urban settlements in North America and Europe. The images cover 810km^2 of land with a spatial resolution of 0.3m per pixel (Figure 2.3). This consistent spatial resolution of 0.3m and that the images are already orthorectified is useful for comparing results using this dataset across different experimental setups. The images are also of a high resolution for satellite images that are often greater than 0.3m resolution. Additionally, this dataset was chosen due to its variety in environmental features ranging from small alpine towns to densely populated cities like Chicago and San Francisco. The dataset was obtained with permission from the owners.

Figure 2.3: *Samples of the INRIA Aerial Image Labelling dataset. The images presented have been segmented into 512x512 tiles*

The OpenCities dataset[103] is another RGB aerial imagery dataset. This contains drone imagery taken across 10 different regions of Africa (Figure 2.4). Unlike the INRIA dataset, OpenCities contains aerial imagery taken from different heights with varying levels of spatial resolution (0.02-0.3m per pixel) and inconsistent aerial surveying conditions. This dataset has been used as a secondary dataset for the research in this thesis as it is different enough from the INRIA dataset to be used to test generalization outside of INRIA. The differences in aerial height and spatial resolution within the OpenCities dataset makes it a good evaluation tool for testing

Figure 2.4: *Samples of the OpenCities African regions dataset. The images presented have been segmented into 512x512 tiles*

the generalisability of systems trained on the INRIA dataset towards other EO imagery.

2.7 Summary

Image generation is an area of computer vision research that has rapidly advanced and matured into its own field since the original implementation of GANs in 2014. In the following years, there has been many different improvements to the original GAN model and also a wide range of competing architectures such as transformers and diffusion models. Although the field is evolving quickly GANs are still one of the most popular architectures to use for image generation tasks and are now capable of producing highly realistic images from various types of image data.

For the detection of machine learning generated imagery, the current literature provides examples of several different approaches, each with different strengths and limitations. Despite the existence of such models, the work on detection is greatly outweighed by the work into novel generation models themselves, presenting a problem for data security as the improvements to generation methods outpace the advancement of novel detection techniques.

EO imagery such as RGB satellite data is a crucial data source in a variety of

technical fields. Despite its significance there is comparatively little research into investigating the generation and detection of synthetic EO image data. This leaves the field potentially vulnerable to malicious misinformation attacks.

Chapter 3

GAN Benchmark Comparison

3.1 Introduction

As previously discussed in Chapter 2, one of the issues prevalent in the field of GAN image generation is the lack of domain specific testing on image data such as EO images. To gain insight into the capabilities of the SotA GAN models towards EO images a benchmark study was conducted which measured three of the highest performing unconditional image synthesis models in addition to a basic DCGAN to function as a baseline. The aim of this benchmark comparison study in the wider PhD project is to provide a systematic evaluation of recent, state-of-the-art unconditional GAN models for the task of generating synthetic aerial imagery, providing a foundation specific knowledge to inform further research. Although other generation models exist and can offer comparable results in many instances to that of GANs (2), the focus of this PhD is on the detection and generation of GAN specific EO imagery, thus only GAN models were included in this benchmark analysis.

The main contributions in this chapter is the evaluation of current GAN models and metrics in the context of EO imagery. A secondary contribution is the generation of synthetic EO image datasets for future research into synthetic image detection methods.

3.2 Methodology

3.2.1 Unconditional GAN Models

Unconditional GANs do not require labels and learn unsupervised. This chapter focuses on the task of unsupervised image generation and an evaluation of the abilities of recent unconditional GAN models in the context of aerial imagery. GAN models share the same underlying principles of adversarial training, comprising two opposing deep neural networks: a generator and a discriminator [8]. The generator network G generates "fake" images by up sampling a random noise vector z. The produced image from G is then passed on to the discriminator network D. D is then tasked with classifying the given image $G(z)$ as real or fake, based on the distribution of the training dataset. The result is then used to optimize both networks simultaneously.

Progressive Growing GAN (PGGAN)

PGGAN is a popular unconditional GAN model for image synthesis and is one of the first GAN models to consistently produce high quality images at a resolution (1024×1024) (Karras et al., 2018). The images of faces generated by the authors of the model were also widely reported in the media (Vincent, 2017), giving exposure to the technology, and stimulating discussion on the potential implications this technology could have.

PGGAN achieves high-resolution and high-quality images primarily through the use of a progressive growing feature during training [17]. When training is initiated, the network starts with an input of low-resolution images (e.g. 4x4 pixels) and then gradually increases the resolution until it reaches its desired output size (e.g. 1024x1024 pixels). The model learns to produce the distribution of the images at a low resolution and then refine its parameters for increasingly finer details when moving to the next resolution. Other additions to the standard GAN architecture include pixel normalization in place of batch normalization, minibatch standard deviation and a Wasserstein loss function.

The Wasserstein loss function [22] is used as it provides more stability than the original, cross-entropy based minimax loss [8] becoming the new standard loss function for GANs. This function calculates the Wasserstein-1 distance (earth mover distance) between two probability distributions and is commonly used in many recent GAN models including StyleGAN2 and CoCoGAN. Instead of a minimax method of using the discriminator to predict the probability of a given image of being real or fake, this new loss function replaces the discriminator with a "critic" which instead gives a numerical score on the how real or fake the given sample is. The generators' goal is to then use this score to minimize the distance between the training data distribution and the distribution of the generated samples. The Wasserstein loss function benefits from a decreased chance of mode collapse and avoiding problems caused by vanishing gradients when compared the minimax loss function as the model is still able to learn independent of the generator's performance. This combined with the other architecture improvements gave PGGAN increased stability, training speeds and lower computational costs over that of previous models.

The original paper for the model presented state of the art Inception score results on a variety of commonly used object focused benchmark datasets (CIFAR10 [104]), LSUN ([105]) and CelebA-HQ[17]). These datasets contain a wide array of images. While the images generated in the model are almost photorealistic, they do contain some noticeable image artefacts which are most obvious in the backgrounds.

Since its release in 2017, PGGANs have been applied to a variety of different tasks in various areas of research. The main use for this model has been to generate faces, often for the purpose of testing fake image detection methods[43, 106, 107]. These studies cover different approaches to trying to detect GAN generated images, using PGGAN and other unconditional models trained on benchmark datasets to test their detection methods. As well as being used in image generation, the key progressive growing architecture that underlies the model has also been successfully adapted for data types, such as music and 3D MR images of brain volumes[108].

When trying to distinguish whether an image is real or generated from PGGAN the most telling sign is the incomprehensible backgrounds behind the focal image

object. This is particularly noticable in the face images from the PGGAN paper, as the model was trained on portrait style face images, the generated faces appear realistic but the backgrounds lack cohesiveness and continuity. For the task of generating convincing aerial imagery this could be a hindrance as it suggests the model struggles with generating cohesive images when there is no single focus (e.g., faces and objects).

StyleGAN and StyleGAN2

The original StyleGAN model is an update from PGGAN that enabled the model to learn unsupervised separation of image attributes. This led to the network being able to have more control over the image output[25]. In addition to control over different "visual styles," the model also achieved new state of the art Inception scores on benchmark datasets (CelebA HQ[17], FFHQ[1]).

StyleGAN has been used in many of the same areas as PGGAN, including detecting GAN fingerprints[52, 109] and face generation[110]. The main difference between PGGAN and StyleGAN is the latter's ability to learn conditional data in addition to unconditional data, leading to it be used for a larger number of tasks such as style transfer and image editing[111–113]. StyleGAN has been trained to produce diverse types of images, with the most common test datasets being faces (FFHQ, CelebA[114]), as well as commonly used datasets containing different object categories (LSUN[105]).

StyleGAN2 is the latest iteration to be released[1]. It includes significant changes in the architecture which allow it to obtain state-of-the-art performance in image generation and style transfer tasks. StyleGAN2 removes the progressive growing structure found in the previous models as the authors stated while this method was able to produce high resolution images it was prone to causing image artefacts in the output images.

Another problem observed with the progressive growing technique was that when trained on objects such as faces there is a strong location preference for certain features such as the nose and eyes which leads to lower variation in the final output

images. StyleGAN2 overcomes this problem by replacing the progressive growing with skip connections between layers after seeing a similar structure utilized by MSG-GAN [115]. Like in residual neural networks [57], the skip connections in StyleGAN2 connect lower feature maps in the network directly to the final generated output. This gives a similar advantage in training as previously given by progressive growing, which is for the generator to initially focus on low-resolution images and then shift its focus to changing the finer details of the image at the latter stages of training.

StyleGAN2 also made other changes to the network which contributed to improvements in performance. The model introduces a new normalization method to the loss, path length regularization. This results in smoother interpolations in the final image when the latent vector z is changed. The authors note that smoother interpolations correlate with increased stability and consistency in shapes, leading to improved visual quality. An adaptive instance normalization layer [116] is also incorporated which increases diversity in the final samples and avoids "water drop" image artefacts that were found in ProGAN and StyleGAN samples.

Like previous models, StyleGAN2 has so far been used primarily for generating realistic faces and other object categories [117] and has not been applied to the generation of aerial images. In comparison with previous versions (PGGAN, StyleGAN), StyleGAN2 can generate more photo realistic and varied images that lack GAN image artefacts that were present in past models. With the updates in architecture bringing increased visual performance and output images which are more coherent across the entire image and not just on the main object, StyleGAN2 is one of the more suitable models for aerial image generation.

One potential flaw of StyleGAN2 that is noted by the authors is that despite generating images with higher levels of photo realism than previous models (PGGAN, StyleGAN)[25], these images are easier to detect as synthetic by image classifiers trained to distinguish between real and GAN images. This discrepancy between perceived visual quality and performance against image classifiers suggests that there are inherent differences towards image realism and detection strategies between hu-

mans and deep learning-based models. Additionally, while StyleGAN2 has been presented as an improvement on PGGAN in terms of generated image quality, the model is also much larger and more computationally expensive to train. This leads to questions on how beneficial the improved performance is for using StyleGAN2 over the smaller PGGAN when considering possible limitations of computational resources available.

CoCoGAN

Conditional Coordinate GAN (CoCoGAN) presents another novel GAN architecture with results that rival other high performing GANs (StyleGAN2, BigGAN)[118]. Inspired from the human visual system's ability to perceive an entire visual scene from eyesight despite the limitations of eyesight to only be able to look at a part of that scene at any given point in time, CoCoGAN generates high resolution, photo realistic images in parts by using the spatial coordinates of each part as a condition during training. The authors also put forward the model to be used for the novel task of "beyond boundary generation". This is when the model is asked to extrapolate the image beyond the range that it has been trained on, generating output images that are larger than those in the training set and guaranteed to be novel, as they are not directly based on any real data.

CoCoGAN has been tested on a number of different datasets including object datasets such as CelebA and the LSUN256 dataset. As well as these standard benchmark datasets, the model was also able to achieve low FID scores for the panorama dataset Matterport3D[119]. This presents a different challenge for image generation than compared to object focused datasets, as it requires the model to learn how to create a coherent image with decentralised features, much like those seen in aerial images.

BigGAN and BigBiGAN

BigGAN is another recent GAN model which is capable of conditional and unconditional high resolution image generation[2]. As the name suggests, BigGAN is a

large-scale GAN model, trained using four times the number of parameters and eight times the batch size compared with prior models. The authors report that their results benefited greatly from upscaling the architecture. BigGAN managed to achieve similar levels of visual fidelity at high resolutions as PGGAN and StyleGAN on object and category datasets like ImageNet. BigBiGAN builds on the BigGAN model architecture with a series of updates to achieve greater variety and photo realism in the generated images[120].

Despite being a recent model with competitive performance in generating realistic images[110, 120], neither BigGAN nor BigBiGAN will be included in this survey. These models are much larger than the other models included in this comparison, with 340 million training parameters compared to 58 million in the largest network included in this study (StyleGAN2). To train these models from scratch at the resolution of 256×256 requires over 12 GB of video memory, which exceeds the resources expected to be available for researchers. The results presented in this chapter focus on applicable techniques for the wider community, using models that are able to be trained on consumer level GPUs with 8-12GB of video memory.

Baseline DCGAN

In addition to testing these current, high performing GAN models, a basic deep convolutional generative adversarial network (DCGAN) is also included for evaluation (Figure 3.1. This serves as a baseline model for this study. It was chosen as it is still able to produce 256×256-pixel images, but lacks any of the innovations and updates in structure found in more recent models. Using the original DCGAN base architecture described by Radford, 2015[121]. Iterating on Goodfellow's original GAN [8] by adding convolutions, this model achieved impressive results at its time of release when deep learning image generation was still a novel concept. Due to being a much smaller model than any of the current high performing models, DCGAN is considerably quicker to train with much lower computational requirements.

The specification for this implementation of a DCGAN uses simple convolutional neural networks for its adversarial training. The generator network G up samples

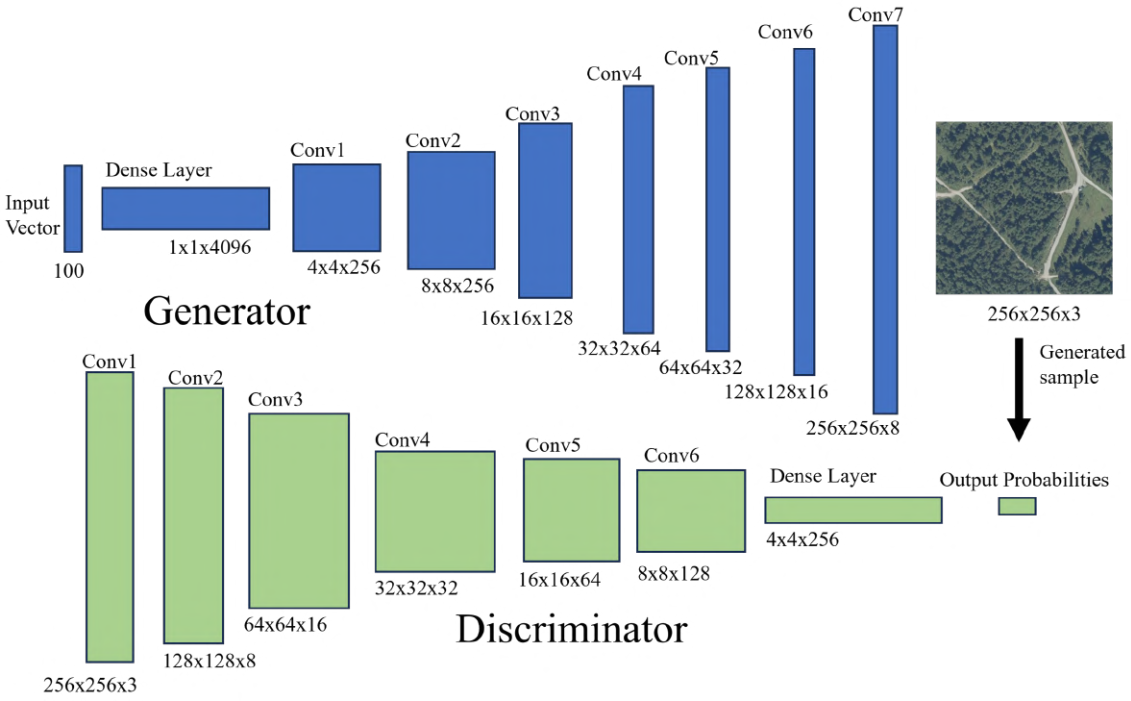

Figure 3.1: *Baseline DCGAN architecture. The generator (top) and discriminator, D (bottom) consist of a series of convolutional layers (Conv) and dense layers for reshaping the data.*

the random noise vector z though 7 convolutional blocks to produce an image that is given to the discriminator D. G uses Relu as its activation function for each layer, apart from the output layer which uses a Tanh function. D is a CNN classifier made up of 6 convolutional layers (including input layer) and a single dense layer. D uses average pooling between each convolutional block and a leaky Relu activation function, as this was found to work better for higher resolution images than the standard ReLu [121]. A Minimax loss function is used for training both networks.

3.2.2 Experimental Setup

The image generation performance was tested for the PGGAN, StyleGAN2 and CoCoGAN networks and, for comparison, also evaluated a baseline DCGAN. In all experiments, the official implementations of the networks were used. These networks were chosen due to their reported state of the art unconditional performance on common GAN benchmark datasets. PGGAN was also selected for comparison as, despite being superseded by StyleGAN2, it remains one of the best performing and most

used GAN architectures for the task of unconditional image synthesis. As mentioned above, despite achieving similar performance, BigGAN has not been included due to the large VRAM requirement during training for resolution of 256×256-pixel images (>12 GB GPU Memory). Other unconditional GAN models (e.g. FineGAN[122], AutoGAN[123], SRNGAN[124]) were not included as they could not scale to the target resolution, while others did not have official code repositories at the time of research.

Each model was trained with the hyper parameters specified in their official implementations. All the models were trained using two GPUs with 8GB of VRAM, aside from the baseline DCGAN which required only one GPU. Each model was trained until a short time after model convergence was apparent, with no further decrease in the model's training loss. For each trained model, a sample dataset of synthesised images was generated to be used for evaluation. Each test generated dataset was of the same size as the training set, and each model was evaluated by comparing 10 random subsets of 10,000 generated images with 10 random subsets of real images of the same number. These comparisons measured the mean and standard deviation for the metrics Fréchet Inception Distance and Kernel Inception Distance for each model.

3.2.3 Dataset

Most of the published research in the area of deep learning methods with Earth Observation data use Map2Sat as a baseline dataset[43]. This dataset is mostly used due to its ease of access: it is included in TensorFlow 2.0. Map2Sat was created for the purpose of demonstrating the performance of CycleGAN[21], and contains 2,000 satellite images with road data extracted from Google Earth. While the image pairs are useful for style transfer tasks, its relatively small size, lack of diversity and low resolution (max 256x256) make it unsuitable for the unconditional generation task addressed here.

In this chapter the models are evaluated using the INRIA Aerial Imagery Dataset[102],

which contains a large number of high-definition images of varied environments. The INRIA dataset contains open access, high resolution aerial images in GeoTIFF format. Originally created for building detection, it covers 810 km2 and is comprised of aerial orthorectified colour imagery at a spatial resolution of 0.3 m per pixel. It includes images from urban settlements from a wide range of geographic locations and with a wide range of characteristics, from densely populated areas such as San Francisco, to alpine towns in Austria. The variety of images offered in this dataset make it an ideal target to evaluate GAN-based EO image synthesis.

3.2.4 Data pre-processing

The original version of the Inria dataset includes 180 colour tiles of 5000x5000 pixels covering a surface of 1500 m x 1500 m. These tiles were then resized to 4096x4096 and each split into 8 tiles of 512x512. Fig. 1 shows a random sample of images extracted from the dataset after being split and resized to 256×256. There was approximately a 50/50 split in terms of rural and urban features present in the final tiles.

Figure 3.2: *Random selection of images from the INRIA Aerial Imagery Benchmark Dataset [102].*

Data augmentation techniques have been successfully applied in deep learning problems to improve performance[125]. In this instance both vertical and horizontal flipping transformations were applied to the original dataset[126]. A mirrored, duplicate dataset was added to the training set, and all of the images were also rotated

by 180°and 90°.

After data augmentation, the dataset is comprised of 34,600 256×256 images. This resolution was selected because it was the highest resolution shared by each of the tested models. In addition to the previous augmentations, a sliding window was used to create more tiles to further increase the dataset size.

StyleGAN2 was selected for additional evaluation as it is the most widely used out of all the benchmark models and was built specifically for the generation of high-resolution images. For this task, an additional, higher resolution dataset was created. This dataset also used the Inria dataset, but was made up of larger tiles of the original 4000x4000 images than the 256x256 dataset used for the main benchmark evaluation. This high resolution dataset contains 16,500 images at a resolution of 1024×1024 pixels.

3.2.5 Metrics

To assess the performance of each model Fréchet Inception Distance [90] (FID) and the Kernel Inception Distance [88] (KID) were applied. The Fréchet Inception Distance has become one of the most widely used metrics [127] for evaluating the performance of GAN models. Its purpose is to measure the statistical similarities between the original data and the generated data. A lower FID indicates that the two groups of samples are more similar, with a score of 0 indicating both groups are identical.

This metric is measured by embedding a set of generated samples into the feature space of a specific convolutional layer of the Inception CNN model[128]. Then, the distance between the mean and co-variance of each group (real and fake images) is calculated. The Fréchet distance between the two Gaussians is then used as a quantitative measure for visual quality of the generated samples. It is given by 3.1:

$$FID(r,g) = \|\mu_r - \mu_g\|_2^2 + Tr\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}}\right) \tag{3.1}$$

Where μ_r,Σ_r are the mean and co-variance of the real data distribution, with μ_g,Σ_g

being that for the generated data distribution. The euclidean distance is measured between the two means ($\|\mu_r - \mu_g\|^2$) and Tr is the trace operator that sums the diagonal elements of the matrix. FID superseded the previous GAN evaluation standard, Inception Score [98] as it has been shown to be a more robust measurement of image quality [89]. The Inception score has also been shown to inadequately detect overfitting [91], as it only uses the generated samples while ignoring the training set, and is also sensitive to image resolution [90].

Kernel Inception Distance [88] measures the maximum mean discrepancy (MMD) between two probability distributions (P_r and P_g) for some fixed characteristic kernel function k. MMD is a two-sample testing measure that computes the dissimilarity between P_r and P_g using independent samples from each. This metric has been found to be more sensitive to overfitting than FID scores, although as with FID due to sampling variance $M(X, Y)$ it may not be 0 even when $P_r = P_g$ [129]. It is calculated as shown in equation in 3.2:

$$M_k(P_r, P_g) = \mathbb{E}_{x,x' \sim P_r}[k(x, x')] - 2\mathbb{E}_{x \sim P_r, y \sim P_g}[k(x, y)] + \mathbb{E}_{y,y' \sim P_g}[k(y, y')] \tag{3.2}$$

Where k is a fixed kernel function (e.g. Polynomial Kernel $k(x, y) = (\frac{1}{d}x^T y + 1)^3$, with d being the dimension of the Inception representation) and (x,y) refer to the real and generated sample. KID has been found to converge to its true value faster than FID [88], also requiring less n samples.

One problem which is present in all GAN evaluation metrics is that they try and quantify the very subjective factor of image realism, something which is often measured using human evaluation measures [63, 130]. FID has been to found correlate with measures of human perception towards assessing image realism and image quality in GAN samples [89]. This suggests that FID and its iteration KID can be used as metrics for the quantitative analysis of GAN image quality. However no studies have been performed evaluating the fitness of these functions for EO image synthesis. FID and KID suitability as quality metrics has been questioned in the results from

Zhou (2019) [65]. When measuring GAN face generation (CelebA, FFHQ, Cifar-10) using their own human perception metrics (HYPE) they found that there was no significant correlation between humans and the automated metrics. It is important also to note that these Inception score based metrics are derived using a pretrained network which was is trained on Imagenet [87]. For the main evaluation of GAN models we use the pretrained Inception model that is the standard practice in many GAN papers ([1],[120]).

As EO data comprises different features beyond common images with objects usually centred, the use of an Inception model trained on ImageNet does raise additional questions on the reliability of these benchmark metrics beyond ImageNet models [91, 93]. In addition to the standard Inception Network we also present FID scores using an instance of this Inception model after fine-tuning on a section of the Open Cities Dataset [131], which consists of high quality urban and rural images of African cities.

