ISPRS Journal of Photogrammetry and Remote Sensing 190 (2022) 231–251

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Evaluation of synthetic aerial imagery using unconditional generative adversarial networks

Matthew Yates[a,*], Glen Hart[b], Robert Houghton[c], Mercedes Torres Torres[a], Michael Pound[a]

[a] School of Computer Science, University of Nottingham, United Kingdom
[b] Defence Science and Technology Laboratory, United Kingdom
[c] Human Factors Research Group, Faculty of Engineering, University of Nottingham, United Kingdom

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Generative adversarial networks
Aerial imagery

ABSTRACT

Image generation techniques, such as generative adversarial networks (GANs), have become sufficiently sophisticated to cause growing concerns around the authenticity of images in the public domain. Although these generation techniques have been applied to a wide range of images, including images with objects and faces, there are comparatively few studies focused on their application to the generation and subsequent evaluation of Earth Observation (EO) data, such as aerial and satellite imagery. We examine the current state of aerial image generation by training state-of-the-art unconditional GAN models to generate realistic aerial imagery. We train PGGAN, StyleGAN2 and CoCoGAN models using the Inria Aerial Image benchmark dataset, and quantitatively assess the images they produce according to the Fréchet Inception Distance (FID) and the Kernel Inception Distance (KID). In a paired image human detection study we find that current synthesised EO images are capable of fooling humans and current performance metrics are limited in their ability to quantify the perceived visual quality of these images.

1. Introduction

With the rapid increase in the availability of image data and the continued development of computer vision technology, image synthesis has become an emerging area of research due to its many different applications. These range from editing photographs, to creating synthesised datasets for training models and creating highly realistic fake imagery (Karras et al., 2019; Gui et al., 2020; Brock et al., 2016). Despite the wide variety of different public imagery available for training, much of the current research has focused on facial imagery for both editing existing photos and generating completely new and artificial faces (Wang et al., 2018; Yin et al., 2017; Tolosana et al., 2020).

Due to the large amount of public personal photos on social media platforms, it is perhaps not surprising that faces have been the main focus of most of the research in image synthesis (Karras et al., 2018; Zhang et al., 2020), with object generation being a close second (Singh et al., 2019). An area less explored is that of the synthesis of Earth Observation (EO) data. With increasing concerns about the generation and spread of fake information (Shu et al., 2017; Scheufele and Krause, 2019), commonly referred to as *"fake news"*, and with clear evidence on the effects that this misinformation has on the public (Cao et al., 2020), it is not unreasonable to think that it is a matter of time before these techniques for fake information generation start making use of additional sources of data to expand their reach and impact. EO data is a prime candidate for these machinations, as its main uses rely on it being objective and trustworthy. Therefore obtaining an objective and thorough evaluation on the current capabilities of image synthesis methods for EO data is an important goal.

Scientifically, EO data presents challenges that are not present in face or object image synthesis. Images containing faces or objects tend to require high levels of detail on centralised objects, indeed the majority of synthetic face generation solutions operate on aligned face images. Instead of having only one focus or feature that must be generated within the image, the challenge in EO imagery synthesis resides in generating a set of features and patterns that both individually and collectively can be deemed as "real". The ability of Generative Adversarial Networks' to generalise beyond object and face synthesis is underexplored.

There are also practical reasons that EO data is an important target for image synthesis. There is extensive evidence on the benefits of using GAN-generated synthetic data for data augmentation purposes (Sandfort et al., 2019; Tanaka and Aranha, 2019). This has positively impacted the

* Corresponding author.
E-mail address: matthew.yates1@nottingham.ac.uk (M. Yates).

https://doi.org/10.1016/j.isprsjprs.2022.06.010
Received 8 October 2021; Received in revised form 18 May 2022; Accepted 16 June 2022
Available online 28 June 2022

performance of machine learning models in other areas, such as medical imaging (Waheed et al., 2020) and autonomous vehicles (Lee et al., 2020). Synthetic image generation also improves performance in problems with limited data (Zhang et al., 2019), or where data contains class imbalances (Mariani et al., 2018a). Synthesised EO data could be used to generate challenging datasets that are hard to obtain naturally, such as synthetic landscapes containing camouflaged buildings, structures or vehicles to be used to train models in the detection of these objects.

The most successful approaches to image synthesis in recent years are based on deep learning, with the most popular and successful models using a Generative Adversarial Networks (GANs) framework (Goodfellow et al., 2014). These models require large amounts of image data to train, and often comprise two competing neural networks, a generator and a discriminator network, pitted against each other in a competitive zero-sum game. The reason for their popularity is that recent GAN models are able generate photo-realistic, high-resolution images across different image data sets (Gui et al., 2020). However much of the current research and testing of models relies on the use of large, popular benchmark datasets such as ImageNet (Donahue and Simonyan, 2019a) or FFHQ (Karras et al., 2019b). The reliance on using these datasets for testing models means that the generalisability of these approaches has not yet been proven. Quantitative and qualitative evaluation of the performance of image generation in more domain specific image types, such as those seen in EO image data, is required. It is also important to test the visual quality of the samples using user studies, rather than relying on common machine learning metrics alone. While popular metrics such as Fréchet Inception Distance (FID) are useful in determining if the distributions of generated images are broadly similar to the original dataset, they do not necessarily reflect all the visual features needed for the human eye to perceive an image as realistic.

Human perception of EO data is important; increasingly realistic fake imagery could present a security threat. In the last few years, there has been growing evidence for the use of machine learning and data science for fake information generation and communication (Bastos and Mercea, 2019). There has been a documented erosion in public trust of mass communication (Chesney and Citron, 2019). The injection of fake imagery into this ecosystem, created with the sole purpose of "adding veracity" to this fake information, represents a real and long-term danger. There are particular risks in the earth imagery domain: it is a domain that typically promotes the re-use of open-source imagery that is both widely shared and shareable (Malarvizhi et al., 2016; Haslett and Wong, 2019). These data sources are trusted by the public, and used extensively by organisations including news organisations, human rights organisations and private industry. Synthetic EO data could also be used to negatively affect intelligent vehicles (Manderson et al., 2020), or to affect land-use (Oubrahim et al., 2018) or environmental-change policy through fake photographs with incorrect or misleading information. From a Human Right's perspective, they could also be used to mask evidence of human rights violations and modern slavery, where detection EO data has provided extremely useful in recent years (Boyd et al., 2018). Even in closed systems working with a carefully sourced private data pipeline, there remain a myriad justifications (e.g. ideological, political, espionage, etc.) for tampering, insider threat, and in general inserting fake imagery as the goal of a cyber-security breach, as pointed out by Mirsky et al. (2019).