3.3 Results

FID and KID were compared for each model across 10 k-fold random subsets of 10,000 generated images, with the average FID/KID being reported. Each subset was compared against the same number of randomly selected real images, the mean and standard deviation was then recorded for each model. The results of the comparison can be found in Table 2. Additional sample images from each model can be found in the appendices 7.4.

3.3.1 Model Comparison

When comparing performance across models, it is first important to note that comparing FID scores between papers can be difficult due to the FID's sensitivity towards the number of test samples [88], meaning that FID can only be fairly compared in tests with an equal n value (as FID is a measurement between two distributions), hence an equal number of images were generated for each model.

Table 3.2 shows the FID and KID scores for the various models trained on the INRIA dataset, a random selection of generated samples visual image quality can be seen in Figure 3.7, with additional images found in the appendices.

Model	**Number of Trainable Parameters**	**Training Time**
DCGAN	(920K generator, 1.8m discriminator)	2 days
PGGAN	(23m generator, 23m discriminator)	4 days
StyleGAN2(256)	(30m generator, 28m discriminator)	10 days
CoCoGAN	(24m generator, 29m discriminator)	7 days
StyleGAN2(1024)	(32m generator, 30m discriminator)	13 days

Table 3.1: Number of trainable parameters and training time for each tested network.

Model	**FID (Mean ± SD)**	**KID (Mean ± SD)**
DCGAN	283.72 ± 1.32	312.32 ± 2.21
PGGAN	27.24 ± 0.30	12.45 ± 0.63
StyleGAN2	**16.59 ± 0.18**	**7.28 ± 0.61**
CoCoGAN	141.10 ± 0.56	104.39 ± 0.09

Table 3.2: Metrics for Baseline and state-of-the-art models. Best performance is shown in bold (lower values are better).

StyleGAN2 was found to produce the highest quality images, both in terms of metrics, as shown in Table 3.2, and in terms of visual results in Figure 3.7. The model did require the longest training time out of the tested models and has the highest minimum requirements for GPU memory (8GB).

The generated samples were more detailed than those from the other networks tested, with the updates in architecture from previous iterations (PGGAN) improving its general generation ability beyond the face datasets (CelebHQ) tested by its original authors. Whilst the progressive growing architecture of PGGAN is still able to produce fairly realistic looking images, with image details such as cars, trees and buildings, the change to using blocks with skip connections in StyleGAN2 makes these details much clearer. StyleGAN2 improvements such as the multi-resolution training, where all image resolutions are trained at once, to help in learning more stable and coherent representations across different scales, makes for more consistent and realistic images.

Despite the generated EO images being of poorer quality than the examples of faces and objects in the original paper [1], they demonstrate that aerial images

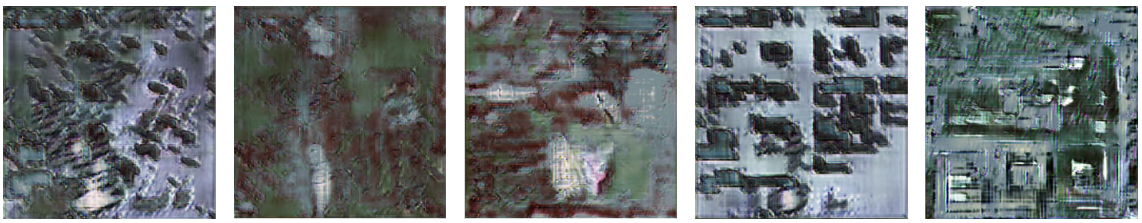

Figure 3.3: *a) Baseline GAN results: The DCGAN has managed to learn the basic shapes and features of the training data but struggles to correctly generate finer details. Convolutional upsampling crosshatch artifacts are also present*

Figure 3.4: *b) PGGAN results: The PGGAN images are quite detailed, baring close resemblance to the real images but still contain obvious deformities at object boundaries (e.g. roads, building edges)*

Figure 3.5: *c) StyleGAN2 results: The generated images bare a strong resemblance to the training data, with improvements in quality over PGGAN and less obvious image artifacts*

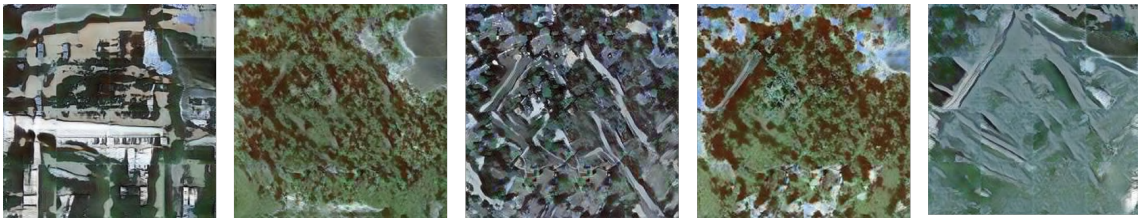

Figure 3.6: *d) CoCoGAN results: CoCoGAN struggles to generate convincing images with large warping artifacts for straight line features (roads, buildings)*

Figure 3.7: *Random selection of results from trained GANs*

can also be successfully generated with high levels of photo realism. The warping artefacts we notice in other approaches are much less pronounced in these images, with roads and roofs being realistically rendered. Overall the images are also much more detailed and clearer, giving them a much more photo realistic look. The model manages to render both rural and urban features reasonably well, although there are more noticeable artifacts in the urban imagery. This is perhaps simply due to

the additional challenge of rendering buildings rather than foliage. Warping around straight edges can be seen in other instances of GAN image generation [58] as the models struggle on replicating hard boundaries between images. The abundance of such features in urban images may explain the differences in visual quality between the generated urban/rural scenes.

The results from PGGAN showed a noticeable dip in visual quality in comparison to the more recent StyleGAN2, but still achieve fairly realistic looking images. The images produced were visually better than those by the baseline DCGAN which was also reflected in the FID and KID metrics. Results look visually more "realistic", as shown in Appendix Figure B.3, with details such as trees and houses present. There are, however, some noticeable artifacts such as warping issues that can be seen where roads and rooftops which should appear uniform and straight do not. The warping issues are most present in the images that depict more urban area and are less noticeable in images with higher amounts of foliage. PGGAN's performance is especially interesting when compared with CoCoGAN. Although CoCoGAN is a more recent architecture, PGGAN outperformed CocoGAN in both metrics and visual fidelity. This suggests PGGAN has a much more robust architecture, better suited to generalisation beyond face and object synthesis. Although PGGAN did not produce images to the same quality of StyleGAN2, PGGAN took considerably less time to train to convergence when trained on the same hardware.

The FID scores presented in the original paper [118] suggested that CoCoGAN would outperform the other networks. However, CoCoGAN produced less visually realistic images, and lower metric scores. FID results dropped by 131.5 between the CelebHQ dataset, which was reportedly used in [118], and the INRIA Dataset. This is a surprising result, as the network has been reported as outperforming other networks on high resolution, non-object focused datasets such as the Matterport 3D panorama dataset [118]. We trained with a reduced batch size of 64, compared to the original 128, due to memory constraints on the large INRIA dataset, but this is unlikely to have caused a notable drop in quality.

Similar to the PGGAN results, the images from CoCoGAN managed to capture

the more basic geometry and colours in the image, but without the detail and clarity of those from StyleGAN2. CoCoGAN struggled with generating convincing urban environments with some images being incoherent. The most prominent image artifact found in the generated CoCoGAN images was a visible grid pattern of the seams between the different macro patches. This grid pattern could be an indication that the model has failed to learn the distribution of features in EO data, causing difficulty in its attempts to merge micro-patches. In addition to generator and discriminator loss functions, CoCoGAN also has a spatial consistency loss. This is used to govern the two spatial coordinate systems (micro and macro) for determining how the individual patches of the image are generated and put together. These grid artefacts suggest that the network has not been fully optimised towards the spatial consistency loss (Figure 3.7). In the original paper this was noted as being a problem that was possible, and this is particularly noticeable in our experiments. In less cluttered images of fields and vegetation the effect is most pronounced.

A limited hyper parameter search (patch size, learning rate, gradient penalty weight) was conducted to try to improve image quality but saw negligible results. Further tuning of the hyper-parameters could potentially improve image quality but is computationally costly. A more likely solution would be to use more data, to better accommodate the optimisation of the spatial consistency systems,. A larger dataset may be, particularly given that the datasets in the original paper contained many more images. In this instance the experiment was limited by the VRAM available. Models such as StyleGAN2 and PGGAN, were shown to be much more reliable in this regard, making them better choices for further experiments.

As expected, DCGAN offered the lowest performance of all methods. It produced average FID and KID scores of 283.7 and 300 respectively. DCGAN is the simplest GAN network, not incorporating modern network design elements present in the other works. Its inclusion in this comparison is still useful in providing baseline results against which we can compare. Visually, as shown in Figure B.2 in Appendix B, the limitations of the network are clear. The model has learned to capture the low level features in the training data, such as broad shapes and colours, but has

struggled to capture the finer details and textures. Earlier GAN models such as this one are known to struggle with producing realistic looking images at resolutions higher than that of toy datasets such as MNIST and ImageNet, producing unclear and blurry images [132]. Certainly, the differences in image realism between StyleGAN2 and DCGAN highlight just how rapidly automated image generation techniques have evolved in a short space of time. The DCGAN images do not contain the detail in the StyleGAN2 images which manage to replicate aerial image features to a much higher level of visual fidelity. Features in the more recent GANs such as the progressive growing in PGGAN and the skip connections in StyleGAN2, enable the GANs produce finer details, such as cars, whilst keeping the low level features and can do so at a higher resolution. The simplistic convolutional layers used in DCGAN are not able to produce these at the tested resolutions.

3.3.2 StyleGAN2 Latent Space Analysis

As the highest performing network, we performed an analysis of the embedded features in StyleGAN2's latent space. This can give us a further understanding to what extent the model has learnt the more uncommon features in the training dataset. We first generate an output image from the StyleGAN2 generator given a starting latent vector z [1]. The output images and a target real image are then both placed in a pretrained feature extractor (VGG16 [133]) which then computes the loss between the features of the images. Using gradient descent the loss is then used to optimize the latent space to generate an approximation of the target image. The latent space images show a noticeable disparity between the learned representations and the target images (Figure 3.8). While the model can approximate the general rural landscape from the target images, it has failed to replicate the building estates in the bottom two examples. The model can be seen to effectively replicate the more global features and repetitive patterns in the aerial images, such as type of terrain or vegetation, but struggles to add local features such as buildings and more fully completed roads and trees. If the model had managed to learn the datasets

Figure 3.8: *Four pairs of real (right) images and their associated latent-space image (left) from the trained StyleGAN2 1024 model. The latent space images are close to the training images, but have some differences in specific features. The model has learnt to generate the structures and features similar to those in the Inria dataset images with out overfitting to the training data. This would likely be the case if the latent images were exact replicas of the training images*

distribution more accurately then there would be fewer differences in the images, although this also suggest that the model has learnt image features rather than simply overfitting to the training data. Additionally, irregular and unique features such as landmarks specific to that are not produced in the generated samples as these represent anomalies in the data distribution that StyleGAN2 is attempting to mimic. In distinguishing between a well generated urban scene, looking for unique landmarks such as a sports stadium could help to quickly determine if the image is authentic or not.

3.3.3 FID Comparison between Inception Models

As previously discussed in section 2, one concern with Inception model-based metrics (e.g. FID, KID) is that the standard way of calculating the metric is to use a model pretrained on the ImageNet [87] dataset. Due to the large variety of image classes included in the ImageNet dataset, this single pretrained model can be useful for evaluation in many circumstances. This becomes a potential issue however when being used for evaluation of image data from domains not included in ImageNet

(e.g. EO imagery).

In the main evaluation section of the GAN models this instance of the Inception network was used, keeping in line with the standard practices in current GAN literature. To further explore the consequences of using an Inception network trained on a different image type (objects vs EO data) a second set of FID metrics were calculated using a ImageNet pretrained Inception Network fine-tuned on a different EO dataset (OpenCities [131]). The choice for using this dataset was that it was another high quality, large EO image dataset that contained different geographical rural and urban locations as those found in the Inria dataset. Although the Inria dataset could also have been used to train an Inception model for calculating FID, using an external dataset allows the results be used in future comparisons with different generative models or EO datasets.

To configure this additional Inception model (ImageNet+Maps) for FID evaluation, the output layer was modified from 1000 output class probabilities to 1001 (n =ImageNet classes + Map class). The classifier was then fine tuned on the task of classifying the OpenCities dataset into the new Maps class against the existing ImageNet classes. This was completed when the accuracy was approximately the same as the ImageNet classes (∼94.8%) .

These results can be seen in Table 3.3. The results show that FID changes significantly when using a model fine-tuned on EO data. The resulting scores being much lower, which indicate closer features found in the distributions between the real and generated datasets. The performance rankings between the different models stayed the same for both evaluation methods.

Model	**FID (ImageNet)**	**FID (ImageNet+Maps)**
DCGAN	283.45	193.45
PGGAN	25.51	0.90
StyleGAN2	**16.59**	**0.69**
CoCoGAN	136.35	27.35

Table 3.3: FID metrics for different Inception models (lower values are better).

3.4 Discussion

The aim of this chapter has been to evaluate the generation of Earth Observation (EO) data using current state-of-the-art GAN models. EO data presents a novel challenge since these models are usually fine-tuned towards the generation of objects and faces. The main motivation behind this work comes from the increasing prominence of sophisticated image-generation algorithms, the lack of current literature and scientific evidence towards their use for generating aerial image data, and the potential concerns associated with malicious use of image synthesis tools which could relate to fake information generation.

Results of this evaluation are both promising and concerning for those addressing the problem of fake EO image generation. When comparing the performance of all models together, all were found to perform worse quantitatively for the purpose of EO generation, than in results reported for their original implementations on other domains. While this comparison was not expected to find the same levels of state-of-the-art scores that were achieved on the various benchmark datasets there are substantial drops in performance (average 61% FID decrease) in this evaluation using EO imagery. One explanation that can be hypothesised is that the data these models were designed with were primarily face and object datasets (e.g. CelebA, FFHQ). These image types are quite different that aerial imagery in terms of the spatial relationships between features. Unlike the defining features in facial data (nose, mouth, and eyes), which are very central and have defined spatial relationships with each other, the features unique to aerial imagery (roads, foliage, buildings) are much more decentralized presenting a different spatial relationship of features.

The importance of differences between image types can be seen in the disparity of FID scores (Table 3.3). The lower scores for the EO trained Inception Model show that even without being trained from scratch, the addition of fine tuning on domain specific data forces the network to embed different predictive features than those in the standard ImageNet model. This has also been observed in previous experiments [134] of changing the underlying data the Inception network is trained on. Changing

this dataset causes FID to be effected by different features and biases in the data. This results in an FID metric which is more suited for bench-marking models when dealing with a specific image domain. Due to the bias that changes to the Inception dataset create, comparisons can only be made with FID results that use the same FID implementation.

Another contributing factor is that the dataset used to train the models was relatively low at 36.4K samples, in contrast, many commonly used datasets have well over 50k samples. This does, however, highlight the ability for some models such as PGGAN and StyleGAN2 to be able to generate realistic looking images from smaller training sets. This ability to perform well in terms of KID and FID with smaller datasets makes these models suitable for tasks where existing data for training is limited or unbalanced. As previously discussed in this chapter, this provides further evidence of the advantages of GANs as a data augmentation tool to extend training sets for classifiers which may need larger or more balanced training datasets.

StyleGAN2 produced visually impressive samples despite a smaller training set. This model is sufficiently capable on EO data to merit further research, both as a tool for generating training data for detection systems, and in assessing the level of threat that it poses towards current systems. This ability to generate data that could potentially fool detection systems, both automated and human presents an immediate concern, especially when this technology is developing at a rapid pace as seen in the improvements between current models (StyleGAN2, PGGAN) and ones from only a few years prior (DCGAN).

3.5 Conclusion

This chapter presented a thorough benchmark analysis of the generation of fake aerial imagery using unconditional adversarial networks. Results from this benchmark study show that SotA GAN models, such as StyleGAN2, can successfully generate examples of EO imagery although replication of the low FID scores presented in their accompanying papers is difficult to replicate. Furthermore, the ex-

ploration of different implementations of FID based metrics lend further evidence to the argument that these metrics are not consistent enough to be employed for empirical evaluation of GAN performance. The FID results (Table 3.3) obtained from the OpenCities Inception Model show how these scores are subjective to the types of image data being used and making score comparisons between GAN papers unreliable. The differences we found in FID when using a different dataset for the Inception model support the concerns that these are flawed metrics for measuring GAN image quality, especially the industry standard to rely on Imagenet based models regardless of the generated image type[93].

The accurate generation of Earth Observation data such as photo-realistic aerial imagery presents concerning implications regarding the security and validity of digital imagery. The ability to rapidly generate enormous quantities of false information gives Security and Defence research a unique challenge to tackle. The work in this chapter has explored EO image generation using mathematical GAN evaluation metrics as well as the creation of GAN generated EO datasets to use for future work. Following on from the results and conclusions drawn in this chapter the research in the following chapters will evaluate the best performing model (StyleGAN2) using human evaluation methods. As GAN images are often generated for the purposes of being viewed by humans, particularly in terms of deception, it is important to examine just how realistic they appear using human evaluation methods, rather than automated metrics such as FID and KID.

Chapter 4

GAN Image Detection

4.1 Introduction

The GAN benchmark study in the previous chapter found that StyleGAN2 achieved the highest performance in the unconditional generation of synthetic aerial imagery when measured using FID and KID. Using samples generated from this model, this chapter explores human detection performance on synthetic aerial imagery and the differences with automated methods. As discussed in the literature review section on human detection (Chapter 2), it is important to understand this aspect of fake image generation in addition to automated detection methods as humans are often the target viewer for fake photo realistic imagery, malicious or otherwise.

This chapter argues that research into the security and authenticity of image data, given the rise of machine learning driven generation algorithms, requires a balanced understanding of human and automated generation methods. Evidence is given to support this by showing how both humans and pretrained CNN detection models achieve sub optimal performance in differentiating between real and synthetic EO imagery. This is presented in the form of a comparison between a CNN detection model and the results of an online human detection study.

The research presented here aims to address how hard it is for humans to identify between real and synthetic aerial imagery and whether this is correlated with previous experience in using similar data (e.g. satellite imagery).

The results find that previous experience (self-reported) positively correlates with task performance, this indicates that self-perception of expertise in this context can be used to reliably group participants when using expert/novice experimental designs.

Additionally, this chapter also examines the current Inception model-based evaluation metrics and their differences with human evaluation in the context of synthetic aerial imagery.

4.2 Experiments

4.2.1 User Study

To investigate human detection towards synthetic aerial imagery a user study was conducted online using images from the Inria aerial imagery dataset and samples generated by StyleGAN2, the best performing model in the previous chapter.

Design

A within groups, 2-alternative forced choice design (2AFC) was used for the study. In this experimental design, participants from independent groups are asked to select between two given stimuli. This methodology was selected for being a reputable and established experimental design ([135]) for decision based visual search studies. Although this study design does not indicate participant confidence, as the forced choice may result in some decisions being guesses, this design does negate the option of non-answers or neutral responses from participants.

The purpose of this study is to investigate if participants can distinguish between real and generated images as well as any correlations between experience levels and accuracy. Due to this study was kept as simple as possible, and additional measures for confidence were not included. Doing so, such as, asking participants to rate the confidence of there answers, may have reduced participant engagement, lowering the overall turn out for the study.

2AFC was chosen rather than a sequential design for displaying images, as it removes implicit biases towards the visual stimulus (real/generated images) that could impact decision making [135–137]. If a single image is presented to the participant which they are unsure of, they may be biased in one direction to chose a certain answer. As the participant has been briefed on the purpose and context of this study, they may be biased towards believing the image is fake, skewing the final results. Additionally, sequence of single images can also skew a participant's perception as they will be influenced most by the previous image they saw. This could cause their perception of real and generated to become more warped over the course of the study. 2AFC, with two images presented simultaneously tries to prevent this from occurring.

Figure 4.1: *Screenshot from the forced choice study. One image is real and the other is synthetic. Participants are asked to indicate which one is synthetic. In this example the left image (real) depicts an urban scene and the image on the right (synthetic) depicts a rural scene, urban/rural combinations are presented randomly.*

The experiment was created using PsychoPy3 [138], a Python3 software package for creating interactive cognitive psychology experiments. The study was hosted on the online psychology experiment platform, Pavlovia.org [139]. This was done as at the time of research, face to face studies were not an option due to restrictions in response to the COVID-19 pandemic. Although online studies are not able to be as rigorous and precisely monitored as in lab conditions, it did allow a larger sample of participants to be recruited in a relatively short amount of time. To dissuade repeated attempts, only one full completion of the task could be made before being

locked out via website cookies. Participation was open to all but the final population (N= 94) was generally made up of students, academics and data scientists based in the UK.

In an initial practice task, participants were given a set of image pairs (4.1) and for each pair asked to identify which image is fake. Each pair consisted of 1 StyleGAN2 generated image and 1 real INRIA satellite image. It was made clear to the participants that for these pairs there would always be one real and one fake image. The images from both sets (fake/real) were a mix of urban and rural images and presented at random. After answering they were then given feedback if they were right or wrong. At the start of the experiment participants were asked to give their level of previous experience (low, moderate or high) at looking at similar types of EO data and images. After this initial practice task (10 image pairs) was completed, the main task was given. This followed the same format as the practice task but without feedback and more image pairs (100 image pairs in blocks of 25). The image pairs were generated randomly for each trial from a dataset of 250 StyleGAN2 images and 250 INRIA images, all at a resolution of 256×256. These pairs consisted of random combinations of urban and rural images. Randomisation of the images ensures that the data obtained during the study is from the difficulty of the images rather than specific pairings. Although some participants may encounter more challenging pairs than others, this effect is minimised over multiple trials and participants.

Participants were given as much time as they would like to answer. Previous works ([65, 140]) have opted to implement time constraints for participants. The current experiment measures the participants' ability to distinguish between real and fake images regardless of a time taken. The only time variable that is controlled is that the images are shown for a minimum of 500ms before the participant is allowed to answer, this is to avoid mis-clicks or the participant making guesses without looking at the stimuli first.

Hypotheses

H_1 : Participants will not be able to consistently distinguish real EO images from generated images. This will be reflected in a low task accuracy (e.g., less than 75% accuracy average).

H_2 : Participants that self-perceive to have higher expertise will show higher accuracy than lower expertise.

4.2.2 CNN Detection Model

In addition to testing the generalization capabilities of generation models, using less common GAN datasets such as earth observation images can be used to evaluate current GAN detection models. Currently there are a number of different models for the purpose of detecting GAN generated images but many of these are limited to looking at specific image types such as faces. One recent paper by Wang et al.[45] (Previously discussed in Chapter 2) claims that GAN detection is currently a solved problem and presents a trained ResNet model which is capable of classifying GAN images with high accuracy across datasets generated by a variety of SotA GAN image synthesis models including ProGAN and StyleGAN2. While the authors show that the model can generalize to a variety of different image types by different models only common GAN benchmark datasets were used. This makes the results of paper hard to generalize to other unseen forms of synthesised data such as EO data. Despite this potential limitations of the model, further studies on GAN image detection have used this model as a method of evaluation[39] making it one of the most currently used methods of GAN detection that evaluate the visual differences between real and synthetic images.

For this study, the ResNet model from Wang et al.[110] was selected for evaluation, due to its reported high detection accuracy and use in other GAN papers. This model consists of a pretrained (ImageNet) ResNet-50 model that the authors had further trained on additional image data specifically for the purpose of being a generalized CNN generated image detector. The model was trained on real/fake,

object and image datasets generated by PGGAN, up to a resolution of 1024x1024. During training the authors used data augmentation techniques of adding Gaussian blur and jpeg image artefacts were applied to increase generalization ability and robustness to real life scenarios.

For the purpose of the current experiment, two variants of the model were used (trained by Wang et al.). One which was trained with each training image having a probability of JPEG compression and a Gaussian blur of 10% (jpeg and blur 0.1) and another at 50% (jpeg and blur 0.5). The purposes of this training augmentation is to test how ecologically valid the model would be when employed on common image types seen online. Both models were tested using 10000 images of each class (Inria real images, StyleGAN2 generated samples) and were from the same datasets as the images used in the human study.

4.3 Results

4.3.1 User Study Results

Image Type	Correct Response	FID	KID
All Images	68%	17.51	43.01
Urban	70%	17.48	43.65
Rural	66%	13.54	36.88

Table 4.1: Metrics table for urban and rural generated samples.

In this chapter a user detection study was carried out. The aim being to discern the extent to which users are fooled by state of the art, synthetic EO images, and the extent to which FID and KID are useful predictors of human performance on this task. It was found that participants (N=94) were able to correctly identify the fake image from each image pair shown on average 68% of the time, the distribution of user scores can be found in Figure 4.2. The results also showed that self-reported user experience did have a positive correlation with accuracy (Table 4.2) but no significant conclusions can be drawn from this due to the low sample size of users answering 'High' experience. The data was found to be non-parametric as it did

not follow as normal distribution, so a Kruskal-Wallis H test carried out. The test found $\varepsilon^2 = 0.223$ ($p < 0.001$) indicating only a weak positive correlation. the pairwise comparison can be found in the appendices (C.1).

While these results may initially suggest that synthetic aerial imagery is not yet at a level to cause concern, it is important to note that this was under specific forced choice conditions in which participants were aware that exactly one of the pair of images was synthetic. There was also only two choices for the participant to select from in each trial so even if every choice was completely random this would still result in 50% accuracy. If fake images were deployed in the wild against a less prepared users, it would be expected that a lower level of detection would occur.

The participant accuracy results show favour for H_1, that participants have difficulty in consistently distinguishing the fake EO images from the real ones. H_2 is also favoured as the results show a small but significant correlation between expertise and task accuracy.

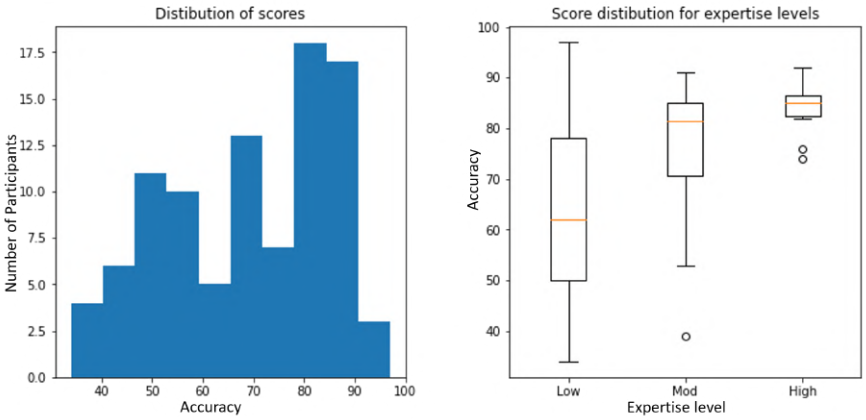

Figure 4.2: *(Left) Distribution of accuracy across participants and (Right) distribution of scores between different expertise*

For analysis, the synthetic images were manually separated into two groups, containing urban or rural scenes. Rural scenes were defined as containing natural features such as forest across the majority (50%) of the image, with images containing less than this threshold being classified as Urban. Examples of urban and rural scenes can be seen in Figure 4.1.

Experience level	N	Accuracy
All	94	68.9%
Low	60	62.9%
Moderate	24	77.2%
High	8	83.8%

Table 4.2: Experience level statistics from User study

Participants were able to better identify synthetic images that contained urban environments than those that consisted of primarily rural features. This is likely due to the fact that rural aerial imagery has less obvious and distinct features than those in urban scenes, making it harder to tell if the scene is naturally blurry or is a GAN image artefact. Errors in the generation of straight features such as roads and building edges is perhaps more obvious. The FID and KID were calculated for each image that was shown to participants (250 StyleGAN2 generated images). The average FID of the images shown was 4.02 and the average KID was 4.31, when calculated using a standard pretrained Inception network.

Metrics	**Correlation Coefficient**	**P value**
FID/KID	0.959	0.001
Accuracy/KID	-0.031	0.625
Accuracy/FID	-0.078	0.270

Table 4.3: Pearson's Correlation Coefficient for StyleGAN2 images and participant Accuracy

Correlations between the GAN metrics (FID/KID/ACC) were explored as seen in Table 4.3 using Pearson's correlation coefficient. The results found that while KID and FID had a strong positive correlation against each other, there was no significant correlation found between participant accuracy and either GAN metric. A comparison of means between human accuracy and FID/KID when split into rural and urban found that there was a significant difference between metrics showing that urban images achieved higher FID/KID scores (Lower is better) than rural images, but were more easily identified by participants (4.1). This shows that on the level of an individual image there is a disconnect between Inception distance-based metrics (FID/KID) and the human perception of photo-realism. This implies

that image generation algorithms do not necessarily require low scores for FID and KID for certain image types when the goal is to achieve photo realism as judged by the human eye. It should be noted that FID and KID are more unreliable when comparing the distributions of a single sample dataset and a real dataset as each samples distribution may be very different to the full dataset. This high variance and uncertainty can be seen as the FID/KID scores for each image are much higher than the dataset as a whole.

4.3.2 CNN Detection Results

Model	Acc.	Acc.(Real)	Acc.(Fake)
jpeg blur:0.1	60.98	97.79	24.16%
jpeg blur:0.5	62.58	97.97	27.19%

Table 4.4: Results from pretrained Inception network CNN image detection model. Two models were trained with different levels of blur added to the images. Accuracy for each class in addition to overall accuracy is included.

In the human detection study the mean accuracy was 61.2% (Figure 4.2), this result increasing to 74.5% when only looking at the results from participants with moderate or high previous experience. When the same image dataset was tested on the CNN detector the model achieved an overall accuracy score of around 60%, achieving 97.2% when presented with real images and 24.16% accuracy when presented with generated samples (4.4).

This disparity between the model's ability to correctly classify the two different image classes suggests that the model can learn the distribution of the real images but there is something in the GAN images that is causing the model to perform worse than chance. This could be due to GAN artefacts that are only present in this type of image. While the original paper states that the model was trained on a variety of image datasets, this did not include aerial data.

Comparing the CNN against the human detection results show that the CNN performs worse at the task than people with at least some experience looking at similar images. It should be noted however that while empirical comparisons can be made

of the differences between the human and CNN results, there are large differences in how each of these visual systems operate and require different testing protocols to measure them. The results do indicate that the detection behaviour between the CNN detection model and humans are different enough to justify further exploration with the goal of improving current detection techniques.

Our findings in EO data support current state of the art research on automatic and human detection of synthetic imagery in other problems, such as non-generated forgeries [63, 64] and faces [46]). Like the findings in these other areas, the results from this particular comparison suggest that while it is still possible for detection systems, human and automatic, to distinguish between real and synthetic EO data these methods are not at a level where they could be reliably deployed in a real-world scenario.

4.4 Discussion

The work in this chapter evaluated the extent to which synthetic satellite imagery can fool both humans and a SotA CNN-based detection model. The results of these detection experiments were also compared to the GAN evaluation metrics FID and KID for the same set of images. In addition to highlighting the limitations of human and automated detection towards synthetic EO data, the comparison between computational and human evaluation metrics for visual quality also adds to the growing body of literature[93] that calls for additional scrutiny on the industry standard use of Inception distance-based metrics as measures of sample quality.

The psychometric study we present here shows that current generative aerial images are at a point where they are becoming harder to distinguish from real images. The level of difficulty for detection varies depending on the level of experience of the participant. Further work into what specific expertise is useful for detection is needed to form a more comprehensive approach to tackling the potential issues that could arise with the use of fake satellite imagery for misinformation. Based on our results we speculate the differences in the rural/urban evaluations may arise from

attention differences between methods, although the experimental design we have used does not allow us to confirm this. Future iterations of this study could include additional measures such as human gaze analysis through eye tracking.

Building on our results, future studies using additional measures focusing on visual attention during detection could provide more clarity and insight into this. The disparity between automated (FID/KID) and human evaluation metrics is supported by previous comparisons [65] which found that correlations between human metrics and KID/FID varied between model, dataset and training instances.

The results from the CNN detection model also show that despite achieving $> 90\%$ performance on many of the common benchmark datasets, it still has some issues when applied to domain specific tasks such as the detection of fake EO imagery, with some room for improvement. Although a generalized GAN detection model is ideal, the work shows here that it may also be important to continue working on domain specific solutions that may be more reliable in practice. Future work should aim at improving the results of similar detection models but on smaller, more focused image types.

4.5 Conclusion

Following on from the benchmark comparison of GAN models for generating EO imagery in the previous chapter, the work in this chapter focused on synthetic image detection using samples from the best performing model, StyleGAN2. The work found that both humans and a CNN based detection model had varying levels of difficulty in distinguishing between real and generated images. Correlations for detection accuracy were also found with levels of expertise of human participants and what urban and rural features were present in the images. It was also observed that there were differences between the CNN detection model and humans for which types of images were harder to correctly predict as synthetic.

With these observed predictors of detection accuracy in this chapter, the next stage of research is to further explore if there differences in visual behaviour between

expertise groups and between the selected detection methods (humans and CNN models). This will be done by conducting a second psychometric study with additional measures to analyse visual behaviour during detection such as eye tracking.

Chapter 5

Human Gaze and CNN Attention in Detecting GAN Generated EO Images

5.1 Introduction

The image detection study detailed in the previous chapter 4 found that both humans and automated analysis (evaluation metrics or CNN image detection model) of synthetic EO data had their own limitations and differing results in evaluation. The user study in particular showed evidence that there is a need for further research into synthetic EO image generation as they have reached a point where they cannot be reliably detected in a highly controlled and simple 2 alternative visual search task. This makes the threat of going undetected in a real-life scenario potentially much greater. The initial user study in Chapter 4 was limited in its scope with a simple design and measurements, with user accuracy for correct detection being the primary evaluation metric of performance. This chapter focuses on a follow up study, using a similar 2 alternative forced choice design but this time with the additional capture of eye tracking during the duration of the experiment.

5.1.1 Hypotheses

The experiments in this chapter will explore two different comparisons using visual attention correlates such as eye tracking. The first is between human expert and non-experts and the second is between humans and a CNN based GAN image detection model. Based on the findings on the previous chapters (Chapters 3 and 4), and also in previous literature on visual attention summarised in this chapter, the hypotheses for the experiments are as follows:

H_1 : There will be significant difference in visual search behaviour and task accuracy between high and low experience groups in detecting synthetic aerial imagery.

H_2 : There will be a difference in ROIs for identifying synthetic aerial imagery between humans and CNN based detection methods.

5.1.2 Visual Attention in Human Cognition

Visual attention is an important area of research in both cognitive and deep learning based perception research[141]. Understanding the mechanism and hierarchies involved when a vision system, human or computational, interacts with a complex visual scene is important to many tasks in vision research such as object recognition, classification, segmentation, and detection.

For research into synthetic image detection and generation, visual attention is of particular importance. Regions of Interests (ROIs) and detection strategies of visual systems should be evaluated to better understand what factors may contribute to EO images being perceived as real or fake. Visual attention is often inferred by eye movements, although these two systems are partially independent, eye movements are typically preceded by shifts in attention [142]. Based on this relationship, visual attention in the context of human detection can be estimated by studying the movement of the eye in regard to the presented stimulus (e.g. real/synthetic EO images.)

5.1.3 Eye tracking

Eye tracking (oculography) is a common research method for accurately measuring visual attention during a given task by recording a person's gaze and eye movements[143]. These eye movements are usually recorded as either fixations or saccades[144]. Fixations are recorded when the eye tracker detects the eyes focuses on a specific region of interest for a fixed amount of time. A saccade is a short, rapid movement between different fixation points [145].

The precise values that define the parameters of fixations and saccades (e.g. minimum fixation time, minimum saccade distance) are often defined in context of the recording scenario, the task given and the accuracy of the eye tracking hardware. In the case of fixations, it has been generally found[146] that a longer fixation duration indicates a higher cognitive workload and a focus of visual attention on that particular ROI. Typically, fixation duration values are recorded between 200 and 350ms[147], fixations significantly shorter than these values are typically not long enough to be embedded into long term memory. Individual differences in fixation duration have also been suggested to be determined by information processing and cognitive processes.

Saccades occur when the eye movements recorded show a sequence of rapid fixations across the target image[144]. During a saccade, no visual information is processed making them less of a useful metric for visual attention than fixations. Unlike fixations, saccades can be used to determine search behaviour and the specific sequence viewing patterns towards a given stimuli, which can be useful in determining the visual hierarchy of an image by analysing in what order a series of ROIs are viewed.

The use of eye tracking research has been applied successfully in many different fields, like, improving the gaze behaviour of pilots[148], a training tool for medical professionals [149] and also for assessing image quality[150]. These applications of eye tracking analysis give useful insights into the visual attention differences for differing groups of domain experts. One eye tracking paradigm which has been used in a variety of research fields is in investigating the differences in visual attention

between experts and novices[151]. This paradigm works under the assumption that in a visual search task that benefits from domain knowledge, the visual attention of experts will be informed by their prior domain knowledge. The resulting differences in the recorded eye movements in comparison to that of novices highlights the most relevant ROIs and search patterns to success at that particular visual task. This can give informative insights into how experts use their domain knowledge to achieve superior performance in a way that the experts themselves may not be consciously aware of enough to articulate themselves.

Observed eye movement differences between experts and novices have been studied across different skills and domains. One example is in chess players, expert players have been found to fixate their attention on pieces more relevant to their next move than novices [152]. Another study observed aircraft pilots, the more experienced pilots spent more time looking out of the cockpit than novice pilots, they also fixated on a wider array of cockpit instruments but for shorter amounts of time than less experienced pilots [148]. Expert vs novice studies have also been applied to earth observation imagery. One study which looked at the differences between expert and novice geoscientists when viewing data visualizations[153] found that experts spent a larger proportion of their time focusing on the latitude and longitudinal axes on EO images, while the students (novice group) spent more time looking at the colour bar key. Both groups had similar behaviour patterns when looking at the actual geomorphological data itself but when interviewed after the study had different interpretations on what they were viewing.

In the study of GAN generated imagery, there is relatively little literature that looks at eye-tracking data towards generated samples. One study that has explored this area is Caporusso et al. 2020 [154], this work looked at the visual attention of participants when given images of either a real human face or a synthetic face generated by a StyleGAN implementation. They found that the participants with the highest detection accuracy had covered a much larger area of the image, with their visual attention spread out throughout the image. This contrasted with the lowest performing group who focused only on the main facial features (eyes, nose

and mouth). In this case, the results identified that it was a more accurate search method to look at the entire image rather than just the face region as one of the issues with StyleGAN generated faces are the often-incoherent backgrounds[1] as the GAN only learns to generate realistic faces, with little regard for the peripheral parts of the images.

5.1.4 Evaluating Visual Attention in Deep Learning with Post-hoc Attention

Attention in deep learning comes with multiple definitions depending on the exact mechanism it is describing. One broad definition given in a 2022 review on Visual Attention Methods in Deep Learning[141] describes attention as "a mechanism that imitates the human cognitive awareness about specific information, amplifying critical details to focus more on the essential aspects of data." Typically, visual attention either refers to some kind of trainable attention mechanism within the model or a post hoc examination of a model's attention focuses such as Grad-CAM which looks at the class activation maps (CAM) of a model regarding a given input image.

Gradient-weighted class activation mapping (Grad-CAM)[155] is a popular method for looking at the visual attention of trained network models. This method works by using the gradients for a target class in the final convolution layer of a network to create a coarse localization map for a given target image. This results in a heat map over the image highlighting regions of importance to the class prediction. This method of analysing a networks visual attention is useful as a model evaluation metric to gain insight into if the model is working correctly and focusing attention on the right parts of the image for prediction. One application for this would be to detect if a model is overfitting to a dataset or producing other aberrant behaviour that typical metrics such as prediction accuracy would not detect.

The original Grad-CAM provides a useful analytical tool but does have some noted shortcomings. One issue is that it will fail to properly localize the predictive regions if there are multiple occurrences of the same class in an image, additionally localization

can often not cover the entire object but will only highlight parts of it. To improve on these points Grad-CAM++[156] was proposed which gives better object localization and is better suited for multiple class instances in a single image.

In addition to Grad-CAM++ there are numerous other CAM models available for evaluation. HiResCAM is one such model that also builds upon Grad-CAM which proposes to provide a more accurate representation of network visual attention. It does this by fixing an issue caused by Grad-CAMs gradient averaging step which causes some of the relative scales and magnitudes of the gradients for the target feature maps to be lost[157]. HiResCAM instead applies an element-wise multiplication between the feature maps and gradients directly, resulting in a better reflection of the regions most used in computation for the model's class predictions. Ablation-CAM[158] is another evaluation model which proposes a "gradient free" method for representing visual attention in networks by using ablation analysis to determine the importance of individual feature maps units. This approach avoids the problem in gradient based methods of "gradient saturation" where the back-propagating gradients are reduced to a point where they will not be visualised on the final heat map, despite being of significant importance. This is also avoided in Score-CAM[159], which is also a gradient free method that calculates the weight for each activation map from its target class score in the forward pass.

Regardless of variant, CAM models provide useful tools for more explainable deep learning models[160]. The move away from "black box" models and the trend towards more explainable AI is becoming increasingly more important as this technology becomes more ubiquitous in everyday tasks. CNN models and other architectures which contain "black box" like inner workings and latent spaces have been found to be distrusted by the public[161], an issue which will surely increase as the reliance on these models increases.

5.2 Experiments

For further insight into the visual attention and behaviour during the human detection of synthetic EO images, a detection study which recorded participant's eye movements was conducted. The visual attention of a CNN image detection model[45] was also estimated using class activation mapping techniques.

5.2.1 User Study

To measure participant's detection behaviour towards synthetic EO images a user study was conducted with a similar design to the study in the previous chapter (Chapter 4). This study again utilized a two alternative forced choice design where participants were presented with a simultaneous pair of images consisted of one real EO image from the Inria Aerial Benchmark dataset and one StyleGAN2 generated image. The original images in both real and fake datasets were 512x512 and enlarged to 768x768 for the experiment. This increase in resolution scale from the previous study (256x256) is to allow the eye tracking camera to more accurately capture changes in fixation, as the images now cover a larger area on the screen. If the images appear too small on the screen, the eye movements of participants may be too small to differentiate fixation points.

Participants were asked which image of the pair was synthetic. Each participant was presented with 300 pairs of images with a break every 50 images to reduce the possibility of fatigue. The 300 image pairs for each participant trial were randomly selected from a dataset of 1800 real images and 1800 generated images. Participant's experience level was recorded, and their task accuracy was measured in addition to eye movement fixations via an eye tracking camera.

Participants were recruited from different departments at the University of Nottingham, in particular the School of Geography and the School of Computer Science. Participant experience was grouped based on their own self-reported experience with satellite image data rather than their respective department, although most of the "high" experience group came from the school of Geography. Overall, 27 partic-

ipants took part in the study, divided into 3 groups, High experience ($N = 12$), Moderate experience ($N = 7$) and Low experience ($N = 8$).

5.2.2 Eye Tracking Setup

For capturing participant's eye movements, a Tobii Pro Nano Camera was used, with participants sitting approximately 60cm from the screen. The Tobii Pro Nano has a sample rate of 60Hz, accuracy of 0.3°and a latency of 17ms. Raw gaze points for the screen were captured for both the left and right eyes of participants and then average was recorded for each gaze point pair. A natural gaze protocol was used as participants were given no instructions on any ROIs during the task. Participants were instructed to return their gaze to central fixation cross between trials. A 9-point calibration was conducted at each 50 trials interval before the start of the next set. Both the image detection task and the eye tracking recordings were implemented using PsychoPy and were conducted in person in a lab setting.

Both fixation and saccades of participants were recorded for each trial, however, the analysis of the results focuses primarily on fixation points with a minimum of 100ms and a maximum duration of 500ms. The decision to use fixation points rather than saccades was due to the short total duration of each individual trial and to make a more valid comparison with the detection behaviour of a CNN detection model which does not include any temporal processing needed to be comparable to saccades.

5.2.3 Gaze Heatmap Calculation

The raw gaze points obtained during the experiments used the Velocity-threshold identification algorithm (I-VT)[162]. This is a velocity-based classification algorithm used to convert the raw camera data relating to the captured eye movements on the screen to gaze data of fixations and saccades. I-VT uses a threshold value to classify the eye movements based on the velocity of the directional shifts of the eye. Once a threshold value is set, any data above that value is classified as a saccade and below

that value as a fixation. After the fixation points for each eye were obtained, each pair of coordinates were averaged to give the final gaze points used for the heatmap generation. As per the I-VT algorithm, the points are then collapsed into fixation group which is then mapped to the centroid of its consisting points. To create the heatmap a Gaussian blur was added to the centre of these groups and plotted over their corresponding images. This process led to the final heatmaps presented in this chapter (Figure 5.1).

The minimum duration for a fixation point to be recorded was 100ms. This fixation threshold value is towards the lower bounds for threshold values in gaze research. The use of a low fixation definition was due to some participants only looking at an image for less than 1 second, making it hard to capture any gaze points at all if a higher fixation value was used.

5.2.4 Gaze Entropy

Gaze entropy refers to a set of gaze evaluation metrics that aim to provide quantitative measures to the uncertainty in scanning behaviour from a given set of gaze fixation points[163]. These measures use the information theory concept of entropy which here, describes the amount of information necessary to produce a given sequence. With a given set of gaze points ordered in a temporal sequence, stationary gaze entropy is calculated using Shannon's entropy equation (5.1) to give the average level of uncertainty in the gaze sequence.

$$H_s = -\sum_{i=1}^{n} p_i \log(pi) \tag{5.1}$$

In equation 5.1, p_i is the probability of viewing the i^th ROI, with n being the total ROIs. p_i is calculated by dividing the number of fixations in an ROI by the total number of fixations, the minus sign is to ensure the value is between 0 and 1, as the log function gives a negative output, therefore H_s is always positive.