Understanding and testing the capabilities of current generation models on EO data as well as current metrics for analysis/detection of these synthetic images is an important step to address the rising problem of fake image generation. This paper evaluates the generation of Earth Observation images using recent GAN models with prior high-quality unconditional image synthesis in other domains. Synthesised image quality is measured using the FID, a standard GAN evaluation metric, and the more recent metric Kernel Inception Distance (KID). These metrics evaluate images by comparing the similarity between the fake and real image distributions using the feature maps of specific layers of the Inception V3 classifier (Szegedy et al., 2016). Typically lower FID and KID scores have been seen as an indicator of image generation performance, however such evaluation has only been performed on common domains such as face images. In this paper we also contribute an evaluation of these metrics against human perception, with a user study providing evidence that common metrics are not suitable as a measure of visual quality in Earth Observation.

1.1. Contributions

This paper explores the current state of synthetic aerial image generation, and the extent to which these synthetic images can fool real human observers. We show that current synthetic aerial imagery can fool human perception, and highlight the limitations of current Inception-based GAN evaluation metrics when used as measures of human perception of visual verisimilitude.

The contributions of this paper are:

- A systematic evaluation of recent, state-of-the-art unconditional GAN models for the task of generating synthetic aerial imagery.
- Paired image detection study showing that current synthetic aerial imagery can fool observers.
- Comparison of human and mathematical metrics for visual quality.
- Discussion on the impacts of urban/rural features in generated samples on human detection.

2. Literature review

2.1. Generative adversarial networks

Since their inception in 2014, Generative Adversarial Networks (GANs) have been used for different computer vision tasks, such as style transfer (Park et al., 2019), super resolution (Wang et al., 2018) and image-to-image translation (Zhu et al., 2017). For each domain, there also exists many different variations (Pan et al., 2019). While early models could learn the distribution of a given dataset and generate new instances of data, these were often limited to low-resolution images, lacking variation and visual quality. Recent models are now used in a much wider context and can produce images of higher quality and resolution (Wang et al., 2020). In addition to image synthesis, GANs have also been used with other data types such as audio (Donahue et al., 2018) and video (Xiong et al., 2018). GANs can be trained by using supervised or unsupervised methods. Conditional GANs learn via supervised learning and require labelled data to train the generator and discriminator. Unconditional GANs do not require labels and learn unsupervised. In this paper we focus on unsupervised image generation, and evaluate the abilities of recent unconditional GAN models in the context of aerial imagery data via both statistical and human evaluation metrics. We compare the statistical, inception based metrics and human detection scores of different image features (rural/urban).

GAN models share the same underlying principles of the adversarial training, comprising two opposing deep neural networks: a generator and a discriminator (Goodfellow et al., 2014). 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. This process is given by the Eq. (1):

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

2.1.1. 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.

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 (Krizhevsky et al., 2014), LSUN (Yu et al., 2015) and CelebA-HQ (Karras et al., 2018)). These datasets contain a wide array of images. While the images generated in the model are almost photo realistic, 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 (Matern et al., 2019; Yu et al., 2019; Marra et al., 2018b). 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 (Eklund, 2019) and 3D MR images of brain volumes (Eklund, 2019).

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. 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).

2.1.2. StyleGAN and StyleGAN2

The original StyleGAN model is an update from PGGAN that enabled the model to learn unsupervised separation of image attributes. This lead to the network being able to have more control over the image output (Karras et al., 2019b). In addition to control over different “visual styles”, the model also achieved new state of the art Inception scores on benchmark datasets (CelebA HQ (Karras et al., 2018), FFHQ (Karras et al., 2019b)).

StyleGAN has been used in many of the same areas as PGGAN, including detecting GAN fingerprints (Marra et al., 2018b; Kim et al., 2019) and face generation (Wang et al., 2019). 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 (Yildirim et al., 2019; Härkönen et al., 2020; Collins et al., 2020). StyleGAN has been trained to produce different types of images, with the most common test datasets being faces (FFHQ, CelebA Liu et al., 2015), as well as commonly used datasets containing different object categories (LSUN Yu et al., 2015.

StyleGAN2 is the latest iteration to be released (Karras et al., 2019). It includes significant changes in the architecture which allow it to obtain state-of-the-art performance in image generation and style transfer tasks (Karras et al., 2019; Wang et al., 2019). Similarly to previous models, StyleGAN2 has so far been used primarily for generating realistic faces and other object categories (Viazovetskyi et al., 2020) and has not been applied to the generation of aerial images. In comparison with previous versions (PGGAN, StyleGAN), StyleGAN2 is able to 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 likely 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) (Karras et al., 2019), these images are actually easier to detect as synthetic by image classifers trained to disinguish 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 humans and deep learning based models.

2.1.3. CoCoGAN

Conditional Coordinate GAN (CoCoGAN) presents another novel GAN architecture with results that rival other state-of-the-art GANs (StyleGAN2, BigGAN) (Lin et al., 2019). 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 few different datasets including object datasets such as CelebA and the LSUN256 dataset (Yu et al., 2015). As well as these standard benchmark datasets, the model was also able to achieve low FID scores for the panorama dataset Matterport3D (Chang et al., 2017). This presents a different challenge for image generation than in object focused dataset as in requires the model to learn how to create a coherent image with decentralised features, much like those seen in aerial images.