This means that for higher values of stationary gaze entropy value(H_s), there is less of an overall fixation on a certain ROI as the higher entropy equates to less

predictability and a higher overall spatial dispersion of gaze points. An individual with a high H_s will focus less on a single ROI than an individual with a lower H_s. Gaze transition Entropy (H_t) is a further metric which measures the individual's unpredictability in gaze patterns when switching between ROIs. This is calculated by modelling the temporal gaze sequences as Markov chains. Previous findings [164] have indicated that predicting the next fixation point in a sequence is more accurate when predicted using the current fixation location rather than the overall probability of the prior locations. With this assumption, the Markov property can then be applied to gaze data and modelled using the Markov chain rule. Shannon's equation of entropy is then applied to the 1st order Markov chain matrices of the fixation transitions to give the transition entropy. This gives a measure of the average uncertainty and therefore unpredictability of the scanning behaviour. This is given by the equation 5.2:

$$H_t = -\sum_{i=1}^{n} p_i \sum_{j=1}^{n} p_{ij} \log(pij) \tag{5.2}$$

Again, like in stationary entropy, p_i is the probability of looking at the i^th ROI, with n being the total number of ROIs. pij is the probability of a fixation at j_th ROI with respect to the previous fixation of the i_th ROI. These probabilities are represented by a transition matrix, with the matrix rows being the source ROI and the columns being the destination ROI. This transition matrix is calculated by the equation :

$$p_{ij} = \frac{n_{ij}}{\sum_j nj} \tag{5.3}$$

In this equation the transitions $\sum_j nj$ are divided by the total number of transitions from the source ROI.

A higher H_t value suggests a less predictable and less structured visual search behaviour between ROIs [165]. In the context of this study, this would indicate the participant is switching back and forth between real and fake images, with a less methodical search pattern than a participant with a low H_t. These metrics have

been applied to gaze data analysis on a variety of tasks such as measuring the effects of sleep deprivation on drivers[166] and evaluating situational awareness in emergencies at nuclear power plants[167]. Together the two metrics (H_t , H_s) are useful measurements for linking top down interference events with gaze pattern behaviour [168].

In this chapters' image detection study, for each participant's data, H_t and H_s are calculated for each trial (real/fake image pair) and then averaged across all trials. The ROIs used in the gaze entropy calculations were the real and fake images, which were each further split into 9 equal sections for the purpose of gaze entropy calculations. The inclusion of these metrics is to provide insight into the differences between groups in visual scanning between the two images on screen (left/right) when they are trying to make a detection decision.

5.2.5 CNN Detection

To compare the eye tracking results to the CNN image detection model, GradCam++ was used to visualise the networks attention in the last convolutional layers of the ResNet model. The model analysed used the pretrained weights from the results cited in the original paper that were made available on GitHub by the model creators. Multiple Cam models were tested but there was not much variation between the different types tested (GradCam, HiResCam, Ablation-Cam) so only the results from GradCam++ were reported.

5.3 Results

A multivariate analysis test (MANOVA) between all dependent conditions (Table 5.1) indicated that there is a large significant difference in the dependent vector (experience) between the different groups, $F(12, 36) = 4.35$, $p < .001$, Wilk's $\lambda = 0.167$, partial $\eta 2 = .59$.

5.3.1 Difference Between Expertise Groups

Experience	Avg Resp Total	Avg Resp Corr	Avg Resp Incorr	Acc	Avg H_t	Avg H_s
High	1.396	1.405	1.191	91.500	0.201	0.488
Moderate	2.363	2.303	3.643	92.833	0.210	0.523
Low	3.960	4.141	4.840	68.000	0.233	0.604

Table 5.1: Gaze results for experience groups. The variable values are average response time (Avg Resp Total), average response time for correct answers (Avg Resp Corr) and incorrect answers (Avg Resp Incorr), task accuracy (Acc), gaze transitional entropy (H_t) and gaze stationary entropy (H_s).

Variable	Experience Pairs (G1, G2)	G1 Value	G2 Value	Mean	SD	P-Value
Acc	High-Moderate	91.500	92.833	1.43	4.57	0.97
Acc	**High-Low**	**91.500**	**68.000**	**23.70**	**4.57**	**<0.01**
Acc	**Moderate-Low**	**92.833**	**68.000**	**25.14**	**5.14**	**<0.01**
Avg resp total	High-Moderate	1.396	2.363	0.96	0.45	0.3
Avg resp total	**High-Low**	**1.396**	**3.960**	**2.56**	**0.45**	**<0.01**
Avg resp total	Moderate-Low	2.363	3.960	1.59	0.51	0.087
Avg Resp Corr	High-Moderate	1.405	2.303	0.9	0.49	0.41
Avg Resp Corr	**High-Low**	**1.405**	**4.141**	**2.74**	**0.49**	**<0.01**
Avg Resp Corr	Moderate-Low	2.303	4.141	1.84	0.55	0.067
Avg Resp Incorr	**High-Moderate**	**1.191**	**3.643**	**2.45**	**0.65**	**<0.01**
Avg Resp Incorr	**High-Low**	**1.191**	**4.840**	**3.65**	**0.65**	**<0.01**
Avg Resp Incorr	Moderate-Low	3.643	4.840	1.2	0.73	0.49
H_t	High-Moderate	0.201	0.210	0.0092	0.0062	0.56
H_t	**High-Low**	**0.201**	**0.233**	**0.032**	**0.0062**	**<0.01**
H_t	**Moderate-Low**	**0.210**	**0.233**	**0.035**	**0.007**	**<0.01**
H_s	High-Moderate	0.488	0.523	0.035	0.017	0.33
H_s	**High-Low**	**0.488**	**0.604**	**0.12**	**0.017**	**<0.01**
H_s	Moderate-Low	0.523	0.604	0.081	0.019	0.018

Table 5.2: Pairwise (G1, G2) ANOVA Results for each experience group. The variable values (G1, G2), absolute difference between means (mean), standard deviation (SD) and P-values are reported. The variable values are task accuracy (Acc), average response time (Avg Resp Total), average response time for correct answers (Avg Resp Corr) and incorrect answers (Avg Resp Incorr), gaze transitional entropy (H_t) and gaze stationary entropy (H_s). The significant results are in bold

An ANOVA post hoc was also performed for the individual DVs between groups (Table 5.2). For accuracy there was a significant difference between High-Low ($p < 0.01$), Moderate-Low ($p < 0.01$) but no significant difference between High-Moderate ($p < 0.97$). Significant differences between response times for High-Low were found;

High/Low variables	F	P-value
Accuracy	16.9 (1,24)	<0.001
Avg Resp Total	13.2 (1,24)	<0.001
Avg Resp Correct	13.6 (1,24)	<0.01
Avg Resp Incorrect	8.3 (1,24)	<0.001
Transition	12.3 (1,24)	<0.01
Stationary	20.1 (1,24)	<0.001

Table 5.3: ANOVA results for merging experience groups High and Moderate vs Low

avg resp ($p = 0.01$), avg resp correct ($p > 0.01$) and avg resp incorrect ($p < 0.01$). Significant differences for stationary gaze entropy were found between High-Low ($p < 0.01$) and approaching significance for Moderate-High ($p = 0.05$). Stationary gaze entropy was found to be significantly different between High-Low ($p < 0.01$) and Moderate-Low ($p = 0.01$). The between groups ANOVA results can be found in the appendices (Table C.1)

When groups high and moderate were grouped together resulting in just two groups (High, Low), a significant difference was found between all measures (Table 5.3).

5.3.2 Task Accuracy

Like the results in the previous image detection study (4) participants that reported their previous experience as either "high" or "moderate" achieved higher task accuracy than those who responded with "low". As with the previous results, this provides further evidence that self-reported experience shares a positive correlation with task accuracy, a useful finding for further studies as it supports the use of self-reported experience as a grouping metric for expert/novice designs. In comparison with the previous image detection study, all participant groups performed much higher on average. This could have been caused by a few differences between the studies. One difference is the size of the images presented to participants, the larger images in this study make it easier to spot anomalies and discern the images as real or fake. The initial study was also hosted online, due to the COVID restrictions at the time, while this current study was held in lab conditions. Online studies have been found to have decreased levels of participant engagement than ones held in the

lab [169], which could negatively effect task performance.

Although there was a significant difference (Table 5.3) between the High/Moderate and Low groups, there was no significant difference between high and moderate. This effect of understating one's experience levels has been seen before in research[170], where participants may have a much higher level of expertise than they believe. Although one takeaway from this could be that future experiments using the same self-report methodology should instead only use the categories of high and low, another interpretation would be that having a middle category (e.g. moderate) is important as it minimises the potential effects of underestimating expertise. With just two categories, this would potentially result in those who do underestimate their expertise self-reporting as "low", which skews the borders between groups. The results from this experiment show that having a "moderate" group gives a better grouping for those with actual low expertise.

Rural and Urban Image Accuracy

Experience	Rural	Urban
High	0.86	0.92
Moderate	0.78	0.83
Low	0.81	0.92
All	0.83	0.90

Table 5.4: Participant accuracy for correct detection of generated urban and rural images

In the previous study (Chapter 4), participants were found to have a lower detection accuracy for generated images that depicted rural scenes than those that depicted urban scenes. This is also seen in the results from this study (Table 5.4). For all experience groups an Wilcoxon Signed-Rank test indicated that there is a significant large difference ($p < 001$, effect size=0.8) between means for rural accuracy (0.83) and urban accuracy (0.90). Although, like with overall task accuracy, the scores are higher than in the previous study, likely due to factors such as increased image size and being conducted in person. There was no significant difference found between experience groups, with rural images being more challenging to correctly identify regardless of experience levels.

5.3.3 Response Times

In addition to differences in task accuracy, the average response times for each detection also differed significantly between groups. Those in high/moderate took significantly less time to make their detection choices ($1.40ms$) than those in Low group ($3.96ms$). This further reinforces that self-reported expertise can be used to differentiate experts from novices as they achieved higher task accuracy and needed much less time to do so. This is supported by the literature as it suggests that experts make more informed searches, to quickly identify key features in the image. Novices are more likely to scan the entire image without noticing any of these visual cues [86]. All groups had higher response times to incorrect guesses than correct guesses. For correct responses, this could be caused by a tendency to double check answers when the participant thinks they may be correct but lacks complete confidence due to a challenging image pair, leading to a higher response time. This could be similar in the inverse scenario, if the participant is unsure and the image is too challenging, they may make a quick guess to move on to the next image pair.

5.3.4 Gaze Entropy

Both stationary gaze entropy (H_s) and transition gaze entropy (H_t) decreased with experience levels (Table 5.1). This implies that expert visual behaviour follows a more structured and less random search pattern in analysing each real/synthetic pair as opposed to the more random and dispersed gaze behaviour of lower experienced participants, despite the fact that the images in this task were novel to both groups. The high H_s value (0.604) recorded in the Low experience group indicates that they more evenly distributed their search across the entirety of the images with not as much of a visual focus on individual features as the High experience group ($H_s = 0.488$). The lower H_s for the higher experience group means that their search strategies were more concise and focused on features that may give away the nature of the image's realness. Transition gaze entropy was also found to be significantly lower($p > 0.01$) for the expert group ($H_t = 0.201$) than the Low experience group

($H_t = 0.233$). This means that their gaze switched less between images, focusing more on identifying one image rather than going back and forth comparing images. Overall, the gaze entropy metrics between groups support previous findings on other expert/novice studies[171] which also find that there is less variance and entropic uncertainty in expert search patterns compared to novices, as experts already have a schema on what cues to look out for that give away an image as either real or fake. Less experienced participants with less developed schemas on what to look out for and are more likely to pass over give away details, resulting in them searching more of the image and making more pairwise comparisons for each given trial.

5.3.5 Gaze Heatmaps

Figure 5.1: *Raw fixation points (left) and the final heatmap (right)*

As the main findings (Figure 5.1) found there to be more measured variables that significantly differed between the High and Low experience groups, a visual inspection of the gaze heatmaps from these groups has been included in the analysis. The samples included here were randomly selected and viewed by the same number of participants from each respective category.

Together with the differences in stationary gaze entropy the visual analysis of the gaze heatmaps (Figure 5.2) show that there is generally a smaller range of ROIs as shown by the spread of fixation points for the high experience group than the low

Figure 5.2: *High expertise (centre) and Low expertise (right) gaze heatmaps over synthetic urban scenes from a combined sample of 7 participants for each group. The high expertise group (centre) shows more focused gaze on few ROIs than the low expertise group (right)*

experience groups. The high experience groups can also be seen to concentrate on the same areas in making their detection predictions.

Figure 5.3: *High expertise (centre) and low expertise (Right) gaze heatmaps over synthetic highway images. The high expertise group exhibits a greater focus on small details such as cars than the low expertise group*

ROIs as defined by the areas of concentrated fixation points show that the high experience groups spending more attention searching the edges of buildings, in particular areas which have the common GAN artifacts of poorly defined straight lines and edges or where roofs of buildings blend unnaturally into other parts of the image. High experience individuals also pay more scrutiny to smaller, objects such as cars, especially in highway images, as shown in Figure 5.3. Like straight edges, cars in the generated images can often lack detail on close inspection.

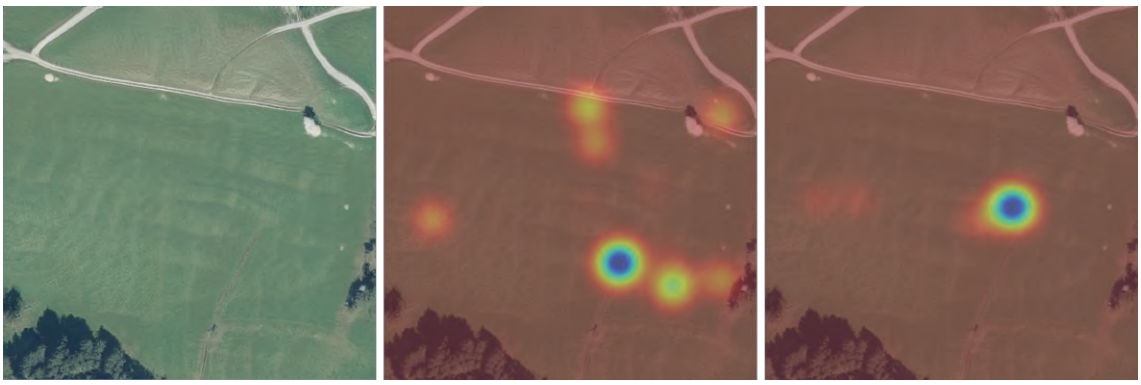

Figure 5.4: *High expertise (centre) and low expertise (right) gaze heatmaps over synthetic rural scene. The low expertise groups shows a less dispersed and more central grouping of fixation points. This could indicate that for images that are harder to identify, such as rural images, the low expertise group are switching, and comparing between both images that have been presented (real/synthetic). The gaze heatmap for the high expertise group (left) conversely feature a higher degree of visual exploration of the image. This could be that the high expertise group is spending longer searching the image for identifiable markers.*

One deviation from the differences in fixation spread and gaze coverage is in rural scenes (Figure 5.4). In these images there was more dispersed spread of fixation points for the High experience group and a smaller number of fixation points near the centre of the image for the Low experience groups. One potential explanation for this is that rural scenes have been found to be harder to identify as real or generated (Table 5.4) than urban scenes, as found in the study in the previous chapter. This could indicate that low experience participants glance at the image once before switching their gaze to the other image on screen (in this case the real image), hence the large grouping of fixation points in the centre of the image but no further searching. High experience individuals, who as shown by their lower transition gaze entropy, are more likely to focus on one image, continue searching the image for any visual cues. Due to rural images being harder to identify, there are no obvious visual cues which results in a low consensus on the ROIs of the image, giving a more dispersed heatmap.

5.3.6 CNN Detection Comparisons

The generalized CNN detection model [45] was evaluated on the same subset of real/synthetic images that were used in the human detection study. The model achieved an accuracy of 82.09% for correctly predicting if an image was synthetic. This is lower than the average detection accuracy of the High and Moderate human groups (92.42%) and slightly lower than all the average of all participants (85.71%). Grad-Cam++ was used to visualize what ROIs in the given images the network found important to classification. Samples were then compared to those from the human detection study. It is important to note these comparisons cannot be thoroughly quantified due to inherent differences in the way the visual attention data has been collected for each method (CNN and human) and also the inherent differences between them (e.g. no temporal aspect for the CNN detection). Comparing the two methods can still give insight into the differences in detection behaviour.

One of the most apparent differences between the two methods was the visualization

of attention in rural images (Figure 5.5). The CNN detection model had much higher activations across the entire images for rural scenes, indicating more features in the image that were useful for prediction. This supports the findings in the previous study which found that computational methods were able to differentiate better between real and synthetic images than humans. From a visual inspection of these images, it also supports the idea that features that are useful for detection for CNN based methods differ from those with people.

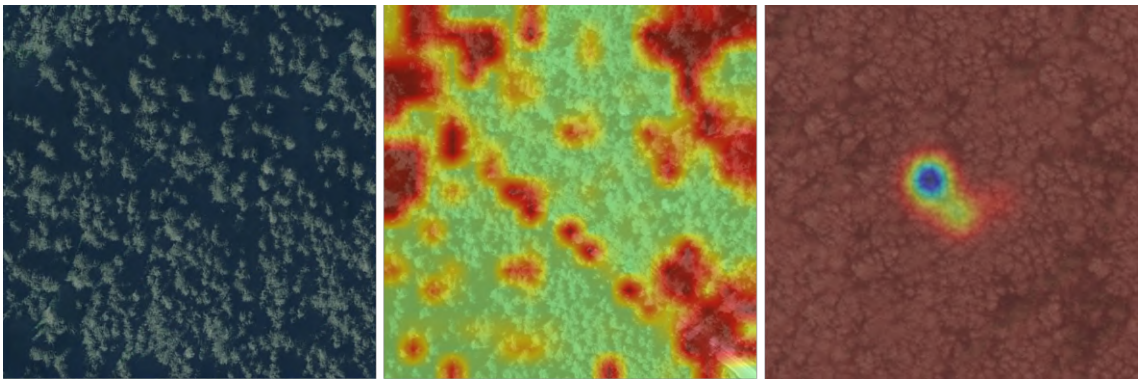

Figure 5.5: *CNN (centre) and Human (all groups) (right) attention heatmaps on synthetic rural imagery. As discussed in Chapter 4, humans find rural synthetic images harder to identify than CNN models. This can be seen in these heatmaps, the Grad-CAM image (centre) shows that large areas of the image contain features that give a high probability of the image of being synthetic. The human gaze-map (right) contains a single central grouping of fixations, indicating that the participants could not identify any obvious features and instead switched their gaze back to the other image on the screen.*

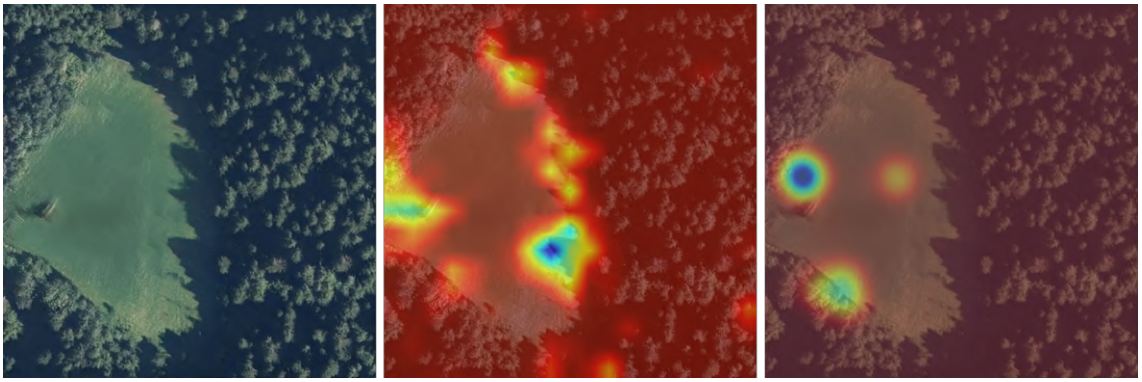

Figure 5.6: *CNN (centre) and Human (all groups) (right) attention heatmaps on boundaries. The CNN detection model places greater attention on boundaries when making predictions than humans. The gaze maps from humans (right) show that while image artifacts around boundaries are still used, there is also a larger visual focus on features (e.g. buildings) when compared to the CNN model.*

For synthetic EO images that depicted more obvious features to the human eye (e.g. buildings, cars and roads) the CNN model showed higher activation for boundaries

between different areas (figure 5.6). While this was also seen in the images from human participants, it was more heavily weighted than other features like buildings in CNN detection. Human participants were found to focus more on objects such as buildings or cars.

Boundary lines can be seen to be the most common visual feature that the model used for its predictions, a lot of the time with much lower activations on buildings. Other areas with high activations, seemingly random to the human eye as they are devoid of anything of visual note, indicate that the CNN can find GAN image artefacts which humans may not be able to perceive. These may include differences in colours or textures between real and synthetic images.

5.4 Fine-tuned CNN Detection Model

Using an unseen section of the Inria dataset (10,000 images), the CNN detection model was fine-tuned to see if exposure towards EO specific images improved its ROI attention (Figure 5.8). The fine-tuned model achieved an accuracy of 88.03%, which is a modest increase of the general GAN detection model (82.09%) and still lower than High and Moderate human groups (92.42%). Although this is a relatively small increase in task accuracy, the GradCam++ images show that fine-tuning the model increases the focus of the network to more specific areas. This is likely due to the network having learnt to identify EO specific features rather than simply GAN artifacts like seen in the original model. This can be seen most prominently for the images in the second and final rows of Figure 5.8. In these images the fine tuned model can be seen to take into account local features (e.g. roads, buildings) more so than the original model which gives greater weight to global features (e.g. GAN specific image textures) in influencing its predictions.

The GradCam++ images comparing the two GAN detection models, demonstrate that for domain specific image detection, there is still value in fine-tuning on similar data, even when the model used is a general-purpose GAN detection model. Another interesting result from this comparison is that despite the fine-tuning increasing the

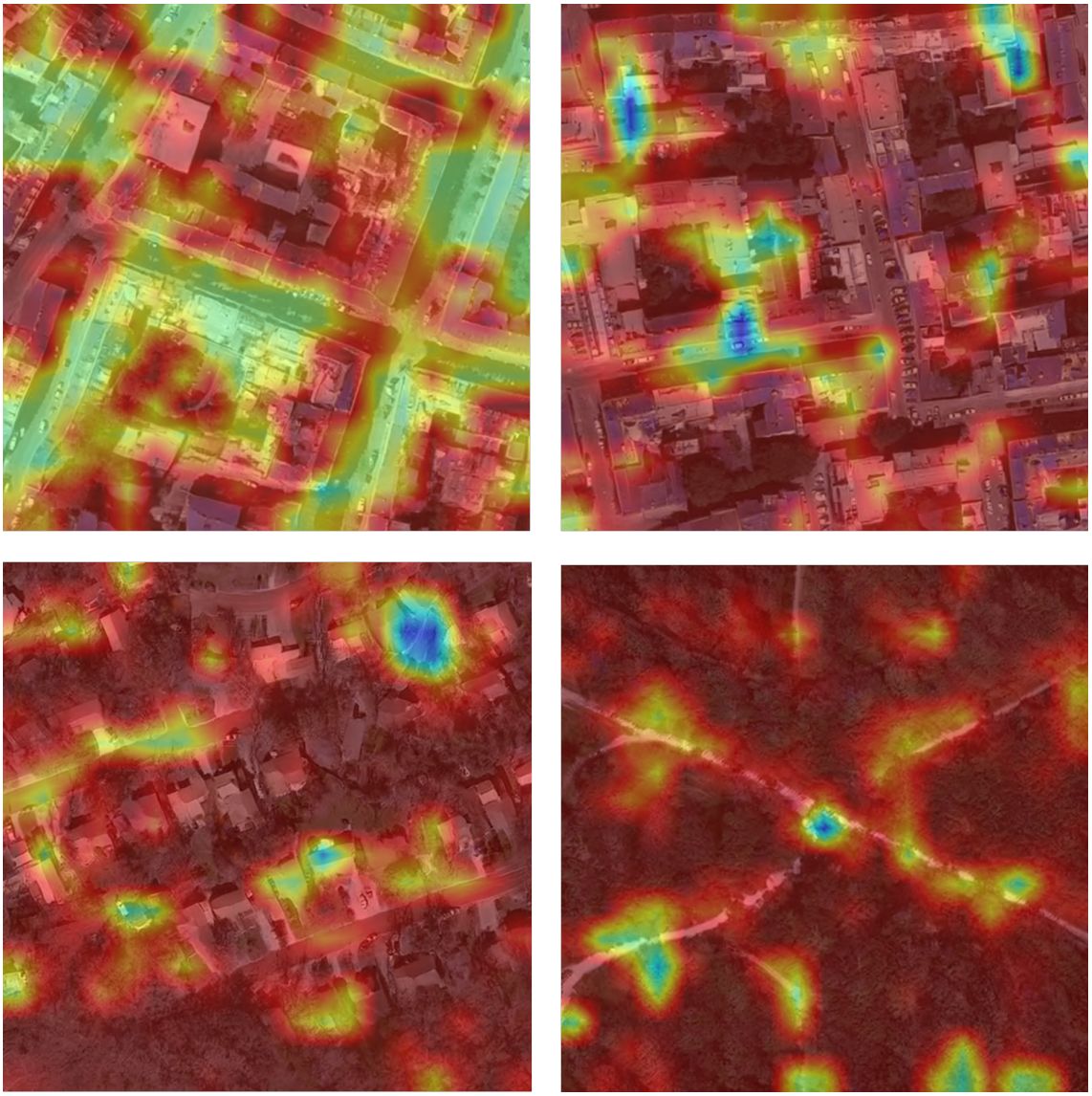

Figure 5.7: *Additional samples of GradCam++ class activation mappings over synthetic EO images. The top two images depict images with primarily urban features. The bottom left image contains more suburban features and the bottom right image consists of just foliage and roads. In all of the images displayed, the heatmaps indicate that boundary lines are the most heavily weighted features for correct predictions of synthetic EO images.*

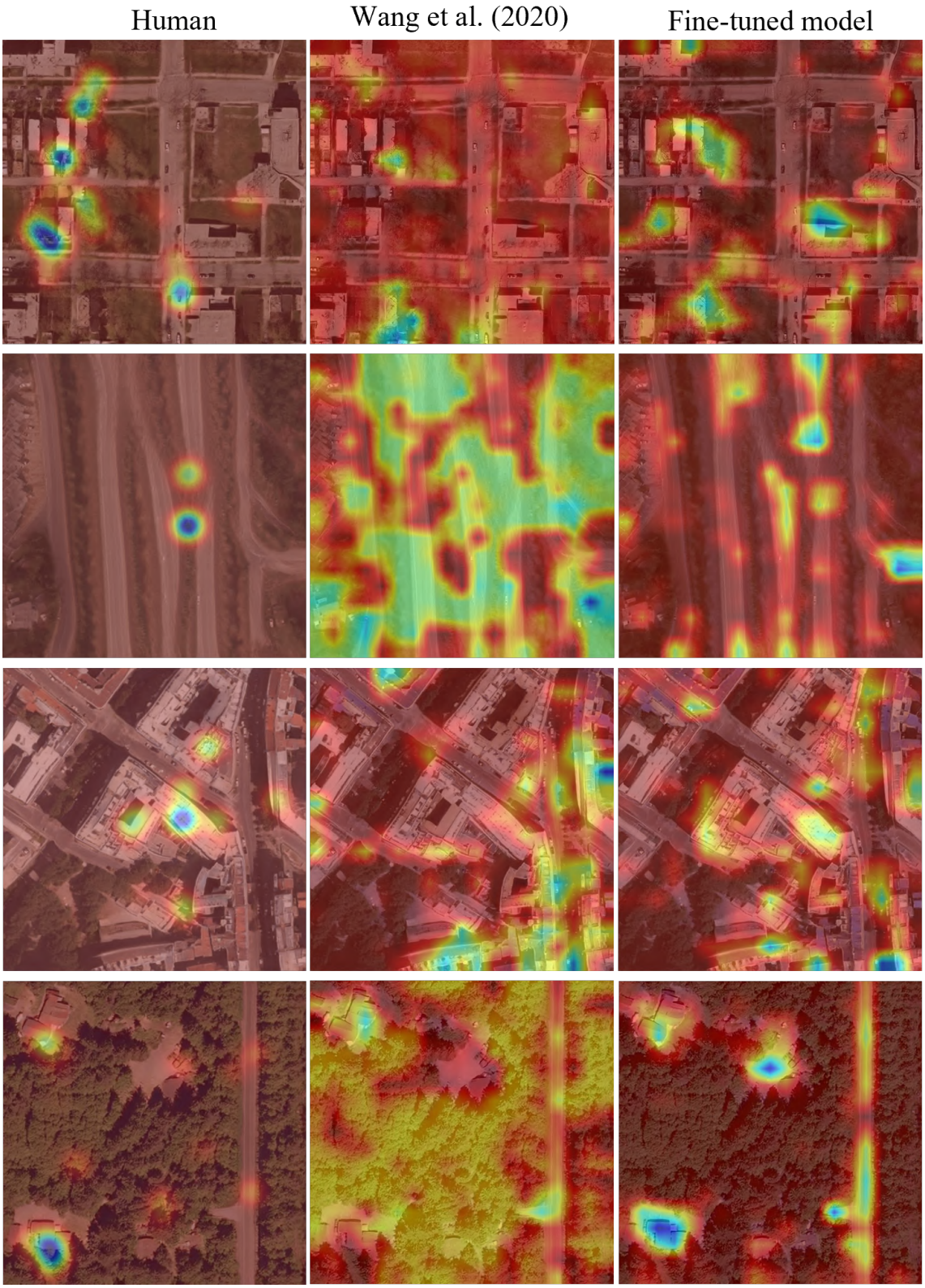

Figure 5.8: *Samples from Human attention maps (left) compared against GradCAM++ samples from a CNN detection model (Middle) and an EO image fine tuned detection model (right). Fine tuning the model on EO images shows an increased focus on particular features in the images compared to the general detection model, although there is still a large disparity between human and CNN ROIs.*

focus on specific EO features, the ROIs for the computational models still differ from those in the human gaze maps. Out of these comparisons only the final row of images shows the fine-tuned model aligning more with human attention, as the model is now looking at EO specific images. This lends further evidence that their is fundamental differences between human and CNN's in the key identifying features for synthetic image detection.

5.5 Discussion

The results from the eye tracking image detection task supported the initial hypothesis that a significant difference would be found in both task accuracy and visual search patterns between high and low experience groups in detecting synthetic aerial imagery. This is evidenced by the differences in gaze entropy metrics and the visual analysis of gaze heatmaps between the groups. It was found that higher experienced individuals applied a more efficient visual search to the images and were more actively looking for image artefacts which appeared abnormal or false. This differed from those with less experience who searched over larger areas of the images and switched their gaze between the two images on screen.

The higher rate of gaze switching suggests that less experienced individuals' detection techniques were more centred around comparing both images on screen to see which one appeared less real, rather than actively looking for GAN markers in individual images. This could also infer a lower level of confidence in the detection strategies in this group. In addition to a higher accuracy in correctly identifying synthetic images, the High experience group were also faster at making their predictions, with a higher average response time further supporting that their search strategies were more efficient and gave a higher level of confidence to their predictions.

The CNN detection GradCAM images also supported the hypothesis that there would be differences in attention between human and CNN detection methods. Although as mentioned before, this comparison is limited by the differences between

the testing conditions and detection methods themselves. The differences between a computational CNN classifier and the human visual system mean that two different testing methods were needed to capture detection performance.

A paired image detection study was used for humans and the CNN model was given single images in batches. This meant that the CNN was given a single image classification while the human participants were performing a paired comparison classification task. The images themselves were kept the same but there was no comparison element to the CNN model testing or any form of temporal dimensions. This is not ideal but is unavoidable, as while the outcome behaviour between methods is the same (predicting real/synthetic), the visual systems (biological and computational) are vastly different despite a few minor similarities. The other major difference is how the heatmaps were obtained and what they represent, the CNN model coming from class activation mappings which show the areas of the image with the highest activations for a class (synthetic) predict as obtained from the gradients passing into the final convolutional layer.

While GradCAM methods have been extensively studied and are presumed to represent visual attention, the degree to exactly how close this does this is still debated, hence multiple different methods that have been developed to improve on the original GradCAM methods. The human gaze heatmaps are also used to visualize visual attention via measuring fixation points on a given target. This method has previously been found to be a valid correlation to visual attention [172] but is limited by factors such as the refresh rate, accuracy, and calibration of the eye tracker. Additionally, there is not an exact agreed upon threshold values for defining fixation points and saccades.

Despite these limitations that come with using individual testing methods, there are still some interesting observations that can be made by looking at the differences between CNN detection models and humans. The gaze heatmaps from humans show that visual attention is drawn towards objects such as buildings much more than in the GradCAM images. One explanation for this is that the CNN detection model has been trained specifically to look for differences between real and GAN

generated images while the human visual attention system is going to naturally be drawn towards features that are more familiar to people, in this case well defined objects.

A common fingerprint for the synthetically generated images for both detection systems was boundaries, a GAN artefact which is present in many of the images is poorly rendered straight lines or boundaries between textures, for example road markings and street edges. This was observed particularly in the case of the CNN GradCAM images, indicating that it was one of the strongest features for assisting the model's predictions. In some cases, there would be very low activations on visually misshapen buildings that the human gaze heatmaps focused on and high activations for boundaries which human visual attention focused much less on. As observed in the results section, the other large discrepancy between the two sets of heatmaps were images that primarily consisted of thick areas of vegetation. The CNN model showed comparatively higher levels of activation throughout much of these images, whereas humans found the images devoid of visual features to assess. This implies that the CNN model can notice textural markers of GAN generation without the need for clearly defined, isolated features that are important for human visual detection.

5.5.1 Limitations and Future Work

Both the human eye tracking study and comparison with a CNN detection model produced interesting insights into the differences between different detection methods regarding GAN generated EO detection. Despite this, there are limitations with the work presented that could be addressed in future work aiming to implement a similar experimental design. As previously mentioned, the eye tracking accuracy and refresh rate was limited to that of the hardware available, a Tobii Nano Pro with a refresh rate of 60 Hz, accuracy of 0.3°and latency of 17ms. It may be beneficial for future studies to use hardware with a higher accuracy and refresh rates to improve the quality of the gaze data obtained.

Another suggestion for follow up studies would be to scale up the population sample used. As well as testing between self-reported experience groups from different university populations, future work could look at finding differences between larger groups of participants from various technical roles such as data scientists, earth observation engineers and remote sensing analysts. Further data surrounding the participants own reasons should also be collected such as a qualitative survey after task completion.

The work here also focused exclusively on EO imagery but the methodology and experimental setup could also be applied to other potential areas that could be potentially vulnerable to GAN image generation technology.

5.6 Conclusion

The work in this chapter measured human gaze and CNN attention towards the detection of EO synthetic aerial imagery. The results established that for the visual detection of GAN generated EO imagery there was a positive correlation between self-reported prior experience and detection accuracy. Through the use of both empirical evaluation metrics (gaze entropy, accuracy, response time) and visual observation (gaze heatmaps) differences were observed in the search strategies used between expert and non-expert groups, with more experienced individuals using more effective and efficient detection methods. It was also found that despite similar levels of overall accuracy on the real/synthetic EO dataset used, there are differences in the key features used for detection between human and CNN based GAN detection methods. The implications for this being that both detection methods, computational and human, use differing visual-spatial features in the images that signify whether the image is real or generated. Future work should follow on from the results found here and investigate novel methods which take advantage of both detection methods studied. Further studies should explore ways to utilize these techniques for a more comprehensive approach to synthetic EO image detection.

Chapter 6

GAN Generated EO Image Detection with Human Gaze Guidance

6.1 Introduction

The previous chapter explored both human and computational visual attention using a mixed methods approach comparing the eye tracking results from a real or fake detection task to that of gradient weighted class activation maps of a CNN based detection network. Through a qualitative comparison of the human gaze heatmaps and CNN detection network GradCAM images of corresponding GAN samples, differences were observed in ROIs across the images, suggesting that the two different detection systems were assigning different visual cues to aid detection. The results also found that experience with EO images was a predictor of GAN image detection ability, with different experienced groups exhibiting different search strategies.

These insights, that human experts show more effective detection strategies than novices and that there are visual differences in attention between human and CNN methods, forms the basis for the work presented in this chapter. The research in this chapter uses the obtained gaze heatmaps from the high expertise group in the previous study, together with their corresponding images to improve the performance

of computational methods by guiding the learning of detection models. This idea of using guided attention to improve image classification networks is supported by previous literature looking at similar methods. One examples is a 2017 study [173] that trained an instance of the CNN model AlexNet to give more human-like prediction behaviour. In this study the researchers fine-tuned the pretrained CNN model on just the ROIs of the training images which were focused on by humans, as obtained from an eye tracking study in a similar setup as the study detailed in Chapter 5 but without the real/fake detection element.

The aim of the work in this chapter is to improve on current CNN detection models for the detection of GAN generated EO images through guided attention using domain specific expertise. Building up on the research conducted in the previous chapters that looked at the generation and different methods of detection of GAN generated imagery, the objective is to combine human and computational detection methods for a single model.

The experiments in this chapter aim to achieve these objectives by using human gaze together with a current SotA GAN detection model. The model used is a pretrained ResNet-50 model [45]. This model has been used throughout this thesis as the benchmark standard for GAN image detection. As a general and non-domain specific GAN image detection model, it performs well, particularly on popular image benchmark datasets such as ImageNet. Despite this, in Chapter 3, it was found that for EO domain specific image data, the detection model performs considerably worse than previous benchmarks for ImageNet and similar datasets. Notably, there was a high false negative rate for correctly classifying synthetic images. This meant that as the accuracy for correct classifications of synthetic images was significantly lower than classifications for real images, it would predict most images as being real, whether synthetic or not. It was also found that using Grad-CAM, the areas of significance differed to the areas that were focused on by humans, as visualized using gaze heatmaps. The experiments in this study focus on improving the pretrained CNN detection model through transfer learning using gaze heatmaps to guide the attention of the model towards areas considered important for human detection.

Based on both the previous literature for synthetic image detection and the results from the preceding chapters the hypothesis is as follows:

H_1 : CNN based EO image detection accuracy will show improved performance when guided by expert human gaze data during model training, compared to models trained without.

The model proposed in this chapter consists of two parts. Firstly, a convolutional U-Net model trained on human gaze data to produce attention masks for unlabelled data. The second part is a pretrained ResNet model [45] which is then finetuned on paired real/synthetic EO image data with corresponding expert gaze data. This results in an end to end model that takes in an EO image (real or synthetic), generates a corresponding attention mask before giving a real or synthetic prediction (Figure 6.2).

6.1.1 ResNet Architecture

The generalized CNN detection model[45] used throughout this paper uses an implementation of the ResNet architecture that has been specifically trained to predict the occurrence of GAN generated images in an input dataset. ResNets, short for Residual Networks[57], refer to a specific class of deep convolutional neural network models that are characterised by their use of a very deep structure of layers and their use of residual blocks. The residual block is a block of network layers which utilizes "skip connections" to facilitate the training of such large models. Skip connections provide a path for information to flow directly from lower layers of the network to higher layers of the network, skipping out any intermediate layers (hence the name). Skip connections are calculated by $H(x) = F(x) + x$, where $F(x)$ is the output of the layers inside the residual block, calculated from the input x. Skip connections ensure that important features learned in the lower parts of the network is not lost as the information flows to the higher layers and avoids the notorious "vanishing

gradient problem" that can be an issue in networks using > 20 layers.

The use of residual blocks throughout the final networks results in a robust model that can learn many complex patterns that can be useful for binary and multi class classification. These attributes have resulted in ResNets being one of the most popular CNN models to use for a variety of tasks and often form the backbone of the image classification component of larger ensemble models.

As ResNets require a large amount of data for training they are often pretrained on large benchmark datasets such as ImageNet then adapted to a target task or dataset through the process of fine-tuning or transfer learning. This enables the network to be used on smaller datasets.

6.1.2 UNet Architecture

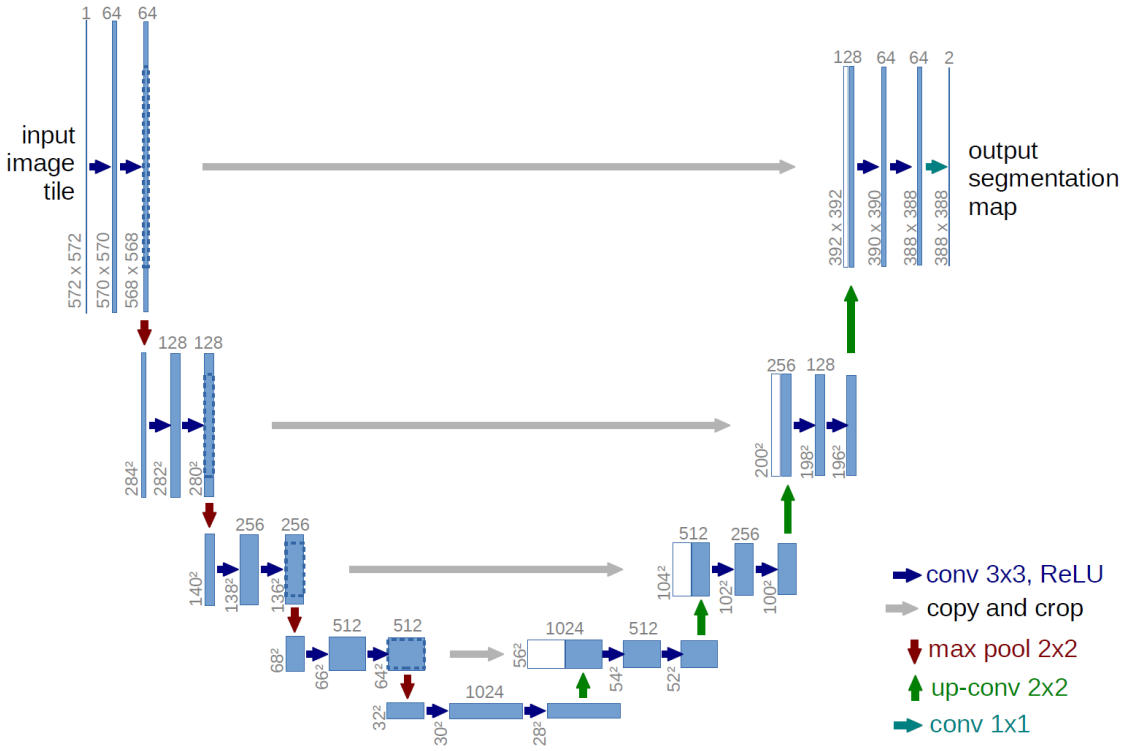

Figure 6.1: *UNet architecture[174]*

The U-net model architecture was created for the purpose of biomedical image segmentation[174]. The network consists of a convolutional neural network arranged in a symmetrical encoder-decoder configuration. The network takes in an input image and outputs a segmentation image of the same size via a series of contracting

then up sampling convolutional blocks. A series of contracting blocks in first half of the network each consist of two 3x3 convolutions using the ReLU activation function and a max pooling layer. The expanding second half of the network uses up sampling convolution blocks with 2x2 up-convolution operations, before two 3x3 convolution blocks, mirroring the contracting section. Throughout the model the feature maps in the contracting section are concatenated to the corresponding feature maps in the expanding section. A final 1x1 convolution is applied to reduce the feature map to the number of channels relevant for the final output image. For classification tasks like in the original implementation, the U-Net then uses a pixel-wise loss function such as cross entropy loss to train the network on the backward pass via backpropagation.

The U-Net model has been widely used for binary and multiclass segmentation for a variety of different applications, for creating segmentation maps for medical imagery. The network lends itself well to this field as it is capable of learning accurate end-to-end image segmentation on only small amounts of data. This makes it a useful model for tasks where only a small training set of masks and images are available.

6.2 Experiments

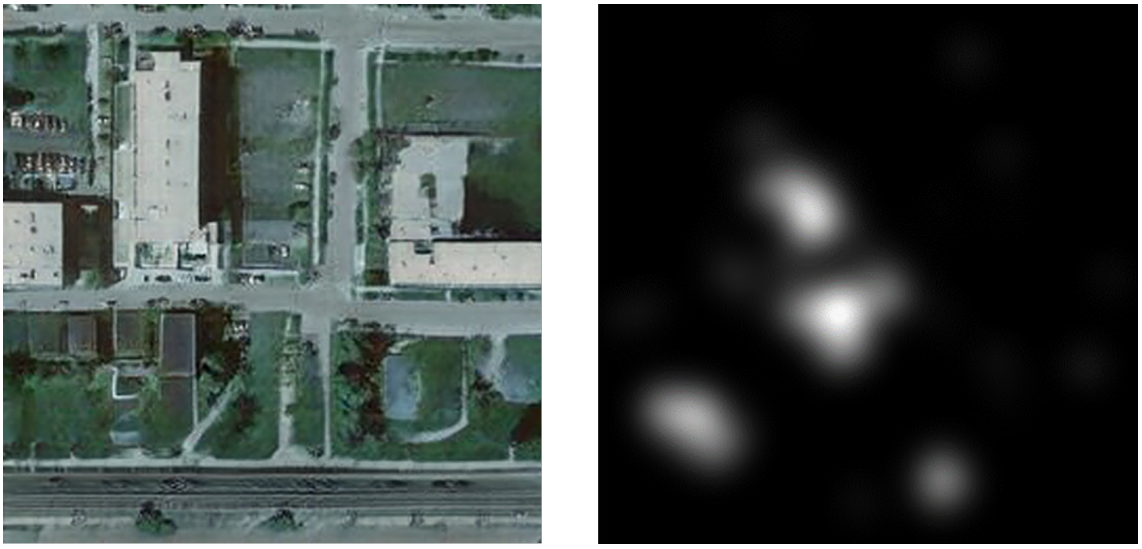

Figure 6.2: *Examples of training image + gaze mask*

6.2.1 ResNet training

A pre-trained CNN detection model using a modified ResNet-50 that took in a 4-channel input image was fine-tuned on the image data obtained in the eye tracking study from the previous chapter. This data consisted of 3200 real (1600) and synthetic aerial images (1600) that had been presented to the participants in the study. Each image was then concatenated with a grayscale gaze heatmap made up of the areas of visual fixation of participants from the high/moderate experience groups. Only images/heatmaps from trials where the participant had correctly identified the generated image were included. The intuition behind this was that the addition of the heat maps would improve the performance of the ResNet classifier during training. The heatmaps would by used to guide the network's attention towards ROIs in the images which were of most importance to visual attention.