2.1.4. BigGAN and BigBiGAN

BigGAN is another recent GAN model which is capable of conditional and unconditional high resolution image generation (Brock et al., 2018). 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 (Brock et al., 2018). 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 in order to achieve greater variety and photo realism in the generated images (Donahue and Simonyan, 2019b).

Despite being a recent model with competitive performance in generating realistic images (Wang et al., 2019; Donahue and Simonyan, 2019b), neither BigGAN or 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 paper focus on generally applicable techniques for the wider community.

2.2. Other generative models

While GANs such as PGGAN and StyleGAN are the currently the most widely used models for unconditional images generation, other methods outside of the GAN architecture can also be applied. Non-GAN methods such as Transformers (Esser et al., 2021) and Diffusion Models (Sohl-Dickstein et al., 2015) may offer better scaliabliy, faster training times and more robust performance than their GAN counterparts.

Transformers (Vaswani et al., 2017) are self-attention models which have become one of primary methods used for natural language processing (NLP) tasks. More recently, transformers have been applied to the field of image generation (Chen et al., 2020) with the resulting images rivaling the quality of state-of-the-art GAN results.

Another set of methods seen in the field of image generation are variational autoencoders (VAEs) (Kingma and Welling, 2013). These models work by encoding a given input, such as an image, and then reconstruct it using its learnt encoding. For image generation VAEs are notably easier and quicker to train than GANs but do not achieve the same image quality that GANs are capable of. This has led to the merging of these two architectures to create VAE-GANs (Xian et al., 2019) which take advantage of the useful features from both systems.

More recently diffusion models (Sohl-Dickstein et al., 2015) have been applied to image generation. These are likelihood-based models which work by gradually removing noise from an input signal such as an image and in some cases can produce results which rival or even surpass those of GANs (Dhariwal and Nichol, 2021).

2.3. Earth observation data

Despite not being as prominent in deep learning research as other types of images such as faces or object datasets (ImageNet, LSUN256, CelebA), Earth Observation data such as satellite aerial imagery provides an interesting challenge for testing the capabilities of models for tasks such as image synthesis and is justified by real world applications.

Aerial and satellite imagery are used in for a range of tasks such as mapping and remote sensing and is found across multiple industry sectors from security and intelligence (Do et al., 2018), economic assessment of regions and disaster warnings (Bredemeyer et al., 2018; Chuvieco et al., 2010).

Previous work involving both GANs and aerial image data is often concerned more with the problem of image to image translation (Isola et al., 2017) for mapping. Researchers have explored training conditional GAN models on satellite data for different image translation tasks. They have been used for generating maps from satellite data (Ganguli et al., 2019) and also for estimating ground level views (Deng et al., 2019). The potential security threat from deep learning based aerial imagery was explored in one 2021 study (Zhao et al., 2021) which investigated detection methods with samples generated from image to image translation methods.

Super-resolution models, where the model attempts to enhance and upscale blurry, low resolution data have also been applied to aerial data (Jiang et al., 2019). However, there has been much less research on the generation of novel photo-realistic, high-resolution aerial images. One of the closest papers in terms of use of Earth Observation data would be Cloud-GAN (Singh and Komodakis, 2018) which used satellite imagery as the training data for the purpose of creating a model to remove cloud coverage from images. Like other previous works surrounding GANs and Earth Observation data, this paper focused on image to image translation rather than the unconditional synthesis of novel images.

Although satellite imagery has not been the focus of unconditional image synthesis, artificial hyper-spectral data has been successfully generated (Audebert et al., 2018) to augment training datasets for tasks where which require more specific hyper-spectral data than is available.

Using Earth Observation images as the training data for GANs is also useful for testing their generalization capabilities beyond the common benchmark datasets. Although there are instances of these models being trained on more novel datasets, the majority of benchmark dataset that get used are of objects and faces, these images all have centralized features that need to be learnt to be able to successfully mimic. Data-decentralized features such as aerial images, which have a larger distribution of defining visual features present, could provide a challenge for architectures fine tuned to generate objects. This difficulty with images lacking a central object can be observed in the results of PGGAN were the images generated from the LSUN bedrooms dataset have much less convincing image realism compared to the results from the CelebA-HQ dataset (Karras et al., 2018).

3. Methodology

3.1. Experimental setup

We tested the performance of the PGGAN,[1] StyleGAN2[2] and CoCo-GAN[3] networks and, for comparison, also evaluated a baseline DCGAN[4] In all experiments, the official implementations of the networks were used.

These networks were chosen due to their reported SotA 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 also most commonly 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 Singh et al., 2019, AutoGAN Gong et al., 2019, SRNGAN Sanyal et al., 2019) 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.

A basic deep convolutional generative adversarial network (DCGAN) serves as a baseline model during this study, 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 GAN base architecture described by Ian Goodfellow (Goodfellow et al., 2014). DCGAN uses simple convolutional neural networks for its adversarial training. The generator network G up samples the random noise vector z though 7 convolutional blocks to produce an image that is given to the discriminator D, G also uses Relu as its activation function for each layer. D itself is a CNN classifier made up of 6 convolutional layers and a single dense layer. D uses average pooling between each convolutional block and a leaky relu activation function. A Minimax (Goodfellow et al., 2014) loss function is used for training both networks.

Each model was trained until they achieved model convergence, and for each trained model, a sample dataset of synthesised images was generated for evaluation purposes. 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.

Samples from the best performing model were then further explored in a user study which measured participant accuracy in correctly identifying which image is synthetic in a series of image pairs each consisting of a real image of aerial satellite photography (INRIA Dataset Maggiori et al., 2017) and a generated sample.

[1] https://github.com/tkarras/progressive_growing_of_gans (Karras et al., 2018).

[2] https://github.com/NVlabs/stylegan2 (Karras et al., 2019).

[3] https://github.com/hubert0527/COCO-GAN (Lin et al., 2019)

[4] https://github.com/t0nberryking/DCGAN256.