To increase the variation in the dataset horizontal and vertical transformations were applied to each image/heatmap input image in addition to image normalization to improve training performance. The model was then trained for 9 epochs (until convergence) with a binary cross-entropy loss function together with a sigmoid layer to obtain the final output prediction.

Two different approaches to model training were tested. The initial approach fine tuned the pretrained ResNet model on the image/heatmap dataset with all trainable parameters available. This was found to led to very poor results as the dataset was too small for the large ResNet-50 model. The second approach, which was used for the final complete model, only fine tuned the initial layer and the final convolutional block of the ResNet model, with the rest of the weights frozen. This is an often used solution to problems with small datasets as it increases training performance and also preserves many of the learnt features from the original pretrained weights. In this process of fine-tuning, the model weights were frozen apart from the input layer, to account for the new input channel (grayscale mask channel), and the final convolutional block. The last convolutional black was left unfrozen (trainable) as generally the higher level features and details of classes are encoded here.

Performance of the final model was tested on two datasets of real/synthetic EO imagery generated from StyleGAN2 as detailed in Chapter 2, one consisting of unseen images from the Inria dataset and the other from OpenCities. Both datasets contained 20K images (10K real/ 10k StyleGAN2) and did not have corresponding human gaze data unlike the training set.

6.2.2 UNet Attention Map Generation

One issue with the ResNet fine tuning and transfer learning experiments is that the networks are trained on 4-channel input data (RGB images + Grayscale gaze attention maps) but are then tested on 3-channel input. This requires removing an input channel in the trained network which results in the final detection model trying to make a prediction on RGB images when the task it has been optimized for is making a prediction on an image + heatmap. This is particularly important as in real world detection scenarios the model would be used for data which does not have any corresponding, expert gaze data available. To solve this problem a U-Net model was trained to generate attention masks for unpaired images, based on expert gaze heatmaps.

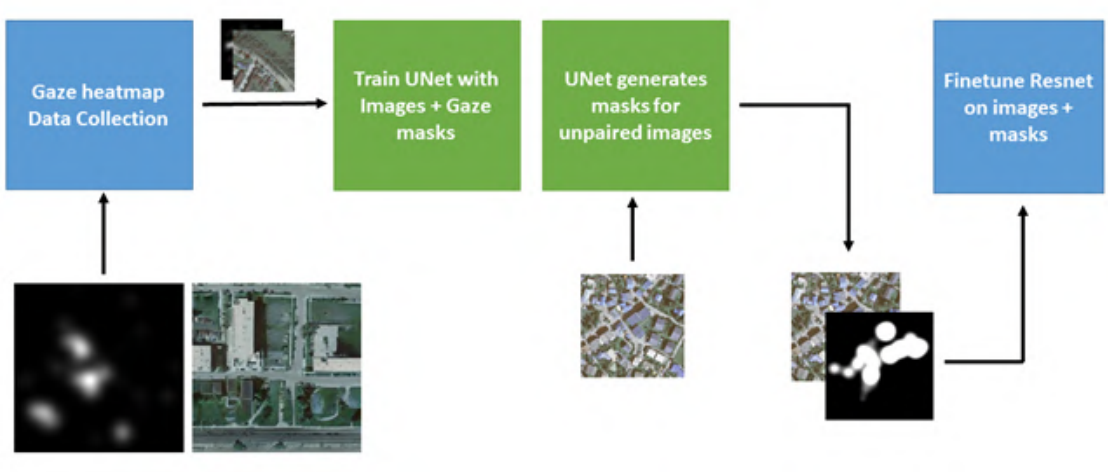

Figure 6.3: *Implementation of expert gaze guided UNet-Resnet synthetic EO detection model*

The U-Net model was trained on the image + heatmap dataset for the task of producing a heatmap based on an input RGB image. The structure of the U-Net

used can be seen in Figure 6.1 and the full model summary can be found in the Appendices (Figure C.2).

The reasoning for this implementation is that the U-Net model would be able to learn areas of the image that correlate with human visual attention and generate corresponding attention maps. These newly generated attention maps would then be paired with the original image (real/fake) and passed through the fine-tuned 4-channel ResNet detection model for a final real/synthetic prediction (Figure 6.3). This results in a complete end-to-end detection model without the need to remove any trained layers of the ResNet model to accommodate for RGB images that are not paired with a human gaze attention map.

The U-Net model was trained on the same dataset as used in the ResNet experiments, with a dataset consisting of 3200 image/heatmap pairs (512x512). The pixelwise loss function used was changed from cross entropy loss to MSE loss to accommodate for the change in task from binary classification of pixels (segmentation) to a regression task of estimating greyscale pixel values of the output image given the input image.

6.3 Results

The results for three ResNet models were compared. The first model is the Wang et al. (2019) general GAN detection model using the weights from the official implementation. This model had been trained on a variety of GAN (PGGAN) generated images from benchmark datasets with different levels of guassian blur added to the images. This was tested against our model which added a UNet attention map generator to the pretrained detection model for further training using paired images (image + attention maps). This test was controlled with a control model which took the Wang et al. (2019) pretrained model and applied the normal finetuning process with real and synthetic EO images (no attention masks/gaze maps).

The control model was added to accurately assess the impact of the human gaze heatmaps on detection performance. Without a control model it would not be pos-

sible to tell if any improvements to performance are due to the addition of heatmaps or from just being trained with EO imagery alone, as this is likely to improve the model's performance on similar data regardless of human gaze. Both the proposed model and control model were trained on the the same 3200 image dataset (1600 real, 1600 synthetic) for a fair comparison. The use of this dataset allows for the evaluation to show if training with the use of gaze heatmaps is more impactful than simply fine-tuning on a large amount of unpaired EO data. If the models were both trained on additional images (e.g. 10K images from the Inria dataset), it would be hard to assess the impact of the small amount of expert labelled data on the prediction accuracy. It is well known that training on large amounts of data can vastly increase prediction performance, but the objective here is to see if gains can be made on a small amount of paired data.

It is important to note that while the model was only trained on this small labelled subset, the experiments for model evaluation used larger (20K images) EO/GAN datasets (OpenCities, Inria) that were unseen by the models.

6.3.1 UNet Mask Generation

The UNet model trained on paired data of EO images (real and synthetic) and expert human gaze data was able to produce attention masks for the images to some success. Looking at examples from the test images (figure 6.4) the trained network could not accurately replicate the real gaze heatmaps when given the corresponding EO image as in input. This was to be expected as only a relatively small number of paired training images (3200 image pairs) were available for training. Due to the nature of the image and gaze data having non uniform features between different samples, it makes it a difficult task for the network to be able to accurately predict a $1:1$ mask. Unlike in image segmentation tasks, this is not necessarily an issue the masks only need to guide the classifiers attention to areas likely to contain image features important to correct predictions.

Another noticeable discrepancy between the real gaze images and the generated

masks are the image artifacts are the lines across the images that are the remains of features in the input EO images, these could also be caused by the upsampling convolutional layers present in the network. Although these UNet results could possibly be improved with larger amounts of trained data, they do manage to produce masks similar to gaze maps that can be used as input for the ResNet section of the proposed detection model.

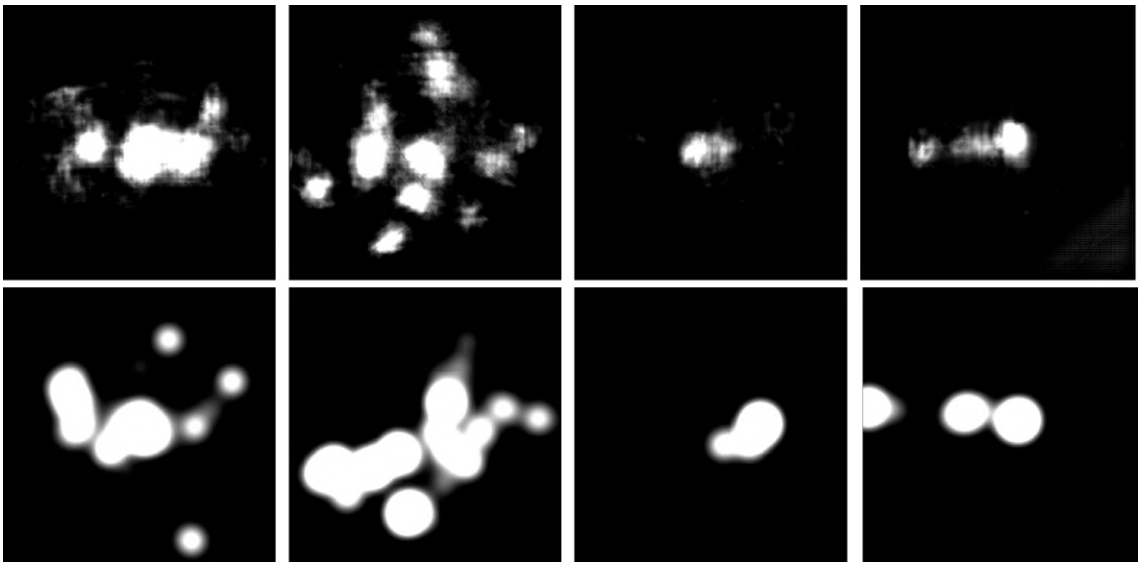

Figure 6.4: *Samples of Unet Masks (top row) generated from true expert gaze masks (Bottom Row)*

6.3.2 Model Detection Performance

Model performance was evaluated using the same datasets as previously used in the first image detection study (chapter 4) for evaluating the CNN detection model before fine-tuning. These datasets consisted of a larger, unseen segment of the Inria benchmark dataset and a segment of the OpenCities dataset. Each test dataset was made up of 10000 real images and 10000 StyleGAN2 generated images based on the respective dataset. The 4-channel models trained with the additional input of human gaze heatmaps were evaluated on each dataset with corresponding mask images generated via the UNet mask generation model. The control model was evaluated using RGB images only.

Using only 3200 images (1600 real, 1600 GAN) with gaze points, the detection model

Model	AP	ACC	ACC (Real)	ACC (Fake)	F1
Wang (2019)	93.28	65.77	99.64	31.90	0.48
Control (no masks)	91.00	63.58	99.75	27.41	0.43
Our Model	95.93	87.54	95.25	79.83	0.87

Table 6.1: Results for StyleGAN2/Open Cities real/fake dataset (20K images). Evaluation metrics are average precision (AP), total accuracy (ACC), accuracy for each category (ACC real/fake) and F1 score

Model	AP	ACC	ACC (Real)	ACC (Fake)	F1
Wang (2019)	94.26	73.42	98.34	46.51	0.63
Control (no masks)	98.02	79.41	99.91	58.91	0.74
Our Model	97.24	85.02	80.04	71.96	0.83

Table 6.2: Results for StyleGAN2/Inria Aerial real/fake dataset (20K images). Evaluation metrics are average precision (AP), total accuracy (ACC), accuracy for each category (ACC real/fake) and F1 score

was able to outperform both the generalized detection network and the control model. The final detection model performed better on both test datasets with a higher AP and ACC (Table 6.1). Although this overall performance was better than previous models, our detection model performed slightly lower than the general CNN detection model at correct predictions for real images, despite vastly improving on prediction accuracy for GAN images.

This can be more clearly seen in the F1 score comparison between models on both the OpenCities dataset (Table 6.1) and the Inria dataset (Table 6.2). This indicates a high overall increase in correct detection. Our model fixes the issues with incorrectly detecting most images as real as seen in the other models. This is likely due to the additional input of the attention maps which highlight areas of the images more likely to contain features indicating whether the image is real or synthetic. Despite the generated attention maps not replicating the real expert gaze maps $1:1$, they were still able to mark areas with a high probability of containing defining image features for correct detection, as evidenced by the increased performance of the final end-to-end model.

All models were further tested in their detection performance on urban and rural classes for both datasets, Inria (Table 6.3) and OpenCities (Table 6.4). Similar to the results found in Chapter 3, which looked at inception scores for generated urban

Model	Image Class	AP	ACC	ACC (real)	ACC (fake)	F1
Wang (2019)	Urban	92.81	54.24	99.97	14.09	0.24
Wang (2019)	Rural	97.62	69.12	99.91	35.77	0.56
Control (no masks)	Urban	85.63	63.28	95.65	34.85	0.50
Control (no masks)	Rural	98.62	78.06	99.83	54.48	0.70
Our Model	Urban	92.81	66.77	99.24	38.25	0.55
Our Model	Rural	99.78	95.15	99.69	90.24	0.95

Table 6.3: Breakdown of model performance for urban and rural images classes from the StyleGAN2/Inria dataset (20K images)

Model	Image Class	AP	ACC	ACC (real)	ACC (fake)	F1
Wang (2019)	Urban	87.75	59.98	99.64	24.80	0.39
Wang (2019)	Rural	93.47	67.98	99.83	29.96	0.46
Control (no masks)	Urban	90.60	61.05	99.22	27.17	0.42
Control (no masks)	Rural	95.43	71.00	99.88	36.52	0.53
Our Model	Urban	92.58	80.61	94.43	68.34	0.78
Our Model	Rural	98.11	93.64	95.81	91.05	0.92

Table 6.4: Breakdown of model performance for urban and rural images classes from the StyleGAN2/OpenCities dataset (20K images)

and rural images, all models performed higher on rural scenes than on urban scenes. Similar to the results for the overall datasets (Tables 6.2, 6.1), the original GAN detection model (Wang 2019) performed the worst and the gaze assisted model (Our model) achieved the best performance. Despite still surpassing the previous model (Wang 2019) and the control model, the gaze-assisted model can still be seen to struggle with fake urban scenes, especially with the low accuracy of 54% (Table 6.3) in the Inria results. This suggests that the improvements from the expert gaze data do not come from the differences in attention between land types and instead more general differences between human and computational gaze methods. This is further seen in the evaluation of the differences between F1 scores (Table 6.5), which shows no pattern of relative class improvements between the models.

6.4 Discussion

In this Chapter gaze data from expert visual detection of real and synthetic EO images was used to guide attention in a CNN based image classifier. The results (Table 6.2) support the hypothesis that expert visual gaze data can be used to

Model	Dataset	F1 Urban	F1 Rural	F1 difference (%)
Wang (2019)	Inria	0.24	0.56	80
Control (no masks)	Inria	0.50	0.70	33
Our Model	Inria	0.55	0.95	53
Wang (2019)	Open Cities	0.39	0.46	16
Control (no masks)	Open Cities	0.42	0.53	23
Our Model	Open Cities	0.55	0.95	16

Table 6.5: Percentage difference between urban and rural F1 scores on both test dataset (Inria, Open Cities)

improve the performance of CNN based image detection models. This new method improves on previous models for prediction accuracy on real/synthetic EO images. By training a U-Net encoder network on a small number of images with gaze data, synthetic gaze masks can be generated for unpaired datasets allowing for further fine tuning of the network without the need to conduct additional studies.

The samples from the UNet generation show that the model was not able to produce completely accurate masks that replicate human gaze data. This could be due to the small amount of labelled samples which due to the nature of the image data varied greatly in ROIs between the samples. Despite this, the results from the end-to-end detection model demonstrate that the additional input of the generated masks into the ResNet model did significantly increase detection performance.

The success of combining detection methods, both human and computational supports the work throughout this thesis that a mixed methods approach is necessary to establish novel solutions to the generation of synthetic EO imagery. Rather than rely completely on algorithmic solutions, domain expertise is a valuable asset to improving current methods and should continue to be considered in further detection methods. this is supported by similar trends in other specialised fields such as medical imaging which have also found success in incorporating domain expertise with SotA deep learning models[175–177].

In this chapter the proposed model is trained to detect GAN EO images and uses UNet and pretrained ResNet architectures. Despite these details the approach used could also be replicated to improve other models and applied to different types of image data. The results from this demonstration show the effectiveness of using expert

human gaze data to improve classifier performance and could quickly be applied to the detection of images generated from other methods such as Diffusion and transformer models. Expert gaze data could also be collected for other domain specific image types that may be vulnerable to misinformation. In a rapidly advancing field where specific architectures quickly become outdated and new, more sophisticated image generation techniques are developed, this work provides methods that can be adapted for such changes and advancements.

6.4.1 Limitations and Future Work

Although the detection model implemented in this chapter increased the overall accuracy and precision of previous methods, it did not perform as well at the accurate prediction of real images. Although it is unclear why this the case, there are a few factors which could potentially explain this drop in performance. One potential reason for this is due to the procedures followed for the data collected in the previous chapter. The gaze data that was used for training was collected from participants in an image detection study where the task was to correctly identify which image was synthetic out of a series of real/synthetic image pairs. By asking the participant to identify the synthetic image could influence the behaviour of the participants to focus their attention more on images that they considered to appear synthetic rather than more realistic looking images. This could lead to less gaze data being collected overall for real images. Additionally, the ROIs in these images may not necessarily correlate with where the images look most real but instead, look synthetic. This could lead to less useful gaze data to be paired with the real images, lowering their effectiveness.

Future data collection should aim to investigate this by conducting an additional series of studies where participants are asked to identify the real image instead of the synthetic image to balance the current set of gaze data. Further research should also investigate using a different study design for data collection, such as showing the images in a sequential sequence rather than as pairs. The paired design was

chosen initially for reasons described in Chapter 4, however, for the sole objective of collecting gaze data for creating image masks this may be a better choice for future studies. This way there is less bias between images as they are shown individually which could led to higher quality gaze heatmaps.

Another advantage of a larger scale data collection would be the increased number of paired samples available to train the UNet mask generation network. As discussed, the masks did not manage to fully replicate the human gaze maps, despite encoding enough information to generate attention maps useful for training. These results could however, be further improved using a larger dataset to produce more accurate gaze maps which would likely result in increased accuracy on the end-to-end model. In addition to expanding the experimental design of the data collection, future work should look at the value of training the model on a larger scale with a higher quantity of image/gaze mask pairs across multiple real/fake EO image datasets. Further studies should also employ novel methods to further investigate the precise differences in human and computational detection. As it was found in this chapter that detection performance improved overall, but there were still differences between urban and rural prediction accuracy.

6.5 Conclusion

The work in this chapter proposes a human expert guided, synthetic EO image detection model. The final detection model improves on accuracy and precision of previous methods and demonstrates how implicit domain knowledge from human experts can be integrated with computational detection techniques for better detection of GAN generated EO images. The model was able to achieve high accuracy after only being trained on a small, paired image dataset ($N = 3200$), demonstrating that this is a viable method in scenarios where it is hard or costly to obtain substantial amounts of expert gaze data or synthetic examples.

In the preceding chapters, it was found that there were difference in performance and behaviour of humans and CNN based detection methods in differentiating between

real and GAN generated EO images. The contributions in this chapter demonstrate an effective methodology of combining these different detection systems to further improve current detection methods.

Chapter 7

Conclusion

As image generation techniques, such as generative adversarial networks (GANs) become more advanced and applied to a wider range of generation tasks and domains, there is a growing need for the development of novel detection techniques. EO imagery, such as satellite imagery is one such area which could be potentially vulnerable to the misuse of image generation.

The research in this thesis investigated the generation and detection of GAN synthesised EO imagery. Following a benchmark evaluation of current GAN methods, studies were conducted to compare differences between automated and human detection methods. The effects of expertise was also measured for human detection of synthetic EO imagery with expert/novice experiments. The results from this research led to the implementation of a new expert guided GAN image detection classifier model.

7.1 Main Contributions

7.1.1 Chapter 3: GAN benchmark comparison

As discussed in the literature review (chapter 2), limited work currently exists for measuring GAN performance in generating EO imagery, compared to other image type such as faces and objects. New GAN models are commonly tested within a small range of benchmark datasets (e.g. ImageNet, CIFAR-10, CelebA), leaving

questions about how they may perform on other, domain-specific image types.

The work in the initial chapter of research evaluated the capabilities of different GAN models for the task of generating synthetic EO imagery and benchmarked them on the Inria Aerial Dataset. The benchmark results found that StyleGAN2 was the most successful at generating high quality synthetic aerial imagery when measured using the evaluation metrics FID and KID. The findings of this comparison show that GAN EO imagery has reached the level that could present a possible security concern in matters relating to authenticity.

In addition to the GAN benchmark evaluation, the effects of how FID is calculated were also explored. Many GAN evaluation papers report FID that has been calculated using an InceptionNet implementation trained on the ImageNet dataset. Although using a standardised evaluation model can be helpful for making comparisions between models and paper results, it may present problems and biases when applied to image classes not covered in the ImageNet dataset. To investigate this, FID calculated using ImageNet was then compared to a novel variant of FID calculated using a network trained on the OpenCities dataset. The results found that EO samples from StyleGAN2 were scored lower (better) for the FID that was based on the OpenCities dataset rather than the standard ImageNet variant. This contributes to the growing amount of evidence that FID should not be used as an empirical measure of quality, particularly in comparing different models as it is heavily influenced by the types of data the core Inception network has been trained on.

In summary the main contributions for this chapter were:

1. The evaluation of GAN models in the domain specific context of EO imagery.

2. The comparison of FID scores for different instances of the Inception V3 model

3. A real/fake dataset based on the Inria Aerial Benchmark Dataset [102] for EO imagery produced using StyleGAN2

7.1.2 Chapter 4: GAN image detection study

Whereas the work done in the first chapter focused on GAN generation of EO image data, this chapter evaluated human and computational detection of generated EO image data. Automated metrics such as FID are useful for performing quick evaluations of model performance but are not representative of how realistic they may appear to humans. Additionally, synthetic images may be used to fool humans and necessarily automated systems, necessitating user studies to evaluate human detection on synthetic EO imagery.

Using the StyleGAN2/Inria dataset produced in the first chapter, an online image detection study was conducted to evaluate the difficulty in distinguishing between real and fake image pairs. The study found that participants achieved varying levels of accuracy correlated with their self-reported previous experience with dealing with similar images (GAN or Satellite images). The study also found that images with higher amounts of rural features such as trees, fields or foliage were much harder to correctly classify than those with urban features (roads, buildings, and other infrastructure). When comparing the FID of rural and urban images with participant detection scores it was found that despite lower FIDs (lower is better) for urban scenes, these were easier for humans to identify, with the reverse relationship found for rural scenes.

A State-of-the-art generalised GAN detection model was also evaluated against the same dataset as used in the image detection study. It was found that the detection model achieved lower levels of accuracy in predicting if an image was real or fake for the StyleGAN2/Inria dataset than on the benchmark datasets that were reported in its original implementation. Although it performed well on classifying real images, the CNN detection model had particularly low accuracy when presented with synthetic images, and overall achieved a slightly lower score than the high and moderate expertise groups from the human detection study.

In summary the main contributions for this chapter were:

1. Detection accuracy for synthetic aerial imagery varies between groups of pre-

vious experience.

2. Rural images are harder for humans to distinguish as real or synthetic than urban images.

3. FID for GAN generated EO images is inconsistent with human detection accuracy

4. Synthetic EO image data presents a novel challenge for current computational detection methods trained only on popular object and face datasets.