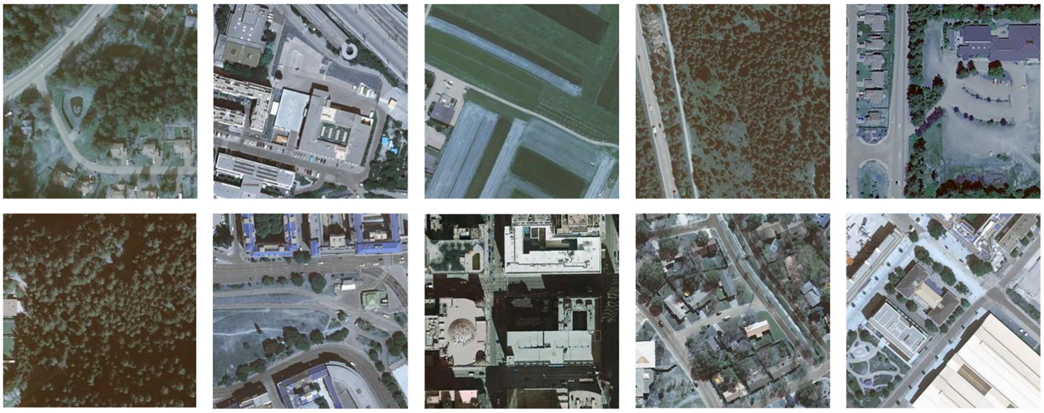


Fig. 1. Random selection of images from the INRIA Aerial Imagery Benchmark Dataset (Maggiori et al., 2017).

3.2. Dataset

Most of the published research in the area of deep learning methods with Earth Observation data use *map2sat* as a baseline dataset (Marra et al., 2018b). 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 (Zhu et al., 2017), 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 make it unsuitable for the unconditional generation task we address here.

In this paper we evaluate the models using the INRIA Aerial Imagery Dataset (Maggiori et al., 2017), 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 km^2 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.

3.2.1. 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 roughly a 50/50 split in terms of rural and urban features present in the final tiles.

Data augmentation techniques have been successfully applied in deep learning problems to improve performance (Taylor et al., 2017). In our case, we applied both horizon and horizonal flipping transformations to the original dataset (Perez and Wang, 2017). 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, our dataset is comprised of 34,600 256 × 256 images. We chose this resolution because it was the highest resolution shared by each of the tested models. Furthermore, we created an additional training dataset for the evaluation of StyleGAN2. This dataset contains 16,500 images at a resolution of 1024×1024 pixels. 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.

3.3. Metrics

To assess the performance of each model we use Fréchet Inception Distance (Borji, 2018) (FID) and the Kernel Inception Distance (Bińkowski et al., 2018) (KID). The Fréchet Inception Distance has become one of the most widely used metrics (Lucic et al., 2018) 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(Szegedy et al., 2016). 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 2:

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

Where $\mu_r, \sum_r$ are the mean and co-variance of the real data distribution, with $\mu_g, \sum_g$ being that for the generated data distribution. FID superseded the previous GAN evaluation standard, Inception Score (Salimans et al., 2016) as it has been shown to be a more robust measurement of image quality (Heusel et al., 2017). The Inception score has also been shown to inadequately detect overfitting (Barratt and Sharma, 2018), as it only uses the generated samples while ignoring the training set, and is also sensitive to image resolution (Borji, 2018).

Kernel Inception Distance (Bińkowski et al., 2018) 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 $P_r = P_g$ (Gui et al., 2020). It is calculated as shown in equation in 3:

$$M_k(P_r,P_g) = \mathbb{E}_{x,x'\sim P_r}[k(x,x')] - 2\mathbb{E}_{x\sim P_r,y\sim P_g}[k(x,y)] + \mathbb{E}_{y,y'\sim P_g}[k(y,y')] \tag{3}$$

where k is a fixed kernel function (e.g. Polynomial Kernel $k(x,y) =$

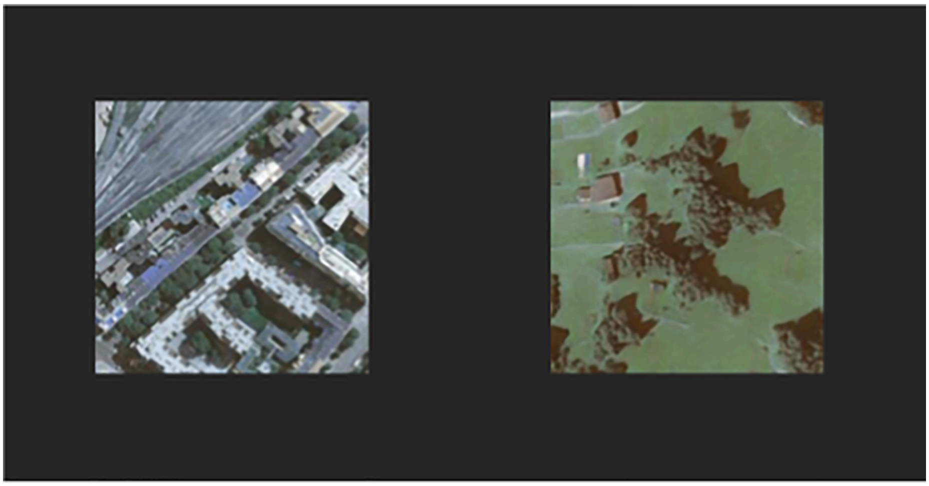


Fig. 2. Screenshot from the forced choice study. One image is real and the other is synthetic. Participants are asked to indicate which one is synthetic.

Table 1
Number of trainable parameters and training time for each tested network.

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

Table 2
Metrics for Baseline and state-of-the-art models. Best performance is shown in bold (lower values are better).

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
Metrics for StyleGAN2 at 1024×1024 resolution (lower is better).

Model	FID (Mean ± SD)	KID (Mean ± SD)
StyleGAN2 (256)	16.59 ± 0.18	7.28 ± 0.61
StyleGAN2 (1024)	32.70 ± 0.254	16.36 ± 0.59

$(\frac{1}{d}x^T y + 1)^3$, 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 (Bińkowski et al., 2018), 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 (Kolchinski et al., 2019; Fan et al., 2017). FID has been found correlate with measures of human perception towards assessing image realism and image quality in GAN samples (Heusel et al., 2017). 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) (Zhou et al., 2019). 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 (Deng et al., 2009). For the main evaluation of GAN models we use the pretrained Inception model that is the standard practice in many GAN papers (Karras et al., 2019b; Donahue and Simonyan, 2019b).