7.1.3 Chapter 5: Human Gaze and CNN attention in Detecting GAN generated EO Images

The image detection study in the previous chapter explored the differences in performance between computational and human detection methods and also found a correlation between experience and synthetic EO detection accuracy. The work in this chapter further investigated these findings through another image detection study including an additional modality of eye tracking. The results were evaluated using gaze entropy metrics and a qualitative visual analysis of gaze heatmaps. The study found that significant differences in the gaze behaviour between self-reported experts and novices when making predictions on real/synthetic EO image pairs. The participants with higher experience made more accurate predictions and showed overall lower gaze entropy, and shorter response times than the participants with low experience. This suggests that the more experienced groups were using more effective and efficient search strategies for making correct detections. These findings were also supported by a visual inspection of the gaze heatmaps.

The CNN detection model evaluated in the previous chapter was also further explored using Grad-CAM to visualize the ROIs that produced higher class specific activations. Both the Grad-CAM heatmaps and gaze heatmaps were then compared as correlates of visual attention for each respective detection method. This qualitative analysis noted that the CNN detection model gave higher significance

to different ROIs for classification than humans, particularly in rural areas and for boundary lines.

In summary the main contributions for this chapter were:

1. Further support for the previous findings that self-reported experience is a correlating factor for synthetic EO image detection ability

2. Experts and novices exhibit differences in gaze behaviour for detecting synthetic EO image data, with experts performing more efficient and accurate visual search strategies.

3. Humans and CNN based detection methods prioritise different ROIs in real/synthetic EO images.

7.1.4 Chapter 6: GAN Generated EO Image Detection with Human Gaze Guidance

Building on the research from the previous chapters the aim of this chapter was to improve current detection methods for GAN generated EO images. This aim was achieved by the implementation of a classifier model guided by expert gaze data. The final model consisted of two sections. The first section of the model used a convolutional UNet architecture that was trained on 3200 real/StyleGAN2 EO images and corresponding grayscale masks of expert gaze data collected in the previous study. Using the images as input the model was then trained to generate corresponding attention masks, using the real gaze masks as the ground truth data. The second section of the model consisted of a ResNet model, modified to accept 4-channel input data and used to classify an image as real or synthetic. For increased model performance the model was pretrained using the weights from a current, generalized detection model [110]. These two sections were then assembled to give the final detection model which would accept an input image, produce an attention map based on human gaze data, concatenate both together before passing it to the ResNet classifier to give a prediction on whether the image was real or synthetic.

The final attention guided detection model was tested on two EO GAN datasets produced in Chapter 1 and based on the Inria Aerial Benchmark dataset and the OpenCities Datasets. To assess the impact of the human gaze data on the classifier performance a control model was also trained on just the 3200 training images and without the addition of the gaze heatmaps, the generalized GAN detection model was also compared. The results found that the attention guided model gave a higher overall accuracy and average precision in detecting synthetic EO images than both the previous GAN detection model and the control model.
In summary the main contributions for this chapter were:

1. A human attention guided synthetic EO image detection model that fine-tuned a GAN classifier using paired image and attention masks generated from heatmaps of expert gaze data.

2. Improvements over current methods for synthetic EO image detection using only a small paired image training dataset

3. Support for the importance of mixed methodologies and the use of domain specific knowledge for implanting future detection models.

7.2 Impact and Implications

In this thesis I have presented an interdisciplinary, mixed methods approach to addressing some of the current concerns that the advancement of generation models presents in relation to the field of EO image data. The first major implication of this work is bringing to attention evidence that image generation techniques have reached a level of sophistication where they can unconditionally produce synthetic EO images to a highly realistic standard. Despite the threats to data security and authenticity, there is still a lack of literature focused on domain specific detection techniques. The work of this thesis has attempted to fill in some of these gaps in current knowledge by looking at human and computational detection methods in a domain specific context. Doing so has found that both detection methodologies

have different advantages and limitations. Future work in this field of research should continue to explore mixed method approaches to detection to fully utilize the different strengths each method provides. The final detection model implemented in chapter 6 is an example of this, as it uses human gaze data to guide the decision making of a CNN based classifier to give more accurate predictions than either a classifier or human alone.

Another outcome from this project is highlighting the importance of utilizing domain specific knowledge for more specific detection methods. The image detection studies conducted in chapters 4 and 5 found that previous experience is a correlating factor with detection performance and experts exhibit different and more effective search strategies than non-experts. This is important to consider for future work that can leverage this experience to improve detection methods (similar to this project) or help train non-experts to be better at distinguishing between real and synthetic images themselves.

In relation to the wider impact on defence and security, the work in this thesis suggests that further research into domain specific detection methods should continue to be prioritised to keep up with the rapidly evolving array of image generation methods. Even over the timespan of this project the landscape of image generation has evolved, with the surge in popularity of Diffusion and Transformer models trained on vast amounts of data presenting new challenges towards image data authenticity. Despite the constant evolution of generation models, the methods and experimental designs depicted in this thesis can provide templates for future detection research into new and emerging generative techniques.

7.3 Limitations and Suggestions for Further Research

The work in this Thesis presents several contributions to EO image GAN detection but also has several limitations of note that are important to note for future

research in this field. Although significant findings were present in the results of the human participant image detection studies, they would have benefited greatly from higher numbers of participants and from a wider range of populations. If similar experiments were to be conducted in the future it would interesting to explore how detection abilities and behaviours differ between different domain expert groups such as geospatial scientists and data scientists, and how occupation correlates with detection accuracy when compared to self-reported expertise.

This thesis has focused on RGB satellite image data, but similar experiments could be applied to other forms such as infrared and multi-spectral data. It is possible that the detection capabilities of current methods have different levels of performance between different EO data types and further work is needed to assess which areas are more vulnerable than others.

GAN models were the primary image generation architecture that was chosen to be the focus of this project, this was due to them being at the top of the image generation field in terms of both their performance and widespread popularity. Over the course of this project and in particular the final year of work, the image generation landscape has changed significantly with the explosion of large Diffusion models such a DALL-E2[37] and Stable Diffusion[36]. If the current trends continue, then it is likely that they will continue to improve and become the dominant models for generative tasks. It is therefore important that future research takes this into account to further detection methods in relation to these models. Although this project has focused on GANs, the experimental designs and techniques used could also be applied to these newer architectures.

Other limitations on the work presented here are partly due to circumstances. The 2019 COVID-19 pandemic caused major disruptions and delays to research in particular the ability to hold lab-based experiments. While the studies were still able to be completed and resulted in meaningful contributions, the scope and design of the studies had to consider the limitations at the time, such as conducting the first study online rather than face to face. A face to face experiment would have been beneficial as it would allow for more control over extraneous variables and higher

participant engagement with the task. Converting the original study to be ran online was also time consuming and caused unforeseen delays to the work. Despite these issues, once the setup was completed, running the study online allowed for the collection of a much higher sample size than originally expected when planning for the initial face to face experiment.

7.4 Summary

This PhD thesis has explored the generation and detection of GAN generated EO images. The capabilities of a range of GAN models were bench marked on EO image datasets, with the best performing model (StyleGAN2) being used to create real/synthetic EO datasets. Using these datasets, different visuospatial detection methods were evaluated, CNN based methods and also human visual perception. These methods were then compared and contrasted within pair of image detection studies which also considered the differences between more and less experienced participant groups using gaze metrics. Finally, the results of these studies were used to inform an improved real/fake image classifier specifically for GAN generated EO imagery.

The work in this thesis has established a link between previous experience in working with EO images and real/synthetic detection ability, both in terms of accuracy and gaze behaviour. The final detection model from this thesis also supports the hypothesis that domain specific knowledge in the form of gaze data can be used to improve the ability of CNN based image classifiers. The methodology described in this thesis also provides a template for further researching detection methods to counter current and future generative models. As AI driven image generation continues to improve, this will become ever more important over the coming years.

Bibliography

[1] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[2] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[3] J. Wang, W. Zhou, G.-J. Qi, Z. Fu, Q. Tian, and H. Li, "Transformation gan for unsupervised image synthesis and representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 472–481.

[4] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] R. Wu, G. Zhang, S. Lu, and T. Chen, "Cascade ef-gan: Progressive facial expression editing with local focuses," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] T. Balmforth, "Satellite images show new russian military deployments near ukraine," Feb 2022. [Online]. Available: www.reuters.com/world/europe/satellite-images-show-new-russian-military-deployments-near-ukraine-2022-02-11/

[7] T. Allen, "Forensic analysis of satellite images released by the russian ministry of defense," May 2015. [Online]. Available: https://www.bellingcat.com/wp-content/uploads/2015/05/Forensic_analysis_of_satellite_images_EN.pdf

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," pp. 2672–2680, 2014.

[9] T. Zhang, H. Fu, Y. Zhao, J. Cheng, M. Guo, Z. Gu, B. Yang, Y. Xiao, S. Gao, and J. Liu, "Skrgan: Sketching-rendering unconditional generative adversarial networks for medical image synthesis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 777–785.

[10] S. Yang, Z. Wang, Z. Wang, N. Xu, J. Liu, and Z. Guo, "Controllable artistic text style transfer via shape-matching gan," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4442–4451.

[11] J. Daihong, Z. Sai, D. Lei, and D. Yueming, "Multi-scale generative adversarial network for image super-resolution," *Soft Computing*, vol. 26, no. 8, pp. 3631–3641, 2022.

[12] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, "Diffusion models for medical anomaly detection," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*. Springer, 2022, pp. 35–45.

[13] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[14] A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network: An overview of theory and applications," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100004, 2021.

[15] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, "Recent Progress on Generative Adversarial Networks (GANs): A Survey," *IEEE Access*, vol. 7, pp. 36 322–36 333, 2019.

[16] C.-T. Lin, S.-W. Huang, Y.-Y. Wu, and S.-H. Lai, "Gan-based day-to-night image style transfer for nighttime vehicle detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 951–963, 2020.

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, oct 2018.

[18] J. Vincent, "All of these faces are fake celebrities spawned by ai," Oct 2017.

[19] H. Thanh-Tung and T. Tran, "Catastrophic forgetting and mode collapse in gans," in *2020 international joint conference on neural networks (ijcnn)*. IEEE, 2020, pp. 1–10.

[20] D. Castelvecchi, "Can we open the black box of ai?" *Nature News*, vol. 538, no. 7623, p. 20, 2016.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[22] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[23] S. Vallender, "Calculation of the wasserstein distance between probability distributions on the line," *Theory of Probability & Its Applications*, vol. 18, no. 4, pp. 784–786, 1974.

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[25] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," dec 2019.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[28] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang *et al.*, "Cogview: Mastering text-to-image generation via transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19 822–19 835, 2021.

[29] C. Yang, Y. Pan, Y. Cao, and X. Lu, "Cnn-transformer hybrid architecture for early fire detection," in *International Conference on Artificial Neural Networks*. Springer, 2022, pp. 570–581.

[30] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[31] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[32] D. P. Kingma, M. Welling *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[33] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "f-vaegan-d2: A feature generating framework for any-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 275–10 284.

[34] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[35] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *arXiv preprint arXiv:2105.05233*, 2021.

[36] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

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

[38] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[39] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 852–863, 2021.

[40] S. J. Nightingale and H. Farid, "Ai-synthesized faces are indistinguishable from real faces and more trustworthy," *Proceedings of the National Academy of Sciences*, vol. 119, no. 8, p. e2120481119, 2022.

[41] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[42] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional cyclegan," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 282–297.

[43] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of GAN-Generated Fake Images over Social Networks," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, apr 2018, pp. 384–389.

[44] M. A. Villan, A. Kuruvilla, J. Paul, and E. P. Elias, "Fake image detection using machine learning," *IRACST-International Journal of Computer Science and Information Technology & Security (IJCSITS)*, 2017.

[45] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot...for now," in *CVPR*, 2020.

[46] N. Hulzebosch, S. Ibrahimi, and M. Worring, "Detecting cnn-generated facial images in real-world scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 642–643.

[47] L. Nataraj, T. M. Mohammed, B. Manjunath, S. Chandrasekaran, A. Flenner, J. H. Bappy, and A. K. Roy-Chowdhury, "Detecting gan generated fake images using co-occurrence matrices," *Electronic Imaging*, vol. 2019, no. 5, pp. 532–1, 2019.

[48] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "{TensorFlow}: a system for {Large-Scale} machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.

[49] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection," in *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, 2017, pp. 159–164.

[50] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[51] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[52] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do gans leave artificial fingerprints?" in *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 2019, pp. 506–511.

[53] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[54] B. Chen, X. Liu, Y. Zheng, G. Zhao, and Y.-Q. Shi, "A robust gan-generated face detection method based on dual-color spaces and an improved xception," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[55] S. Hu, Y. Li, and S. Lyu, "Exposing gan-generated faces using inconsistent corneal specular highlights," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2500–2504.

[56] X. Yang, Y. Li, H. Qi, and S. Lyu, "Exposing gan-synthesized faces using landmark locations," in *Proceedings of the ACM workshop on information hiding and multimedia security*, 2019, pp. 113–118.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[58] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? understanding properties that generalize," in *European Conference on Computer Vision*. Springer, 2020, pp. 103–120.

[59] A. Rai, "Explainable ai: From black box to glass box," *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 137–141, 2020.

[60] K. Chandrasegaran, N.-T. Tran, and N.-M. Cheung, "A closer look at fourier spectrum discrepancies for cnn-generated images detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7200–7209.

[61] C. Dong, A. Kumar, and E. Liu, "Think twice before detecting gan-generated fake images from their spectral domain imprints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7865–7874.

[62] V. Wesselkamp, K. Rieck, D. Arp, and E. Quiring, "Misleading deep-fake detection with gan fingerprints," *arXiv preprint arXiv:2205.12543*, 2022.

[63] S. Fan, T.-T. Ng, B. L. Koenig, J. S. Herberg, M. Jiang, Z. Shen, and Q. Zhao, "Image visual realism: From human perception to machine computation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 9, pp. 2180–2193, 2017.

[64] V. Schetinger, M. M. Oliveira, R. da Silva, and T. J. Carvalho, "Humans are easily fooled by digital images," *Computers & Graphics*, vol. 68, pp. 142–151, nov 2017.

[65] S. Zhou, M. Gordon, R. Krishna, A. Narcomey, L. F. Fei-Fei, and M. Bernstein, "Hype: A benchmark for human eye perceptual evaluation of generative models," *Advances in neural information processing systems*, vol. 32, 2019.

[66] "International Journal of Applied Earth Observation and Geoinformation — ScienceDirect.com by Elsevier — sciencedirect.com," https://www.sciencedirect.com/journal/international-journal-of-applied-earth-observation-and-geoinformation, [Accessed 06-Dec-2022].

[67] "Newcomers Earth Observation Guide — ESA Business Applications — business.esa.int," https://business.esa.int/newcomers-earth-observation-guide#ref_2.2.4, [Accessed 06-Dec-2022].

[68] C. Van Westen, "Remote sensing for natural disaster management," *International archives of photogrammetry and remote sensing*, vol. 33, no. B7/4; PART 7, pp. 1609–1617, 2000.

[69] W. Musakwa and A. Van Niekerk, "Earth observation for sustainable urban planning in developing countries: Needs, trends, and future directions," *Journal of Planning Literature*, vol. 30, no. 2, pp. 149–160, 2015.

[70] C. Chen, X. He, Z. Liu, W. Sun, H. Dong, and Y. Chu, "Analysis of regional economic development based on land use and land cover change information derived from landsat imagery," *Scientific Reports*, vol. 10, no. 1, pp. 1–16, 2020.

[71] F. Dolce, D. Di Domizio, D. Bruckert, A. Rodríguez, and A. Patrono, "Earth observation for security and defense," *Handbook of Space Security: Policies, Applications and Programs*, pp. 705–731, 2020.

[72] J. Vaynman, "Better monitoring and better spying: The implications of emerging technology for arms control (fall 2021)," *Texas National Security Review*, 2021.

[73] J. G. Teicher, "Are These Satellite Images War Propaganda? — newrepublic.com," https://newrepublic.com/article/165910/maxar-ukraine-russia-satellite-images-war-propaganda, [Accessed 06-Dec-2022].

[74] "New satellite images show Russia stealing Ukraine's grain — telegraph.co.uk," https://www.telegraph.co.uk/world-news/2022/05/24/new-satellite-images-show-russia-stealing-ukraines-grain/, [Accessed 06-Dec-2022].

[75] R. Šikl, H. Svatoňová, F. Děchtěrenko, and T. Urbánek, "Visual recognition memory for scenes in aerial photographs: exploring the role of expertise," *Acta Psychologica*, vol. 197, pp. 23–31, 2019.

[76] S. Ganguli, P. Garzon, and N. Glaser, "Geogan: A conditional gan with reconstruction and style loss to generate standard layer of maps from satellite images," *arXiv preprint arXiv:1902.05611*, 2019.

[77] X. Deng, Y. Zhu, and S. Newsam, "Using conditional generative adversarial networks to generate ground-level views from overhead imagery," *arXiv preprint arXiv:1902.06923*, 2019.

[78] K. Doi, K. Sakurada, M. Onishi, and A. Iwasaki, "Gan-based sar-to-optical image translation with region information," in *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2020, pp. 2069–2072.

[79] X. Yang, J. Zhao, Z. Wei, N. Wang, and X. Gao, "Sar-to-optical image translation based on improved cgan," *Pattern Recognition*, vol. 121, p. 108208, 2022.

[80] P. Singh and N. Komodakis, "Cloud-gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 1772–1775.

[81] H. Sinha, S. Kumar, and S. Chaudhury, "A variational training perspective to gans for hyperspectral image generation," in *Soft Computing for Problem Solving: Proceedings of SocProS 2020, Volume 1*. Springer, 2021, pp. 417–429.

[82] B. Zhao, S. Zhang, C. Xu, Y. Sun, and C. Deng, "Deep fake geography? when geospatial data encounter artificial intelligence," *Cartography and Geographic Information Science*, vol. 48, no. 4, pp. 338–352, 2021.

[83] H.-S. Chen, K. Zhang, S. Hu, S. You, and C.-C. J. Kuo, "Geo-defakehop: High-performance geographic fake image detection," *arXiv preprint arXiv:2110.09795*, 2021.

[84] R. R. Hoffman, *The psychology of expertise: Cognitive research and empirical AI*. Psychology Press, 2014.

[85] K. A. Ericsson, R. R. Hoffman, A. Kozbelt, and A. M. Williams, *The Cambridge handbook of expertise and expert performance*. Cambridge University Press, 2018.

[86] J. Wardlaw, J. Sprinks, R. Houghton, J.-P. Muller, P. Sidiropoulos, S. Bamford, and S. Marsh, "Comparing experts and novices in martian surface feature change detection and identification," *International journal of applied earth observation and geoinformation*, vol. 64, pp. 354–364, 2018.

[87] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[88] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, jan 2018.

[89] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in neural information processing systems*, 2017, pp. 6626–6637.

[90] A. Borji, "Pros and Cons of GAN Evaluation Measures," feb 2018.

[91] S. Barratt and R. Sharma, "A note on the inception score," *arXiv preprint arXiv:1801.01973*, 2018.

[92] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[93] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen, "The role of imagenet classes in fr\'echet inception distance," *arXiv preprint arXiv:2203.06026*, 2022.

[94] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing generative models via precision and recall," *Advances in neural information processing systems*, vol. 31, 2018.

[95] L. Simon, R. Webster, and J. Rabin, "Revisiting precision and recall definition for generative model evaluation," *arXiv preprint arXiv:1905.05441*, 2019.

[96] R. Ratcliff, C. Voskuilen, and A. Teodorescu, "Modeling 2-alternative forced-choice tasks: Accounting for both magnitude and difference effects," *Cognitive psychology*, vol. 103, pp. 1–22, 2018.

[97] M. Hodosh and J. Hockenmaier, "Focused evaluation for image description with binary forced-choice tasks," in *Proceedings of the 5th Workshop on Vision and Language*, 2016, pp. 19–28.

[98] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, 2016, pp. 2234–2242.

[99] Z. Wang, Q. She, A. F. Smeaton, T. E. Ward, and G. Healy, "Synthetic-neuroscore: Using a neuro-ai interface for evaluating generative adversarial networks," *Neurocomputing*, vol. 405, pp. 26–36, 2020.

[100] T. Tien, P. H. Pucher, M. H. Sodergren, K. Sriskandarajah, G.-Z. Yang, and A. Darzi, "Eye tracking for skills assessment and training: a systematic review," *journal of surgical research*, vol. 191, no. 1, pp. 169–178, 2014.

[101] D. Massaro, F. Savazzi, C. Di Dio, D. Freedberg, V. Gallese, G. Gilli, and A. Marchetti, "When art moves the eyes: a behavioral and eye-tracking study," *PloS one*, vol. 7, no. 5, p. e37285, 2012.

[102] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017, pp. 3226–3229.

[103] G. Labs, "Open Cities AI Challenge Dataset — doi.org," https://doi.org/10.34911/rdnt.f94cxb, 2019, [Accessed 07-Dec-2022].

[104] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset," *online: http://www. cs. toronto. edu/kriz/cifar. html*, vol. 55, 2014.

[105] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[106] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE, 2019, pp. 83–92.

[107] N. Yu, L. Davis, and M. Fritz, "Learning gan fingerprints towards image attribution," *arXiv preprint arXiv:1811.08180*, 2019.

[108] A. Eklund, "Feeding the zombies: Synthesizing brain volumes using a 3d progressive growing gan," *arXiv preprint arXiv:1912.05357*, 2019.

[109] J. Kim, S.-A. Hong, and H. Kim, "A stylegan image detection model based on convolutional neural network," *Journal of Korea Multimedia Society*, vol. 22, no. 12, pp. 1447–1456, 2019.

[110] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," *arXiv preprint arXiv:1912.11035*, 2019.

[111] G. Yildirim, N. Jetchev, R. Vollgraf, and U. Bergmann, "Generating high-resolution fashion model images wearing custom outfits," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[112] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," *arXiv preprint arXiv:2004.02546*, 2020.

[113] E. Collins, R. Bala, B. Price, and S. Süsstrunk, "Editing in style: Uncovering the local semantics of gans," *arXiv preprint arXiv:2004.14367*, 2020.

[114] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[115] A. Karnewar and O. Wang, "MSG-GAN: Multi-Scale Gradient GAN for Stable Image Synthesis," mar 2019.

[116] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346.

[117] Y. Viazovetskyi, V. Ivashkin, and E. Kashin, "Stylegan2 distillation for feed-forward image manipulation," *arXiv preprint arXiv:2003.03581*, 2020.

[118] C. H. Lin, C.-C. Chang, Y.-S. Chen, D.-C. Juan, W. Wei, and H.-T. Chen, "COCO-GAN: Generation by Parts via Conditional Coordinating," mar 2019.

[119] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.

[120] J. Donahue and K. Simonyan, "Large scale adversarial representation learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 10 541–10 551.

[121] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[122] K. K. Singh, U. Ojha, and Y. J. Lee, "Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6490–6499.

[123] X. Gong, S. Chang, Y. Jiang, and Z. Wang, "Autogan: Neural architecture search for generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3224–3234.

[124] A. Sanyal, P. H. Torr, and P. K. Dokania, "Stable rank normalization for improved generalization in neural networks and gans," *arXiv preprint arXiv:1906.04659*, 2019.