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

Table 4
Metrics table for urban and rural generated samples.

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

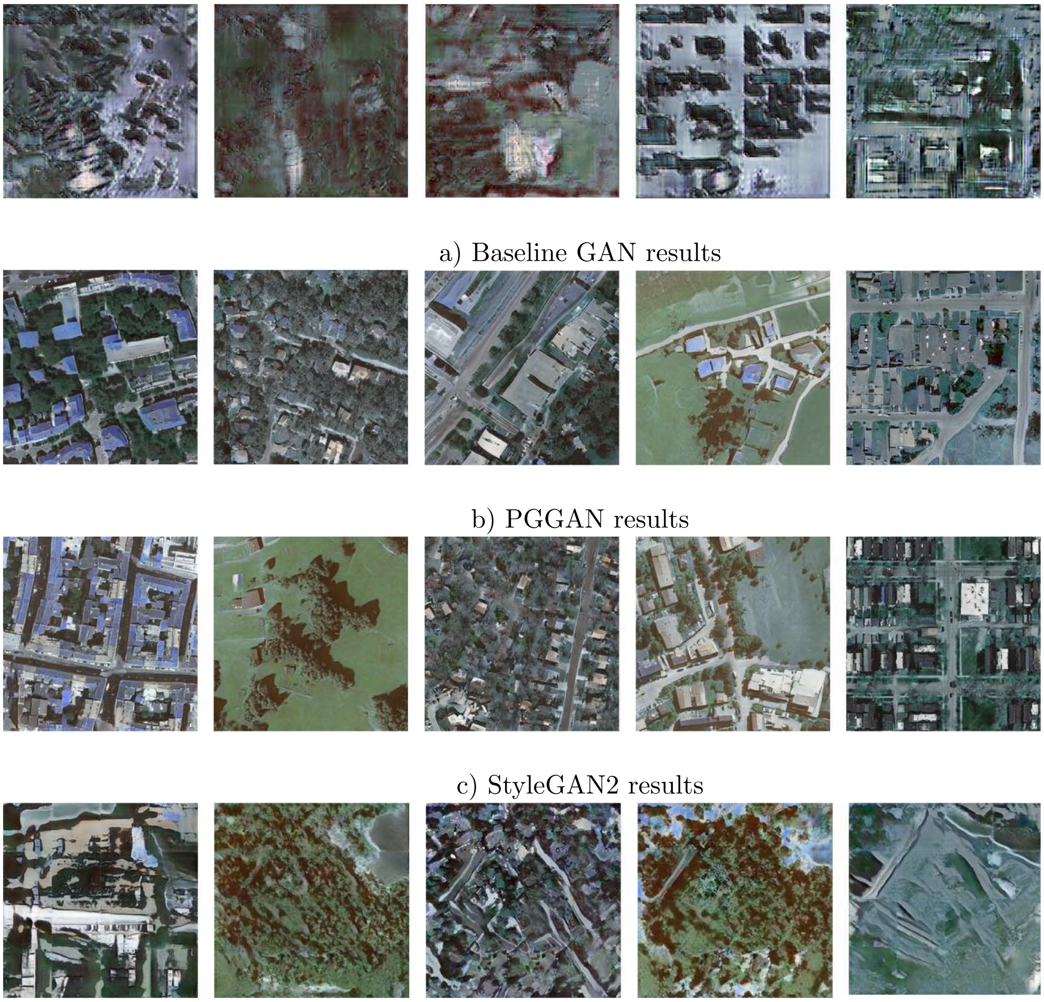
benchmark metrics beyond ImageNet models (Barratt and Sharma, 2018; Kynkäänniemi et al., 2022). In addition to the standard Inception Network we also present FID scores using an instance of this Inception model after fine-tuning on a section of the Open Cities Dataset (Global Facility for Disaster Reduction and Recovery (GFDRR) Labs, 2020), which consists of high quality urban and rural images of African cities.

3.4. User study

To further assess both the abilities to generate photo realistic aerial imagery from the best performing model, and also to explore the relevancy of FID and KID as measurements of image quality, an image detection study was run with human participants using samples generated from StyleGAN2.

A within groups, 2-alternative forced choice design was used for the study. This choice was made as it is an reputable and established experimental design (Hautus et al., 2021) for decision based visual search studies. The experiment was created using PsychoPy3 (Peirce et al., 2019) and hosted on Pavlovia.org (Peirce and MacAskill, 2018). 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.

Participants were given a set of image pairs (2) and for each pair asked to identify which image is fake. Each pair consisted of 1


a) Baseline GAN results

b) PGGAN results

c) StyleGAN2 results

d) CoCoGAN results

Fig. 7. Random selection of results from trained GANs.

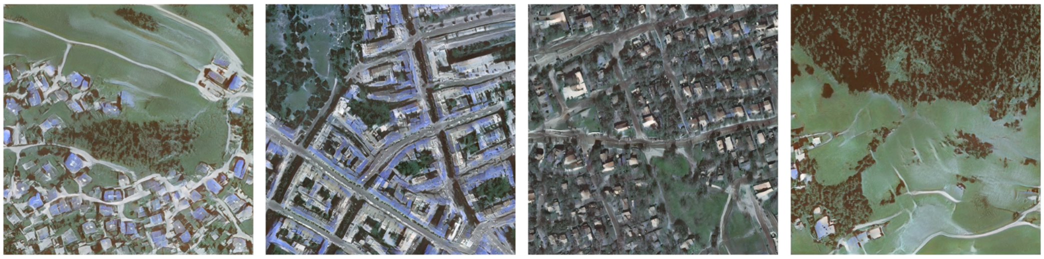


Fig. 8. Random selection of images from the StyleGAN2 1024×1024 model.

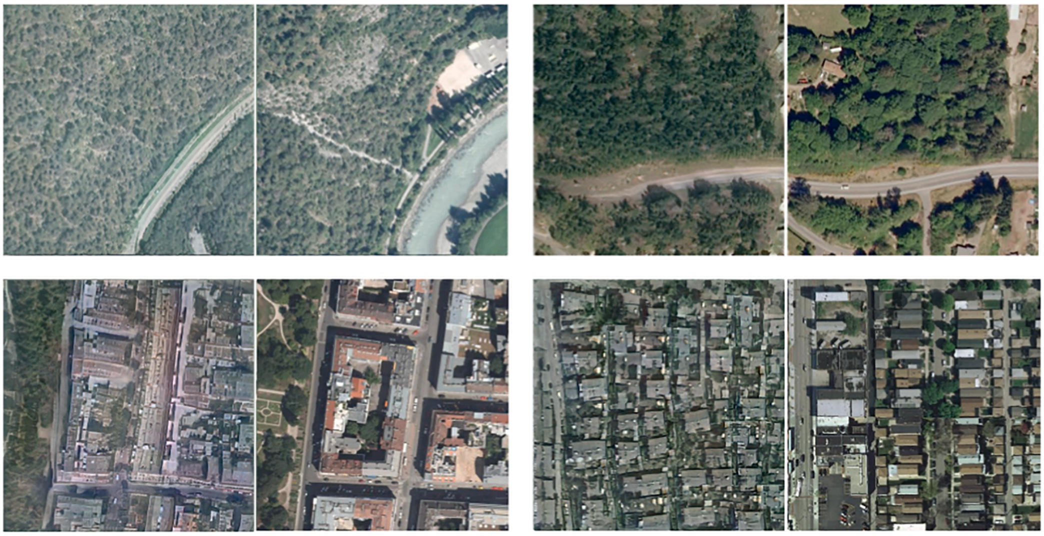


Fig. 9. Four pairs of real (right) images and their associated latent-space image (left) from the trained StyleGAN2 1024 model.

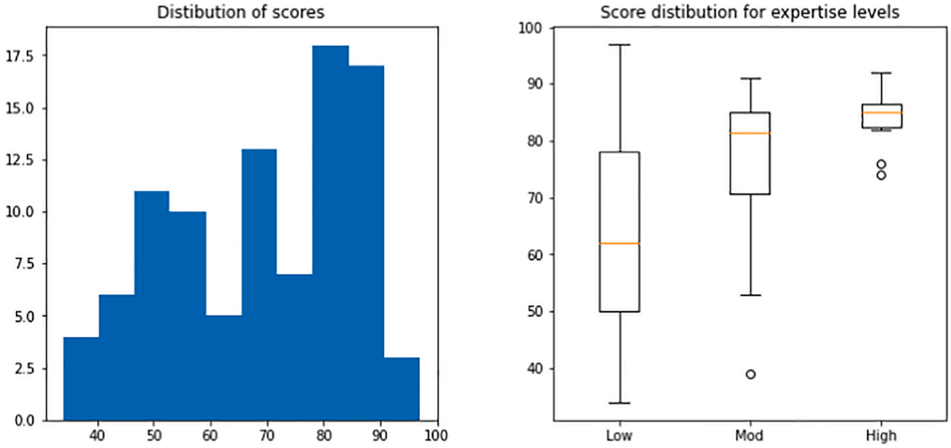


Fig. 10. Distribution of scores for groups of different experience levels.

Table 5
Experience level statistics from User study

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

Table 6
Pearson's Correlation Coefficient for StyleGAN2 images and participant Accuracy.

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

Table 7
FID metrics for different Inception models (lower values are better).

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

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 an initial practice round, the main task consisted of 100 image pairs in blocks of 25, with no feedback given after answering. 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. Participants were given as much time as they would like to answer. Previous works (Zhou et al., 2019; Elsayed et al., 2018) have opted to implement time constraints for participants. In this case we are measuring 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 500 ms before the participant is allowed to answer, this is to avoid mis-clicks or the participant making guesses without looking at the stimuli first.

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 with that self-perceive to have higher expertise will show higher accuracy than lower expertise.

4. Results

In this section we provide a quantitative and qualitative analysis of our results, before presenting further evaluation in the discussion (5).

4.1. Model performance

FID and KID was compared for each model across 10 subsets of 10,000 generated images. 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 Appendix. FID and KID was calculated separately for the 250 image dataset that was used as the stimulus in the user study. (4).

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 (Bińkowski et al., 2018), meaning that FID can only be fairly compared in tests with an equal n value.

Table 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 Fig. 7, with additional images found in Appendix B.

StyleGAN2 was found to produce the highest quality images, both in terms of metrics, as shown in Table 2, and in terms of visual results in Fig. 7. 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. Despite the generate images being of poorer quality than those in the original paper (Karras et al., 2019b), they demonstrate that aerial images can 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 (Chai et al., 2020) 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 Fig. B.13, with details such as trees and houses present. There are, however, some noticable 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.

The FID scores presented in the original paper (Lin et al., 2019) 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 Lin et al. (2019), 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 (Lin et al., 2019). 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 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. Further tuning of the hyper-parameters, or using a larger dataset could potentially diminish the prominence of these and result in higher image clarity.

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 Fig. B.12 in Appendix B, the limitations of the network are clear. The model has learned to capture the low level features int he 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 (Wu et al., 2017). 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.

As StyleGAN2 achieved the best performance at generating images at 256×256 pixels by a wide margin (Table 2), we decided to further test its capabilities by generating high-definition aerial images. As can be observed, at $1024{\times}1024$ pixels, there is a drop in performance in comparison to lower-resolution models (Table 3) with images appearing blurry and features less fully rendered (Fig. 8). This larger model took longer to train than the 256×256 model but the difference was not overly substantial as seen in Table 1.

4.2. 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 (Karras et al., 2019b). The output images and a target real image are then both placed in a pretrained feature extractor (VGG16 Simonyan and Zisserman, 2014) 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 9. 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 distribution more accurately then their would be fewer differences in the images. 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.