[125] L. Taylor and G. Nitschke, "Improving deep learning using generic data augmentation," *arXiv preprint arXiv:1708.06020*, 2017.

[126] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[127] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are gans created equal? a large-scale study," in *Advances in neural information processing systems*, 2018, pp. 700–709.

[128] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[129] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *arXiv preprint arXiv:2001.06937*, 2020.

[130] Y. A. Kolchinski, S. Zhou, S. Zhao, M. Gordon, and S. Ermon, "Approximating human judgment of generated image quality," *arXiv preprint arXiv:1912.12121*, 2019.

[131] Global Facility for Disaster Reduction and Recovery (GFDRR) Labs, "Open cities ai challenge dataset," 2020. [Online]. Available: https://registry.mlhub.earth/10.34911/rdnt.f94cxb

[132] X. Wu, K. Xu, and P. Hall, "A survey of image synthesis and editing with generative adversarial networks," *Tsinghua Science and Technology*, vol. 22, no. 6, pp. 660–674, 2017.

[133] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[134] S. Jung and M. Keuper, "Internalized biases in fréchet inception distance," in *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[135] M. J. Hautus, N. A. Macmillan, and C. D. Creelman, *Detection theory: A user's guide*. Routledge, 2021.

[136] B. Schneider and S. Parker, "Human methods: psychophysics," 2009.

[137] N. A. Macmillan and C. D. Creelman, *Detection theory: A user's guide*. Psychology press, 2004.

[138] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "Psychopy2: Experiments in behavior made easy," *Behavior research methods*, vol. 51, no. 1, pp. 195–203, 2019.

[139] J. Peirce and M. MacAskill, *Building experiments in PsychoPy*. Sage, 2018.

[140] G. Elsayed, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, and J. Sohl-Dickstein, "Adversarial examples that fool both computer vision and time-limited humans," *Advances in neural information processing systems*, vol. 31, 2018.

[141] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," *arXiv preprint arXiv:2204.07756*, 2022.

[142] J. E. Hoffman, "Visual attention and eye movements," *Attention*, vol. 31, no. 2, pp. 119–153, 1998.

[143] M. Borys and M. Plechawska-Wójcik, "Eye-tracking metrics in perception and visual attention research," *EJMT*, vol. 3, pp. 11–23, 2017.

[144] Z. Bylinskii and M. Borkin, "Eye fixation metrics for large scale analysis of information visualizations. etvis work," *Eye Track. Vis*, 2015.

[145] D. Purves, G. Augustine, D. Fitzpatrick, L. Katz, A. LaMantia, J. McNamara, and S. Williams, "Neuroscience 2nd edition. sunderland (ma) sinauer associates," *Types of Eye Movements and Their Functions*, 2001.

[146] U. Ahlstrom and F. J. Friedman-Berg, "Using eye movement activity as a correlate of cognitive workload," *International journal of industrial ergonomics*, vol. 36, no. 7, pp. 623–636, 2006.

[147] S. Pannasch, J. Schulz, and B. M. Velichkovsky, "On the control of visual fixation durations in free viewing of complex images," *Attention, Perception, & Psychophysics*, vol. 73, no. 4, pp. 1120–1132, 2011.

[148] G. Ziv, "Gaze behavior and visual attention: A review of eye tracking studies in aviation," *The International Journal of Aviation Psychology*, vol. 26, no. 3-4, pp. 75–104, 2016.

[149] L. Lévêque, H. Bosmans, L. Cockmartin, and H. Liu, "State of the art: Eye-tracking studies in medical imaging," *Ieee Access*, vol. 6, pp. 37 023–37 034, 2018.

[150] W. Zhang and H. Liu, "Toward a reliable collection of eye-tracking data for image quality research: challenges, solutions, and applications," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2424–2437, 2017.

[151] A. Carmichael, A. Larson, E. Gire, L. Loschky, and N. S. Rebello, "How does visual attention differ between experts and novices on physics problems?" in *AIP Conference Proceedings*, vol. 1289, no. 1. American Institute of Physics, 2010, pp. 93–96.

[152] H. Sheridan and E. M. Reingold, "Expert vs. novice differences in the detection of relevant information during a chess game: evidence from eye movements," *Frontiers in psychology*, vol. 5, p. 941, 2014.

[153] K. A. Kastens, T. F. Shipley, A. P. Boone, and F. Straccia, "What geoscience experts and novices look at, and what they see, when viewing data visualizations." *Journal of Astronomy & Earth Sciences Education*, vol. 3, no. 1, pp. 27–58, 2016.

[154] N. Caporusso, K. Zhang, and G. Carlson, "Using eye-tracking to study the authenticity of images produced by generative adversarial networks," in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE, 2020, pp. 1–6.

[155] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[156] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[157] R. L. Draelos and L. Carin, "Hirescam: Faithful location representation in visual attention for explainable 3d medical image classification," *arXiv preprint arXiv:2011.08891*, 2020.

[158] H. G. Ramaswamy *et al.*, "Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 983–991.

[159] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[160] L. Gorski, S. Ramakrishna, and J. M. Nowosielski, "Towards grad-cam based explainability in a legal text processing pipeline," *arXiv preprint arXiv:2012.09603*, 2020.

[161] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[162] A. Olsen, "The tobii i-vt fixation filter," *Tobii Technology*, vol. 21, pp. 4–19, 2012.

[163] K. Krejtz, A. Duchowski, T. Szmidt, I. Krejtz, F. González Perilli, A. Pires, A. Vilaro, and N. Villalobos, "Gaze transition entropy," *ACM Transactions on Applied Perception (TAP)*, vol. 13, no. 1, pp. 1–20, 2015.

[164] R. S. Weiss, R. Remington, and S. R. Ellis, "Sampling distributions of the entropy in visual scanning," *Behavior Research Methods, Instruments, & Computers*, vol. 21, no. 3, pp. 348–352, 1989.

[165] M. Doellken, J. Zapata, N. Thomas, and S. Matthiesen, "Implementing innovative gaze analytic methods in design for manufacturing: A study on eye movements in exploiting design guidelines," *Procedia CIRP*, vol. 100, pp. 415–420, 2021.

[166] B. A. Shiferaw, L. A. Downey, J. Westlake, B. Stevens, S. M. Rajaratnam, D. J. Berlowitz, P. Swann, and M. E. Howard, "Stationary gaze entropy predicts lane departure events in sleep-deprived drivers," *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[167] Y. Lee, K.-T. Jung, and H.-C. Lee, "Use of gaze entropy to evaluate situation awareness in emergency accident situations of nuclear power plant," *Nuclear Engineering and Technology*, vol. 54, no. 4, pp. 1261–1270, 2022.

[168] S. Lanini-Maggi, I. T. Ruginski, T. F. Shipley, C. Hurter, A. T. Duchowski, B. B. Briesemeister, J. Lee, and S. I. Fabrikant, "Assessing how visual search entropy and engagement predict performance in a multiple-objects tracking air traffic control task," *Computers in Human Behavior Reports*, vol. 4, p. 100127, 2021.

[169] A. Finley and S. Penningroth, "Online versus in-lab: Pros and cons of an online prospective memory experiment," *Advances in psychology research*, vol. 113, pp. 135–162, 2015.

[170] C. Bolton, E. Miskioglu, and M. Roth, "The impact of gender identity on early-career engineer's perception of expertise," in *2022 ASEE Annual Conference & Exposition*, 2022.

[171] N. A. McIntyre, M. T. Mainhard, and R. M. Klassen, "Are you looking to teach? cultural, temporal and dynamic insights into expert teacher gaze," *Learning and Instruction*, vol. 49, pp. 41–53, 2017.

[172] J. A. Brefczynski and E. A. DeYoe, "A physiological correlate of the'spotlight'of visual attention," *Nature neuroscience*, vol. 2, no. 4, pp. 370–374, 1999.

[173] R. Geirhos, D. H. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann, "Comparing deep neural networks against humans: object recognition when the signal gets weaker," *arXiv preprint arXiv:1706.06969*, 2017.

[174] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[175] A. Borghesi, F. Baldo, and M. Milano, "Improving deep learning models via constraint-based domain knowledge: a brief survey," *arXiv preprint arXiv:2005.10691*, 2020.

[176] X. Luo, D. Zhang, and X. Zhu, "Deep learning based forecasting of photovoltaic power generation by incorporating domain knowledge," *Energy*, vol. 225, p. 120240, 2021.

[177] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, "A survey on incorporating domain knowledge into deep learning for medical image analysis," *Medical Image Analysis*, vol. 69, p. 101985, 2021.

Appendix A

List of Abbreviations

EO	Earth Observation
DNN	Deep Neural Network
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long-Short Term Memory
GAN	Generative Adversarial network
MAE	Mean Average Error
AP	Average Precision
ACC	Accuracy
FID	Frechet Inception Distance
KID	Kernel Inception Distance
SOTA	State of the Art

Appendix B

Additional Results from GAN models

Figure B.1: *Randomly selected real images from the Inria Benchmark Aerial Imagery Dataset (256×256)*

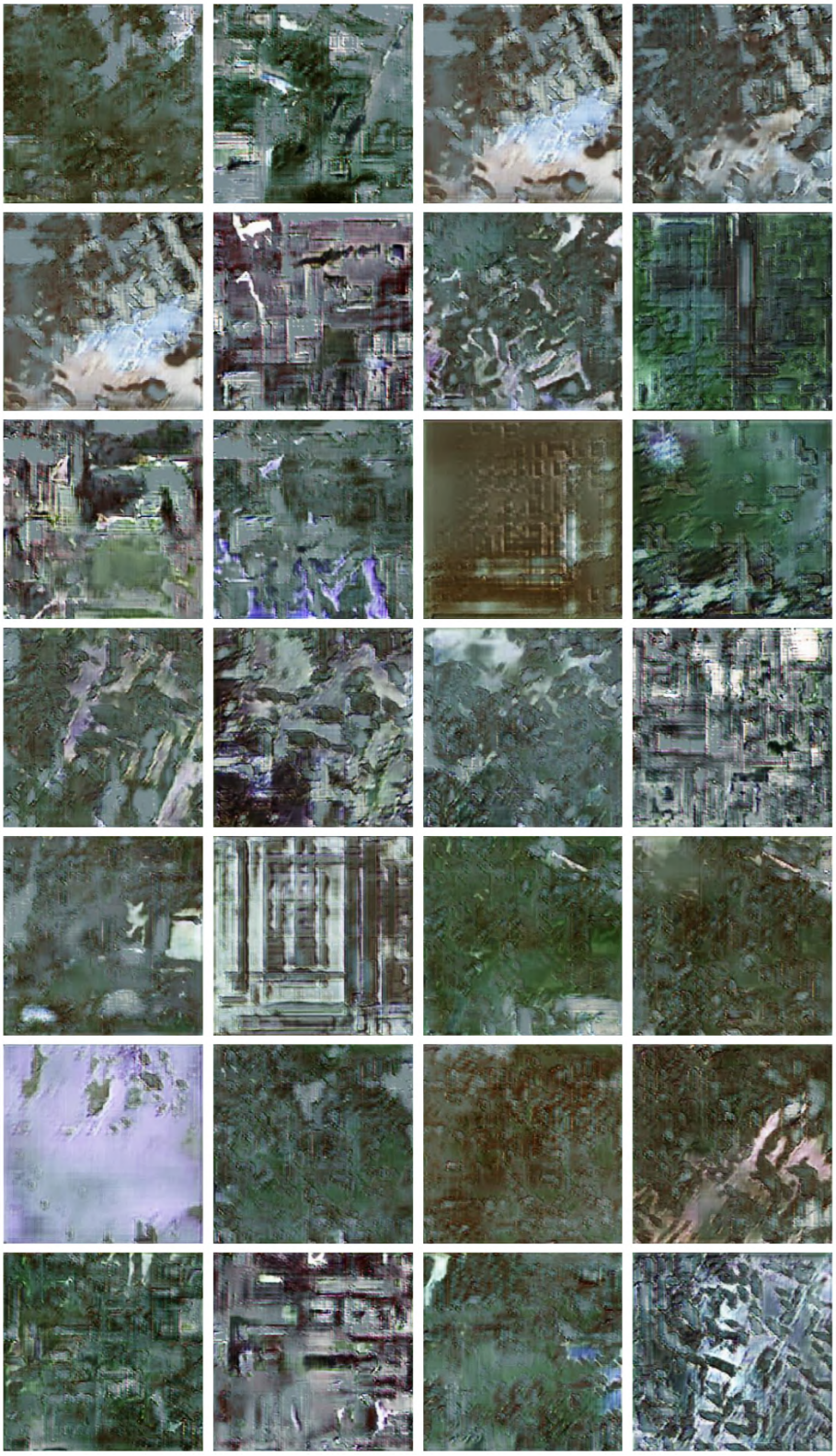

Figure B.2: *Randomly selected images generated from the baseline DCGAN (256×256)*

Figure B.3: *Randomly selected images generated from PGGAN (256×256)*

Figure B.4: *Randomly selected images generated from StyleGAN2 (256×256)*

Figure B.5: *Randomly selected images generated from CoCoGAN (256×256)*

Figure B.6: *Randomly selected images generated from StyleGAN2 (1024x1024)*

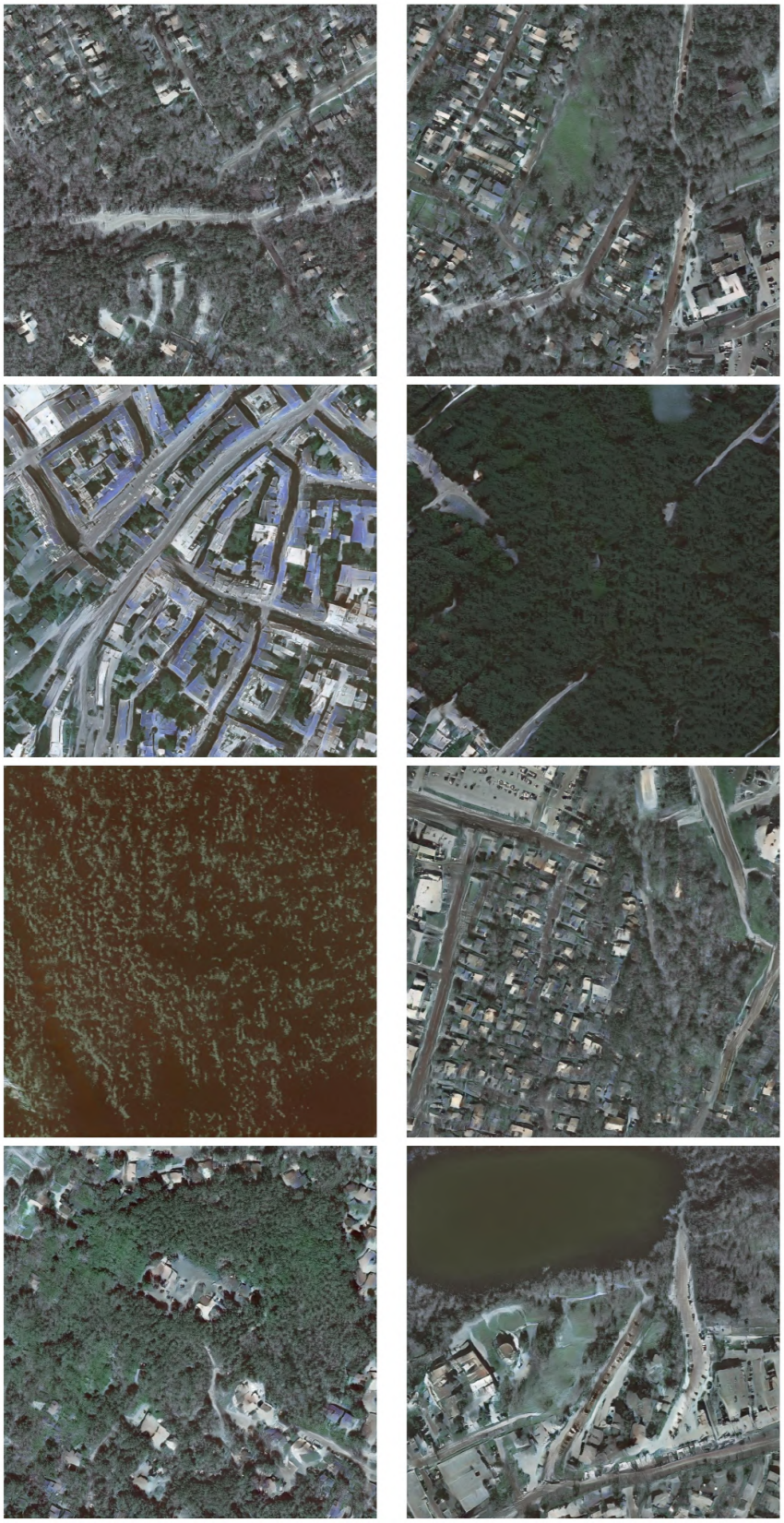

Figure B.7: *Randomly selected images generated from StyleGAN2 (1024x1024)*

Figure B.8: *StyleGAN2 latent space representations (left) of target images from the Inria training dataset (right)*

Appendix C

Additional material

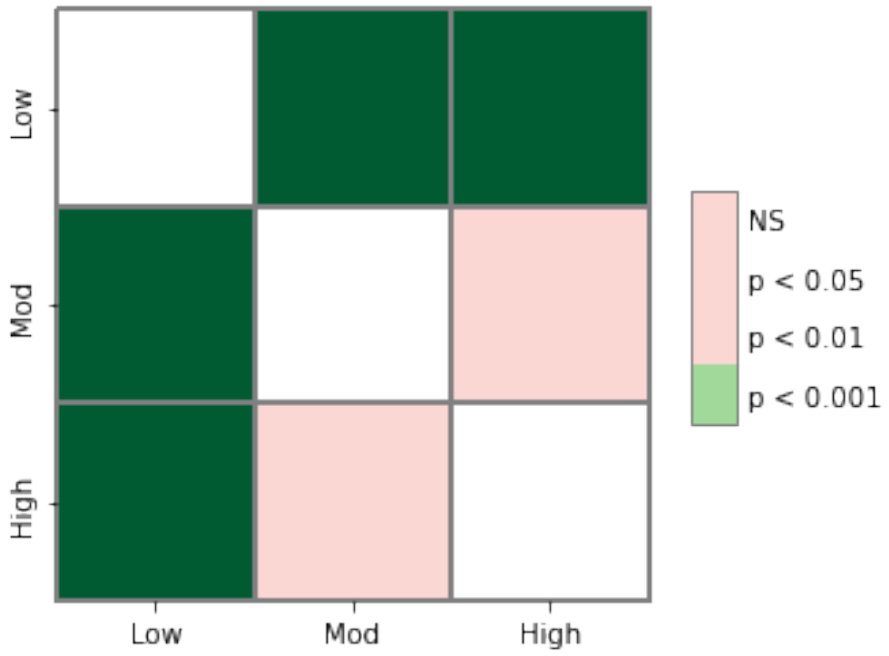

Figure C.1: *Pairwise comparison of experience levels. (H Statistic: 20.086, p* : 0.001*) Different distributions (reject H0)* ε^2 : 0.223

Dependent Variable	F	P-value
Accuracy	8.14(2,23)	0.0021
Avg resp total	8.12(2,23)	0.0021
Avg Resp Correct	7.84(2,23)	0.0025
Avg Resp Incorrect	8.75(2,23)	0.0015
Transition	6.71(2,23)	0.0051
Stationary	11.55(2,23)	0.00034

Table C.1: Between groups Post-Hoc ANOVA results for the eye tracking study in Chapter 5. The variable values are task accuracy (Acc), average response time (Avg Resp Total), average response time for correct answers (Avg Resp Corr) and incorrect answers (Avg Resp Incorr), gaze transitional entropy (Ht) and gaze stationary entropy. The Post-Hoc ANOVA test using a Bonferroni corrected alpha of 0.017

```
==========================================================================================
Layer (type:depth-idx)                             Output Shape              Param #
==========================================================================================
UNet                                               [1, 1, 512, 512]          --
├─ModuleList: 1-7                                  --                        (recursive)
│    └─DoubleConv: 2-1                             [1, 64, 512, 512]         --
│    │    └─Sequential: 3-1                        [1, 64, 512, 512]         38,848
├─MaxPool2d: 1-2                                   [1, 64, 256, 256]         --
├─ModuleList: 1-7                                  --                        (recursive)
│    └─DoubleConv: 2-2                             [1, 128, 256, 256]        --
│    │    └─Sequential: 3-2                        [1, 128, 256, 256]        221,696
├─MaxPool2d: 1-4                                   [1, 128, 128, 128]        --
├─ModuleList: 1-7                                  --                        (recursive)
│    └─DoubleConv: 2-3                             [1, 256, 128, 128]        --
│    │    └─Sequential: 3-3                        [1, 256, 128, 128]        885,760
├─MaxPool2d: 1-6                                   [1, 256, 64, 64]          --
├─ModuleList: 1-7                                  --                        (recursive)
│    └─DoubleConv: 2-4                             [1, 512, 64, 64]          --
│    │    └─Sequential: 3-4                        [1, 512, 64, 64]          3,540,992
├─MaxPool2d: 1-8                                   [1, 512, 32, 32]          --
├─DoubleConv: 1-9                                  [1, 1024, 32, 32]         --
│    └─Sequential: 2-5                             [1, 1024, 32, 32]         --
│    │    └─Conv2d: 3-5                            [1, 1024, 32, 32]         4,718,592
│    │    └─BatchNorm2d: 3-6                       [1, 1024, 32, 32]         2,048
│    │    └─ReLU: 3-7                              [1, 1024, 32, 32]         --
│    │    └─Conv2d: 3-8                            [1, 1024, 32, 32]         9,437,184
│    │    └─BatchNorm2d: 3-9                       [1, 1024, 32, 32]         2,048
│    │    └─ReLU: 3-10                             [1, 1024, 32, 32]         --
├─ModuleList: 1-10                                 --                        --
│    └─Sequential: 2-6                             --                        --
│    │    └─ConvTranspose2d: 3-11                  [1, 512, 64, 64]          2,097,664
│    │    └─DoubleConv: 3-12                       [1, 512, 64, 64]          7,079,936
│    └─Sequential: 2-7                             --                        --
│    │    └─ConvTranspose2d: 3-13                  [1, 256, 128, 128]        524,544
│    │    └─DoubleConv: 3-14                       [1, 256, 128, 128]        1,770,496
│    └─Sequential: 2-8                             --                        --
│    │    └─ConvTranspose2d: 3-15                  [1, 128, 256, 256]        131,200
│    │    └─DoubleConv: 3-16                       [1, 128, 256, 256]        442,880
│    └─Sequential: 2-9                             --                        --
│    │    └─ConvTranspose2d: 3-17                  [1, 64, 512, 512]         32,832
│    │    └─DoubleConv: 3-18                       [1, 64, 512, 512]         110,848
├─Conv2d: 1-11                                     [1, 1, 512, 512]          65
==========================================================================================
Total params: 31,037,633
Trainable params: 31,037,633
Non-trainable params: 0
Total mult-adds (G): 218.47
==========================================================================================
Input size (MB): 3.15
Forward/backward pass size (MB): 2300.58
Params size (MB): 124.15
Estimated Total Size (MB): 2427.87
==========================================================================================
```

Figure C.2: *UNet model summary for images generated in Chapter 6*