4.3. StyleGAN2 user study results

We performed a user detection study, aiming 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. We 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 Fig. 10. This indicates We did find that self-reported user experience did have a positive correlation with accuracy (Table 5) 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.19).

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. If fake images were deployed in the wild against a less prepared users, we might expect a lower level of detection.

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.

For analysis, we manually separated the synthetic images into two groups, containing urban or rural scenes. We defined a rural scene as containing natural features such as forest across the majority (50%) of the image. We found that 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.

Correlations between the GAN metrics (FID/KID/ACC) were explored as seen in Table 6 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 better FID/KID scores on average than rural images but where more easily identified by participants (4). 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.

4.4. FID comparisons between inception models

As previously discussed in (Metrics 3), 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 (Deng et al., 2009) dataset. In our main evaluation of the GAN models we also use this instance of the Inception network, 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) we calculate a second set of FID metrics using a ImageNet pretrained Inception Network fine-tuned on an EO dataset (OpenCities Global Facility for Disaster Reduction and Recovery (GFDRR) Labs, 2020). These results can be seen in Table 7. For this comparison both sets of FID scores for each dataset were calculated using the official Pytorch implementation of the Inception Network which accounts for a slight variation in scores from the previously reported scores which used the Tensorflow model. 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.

5. Discussion

The aim of this paper has been to evaluate the generation of Earth Observation (EO) data using current state-of-the-art GAN models. We have also evaluated the extent to which human observers can spot synthetic images, and whether metric are a meaningful predictor of human visual perception in the EO domain. EO data presents a novel challenge, since these models are usually fine-tuned towards the generation of objects and faces. The main motivation behind our work came 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 be connected with 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, the drops in performance, which average to 61% in terms of FID, are large enough to raise questions over the generality of GANs for image synthesis. 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 features and the spatial relationships between them. 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 7). 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.

Another contributing factor is that the dataset used to train the models was relatively low at 36.4 K samples, in contrast, many commonly used datasets have well over 50 k samples. This does, however, highlight the ability for some models such as PGGAN and StyleGAN2 to be able to generate reasonably realistic looking images from a smaller-than-normal training sets.

This ability to perform well in terms of KID and FID with smaller datasets makes these models suitable for tasks were existing data for training is limited or unbalanced. As discussed in the *Introduction*, 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 also 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, CoCo-GAN) and ones from only a few years prior (DCGAN). Our user study confirms that these networks are capable of tricking humans into misidentifying synthetic images as real, but also that common GAN evaluation metrics are often a poor reflection of human visual perception within the EO domain. While these metrics may provide a good indicator of how well generated samples match the distribution of the ground truth dataset, they do not necessarily account for image artefacts that may stand out more to the human eye. This disparity between automated and human evaluation metrics is supported by previous comparisons (Zhou et al., 2019) which found that correlations between human metrics and KID/FID varied between model, dataset and training instances. 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 (Kynkäänniemi et al., 2022).

5.1. User study limitations and future work

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.

Building on our results, future studies using additional measures focusing on visual attention during detection could provide more clarity and insight into this. Additionally the FID results (Table 7) 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 Exploration of new metrics that are better able to reflect the visual quality of this type of image should be explored.

6. Conclusion

In this paper, we presented a thorough study on the generation of fake aerial imagery using unconditional generative adversarial networks. Results from GAN evaluation metrics and user studies show that state-of-the-art GAN models, such as StyleGAN2, can successfully generate realistic examples of EO imagery to the point were they are hard to distinguish from real images to the naked human eye. We also demonstrate that Inception-based GAN evaluation metrics, namely FID and KID, are flawed measures of the visual quality of samples, as their correlation with human visual perception is dependant on the types of features present in the images.

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 large quantities of false information gives Security and Defense research a unique challenge to tackle. Future work should include training models on a wider range of aerial image datasets, as well as other sources of EO data, such as hyper-spectral imagery. It will also be important to design and evaluate new metrics for measuring quality of generated images that are more strongly aligned with judgement from human visual perception.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Inria benchmark dataset

Fig. A.11

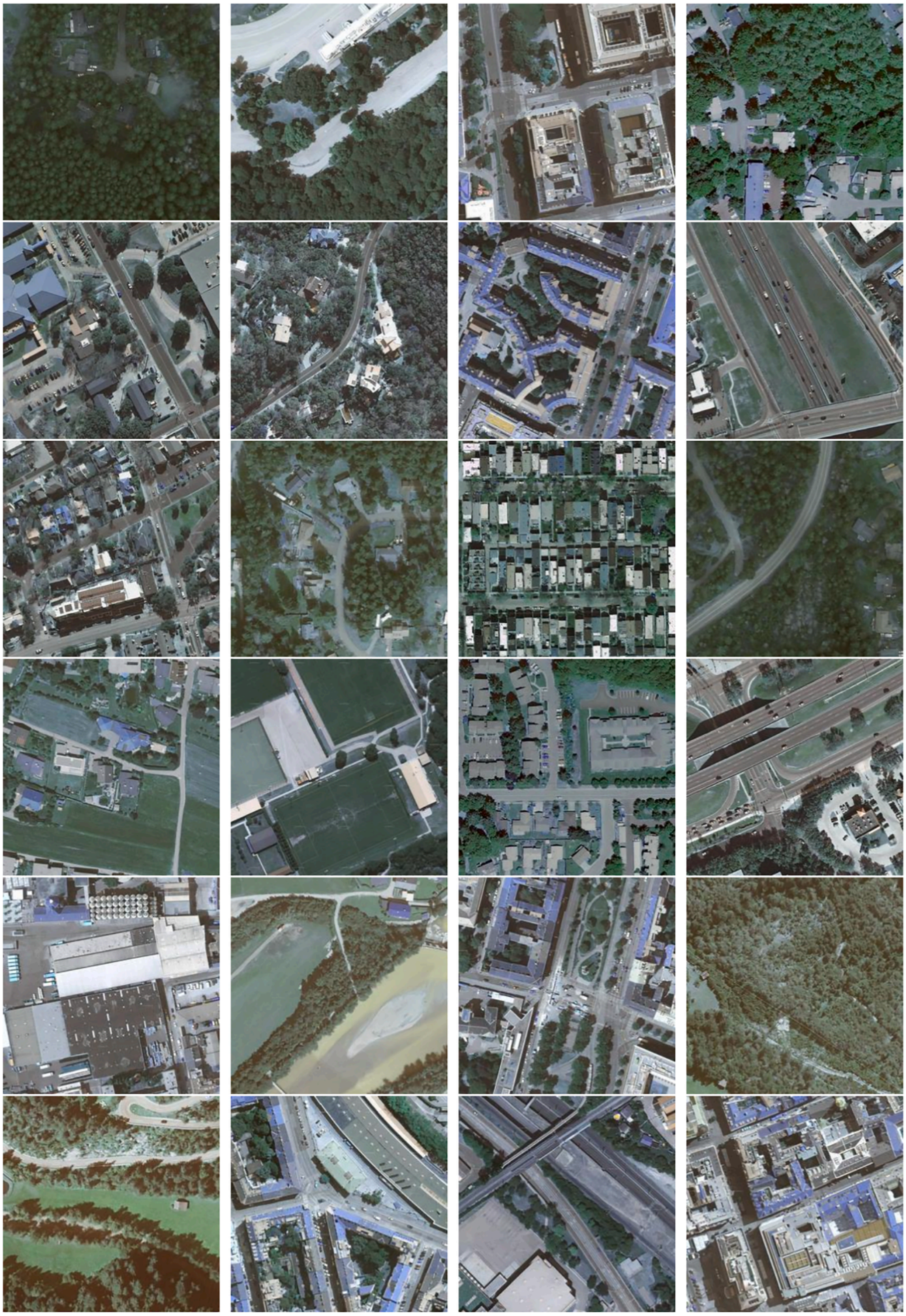


Fig. A.11. Randomly selected real images from the Inria Benchmark Aerial Imagery Dataset (256 × 256).

Appendix B. Additional results

Figs. B.12–B.18

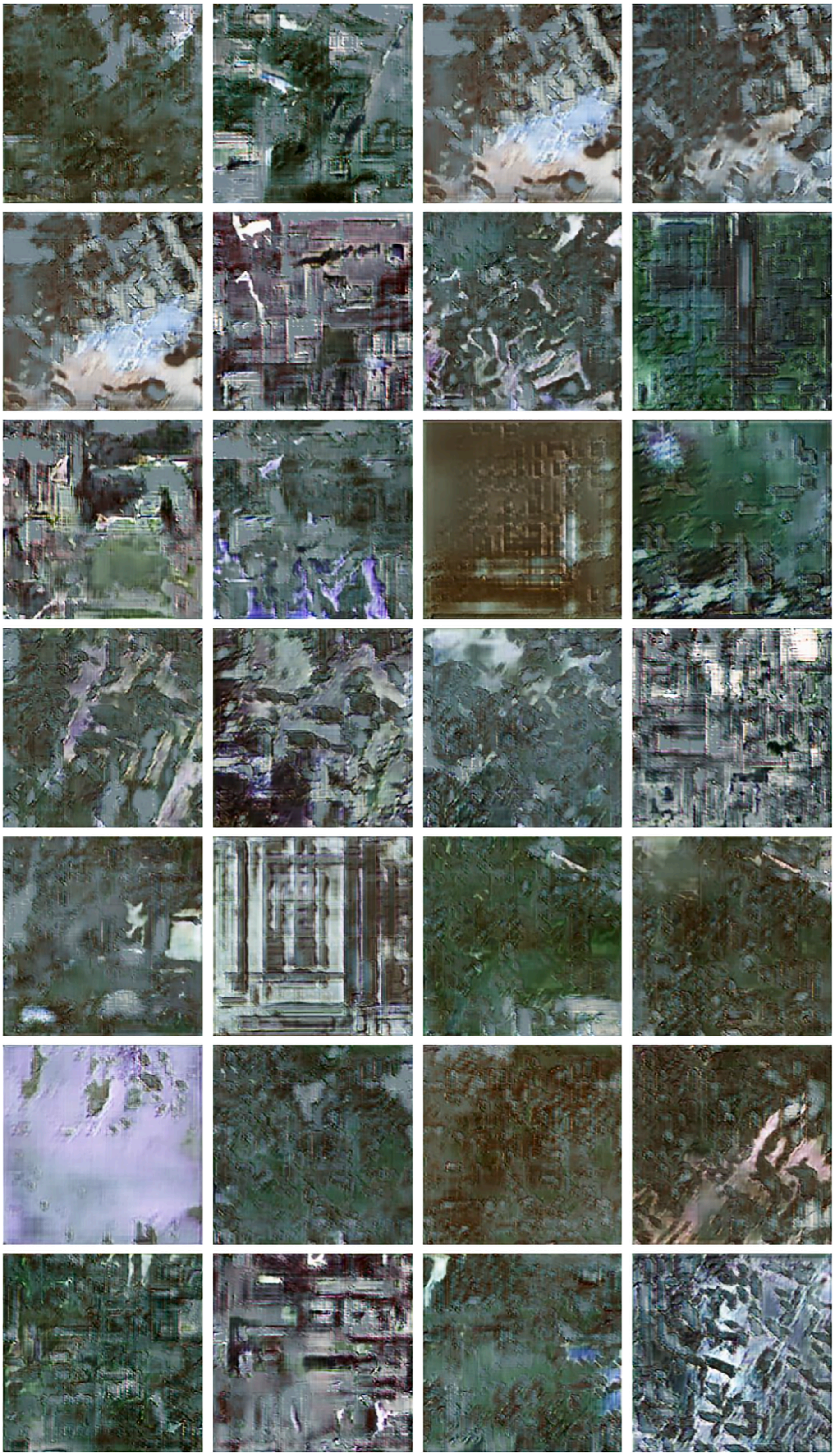


Fig. B.12. Randomly selected images generated from baseline DCGAN (256 × 256).

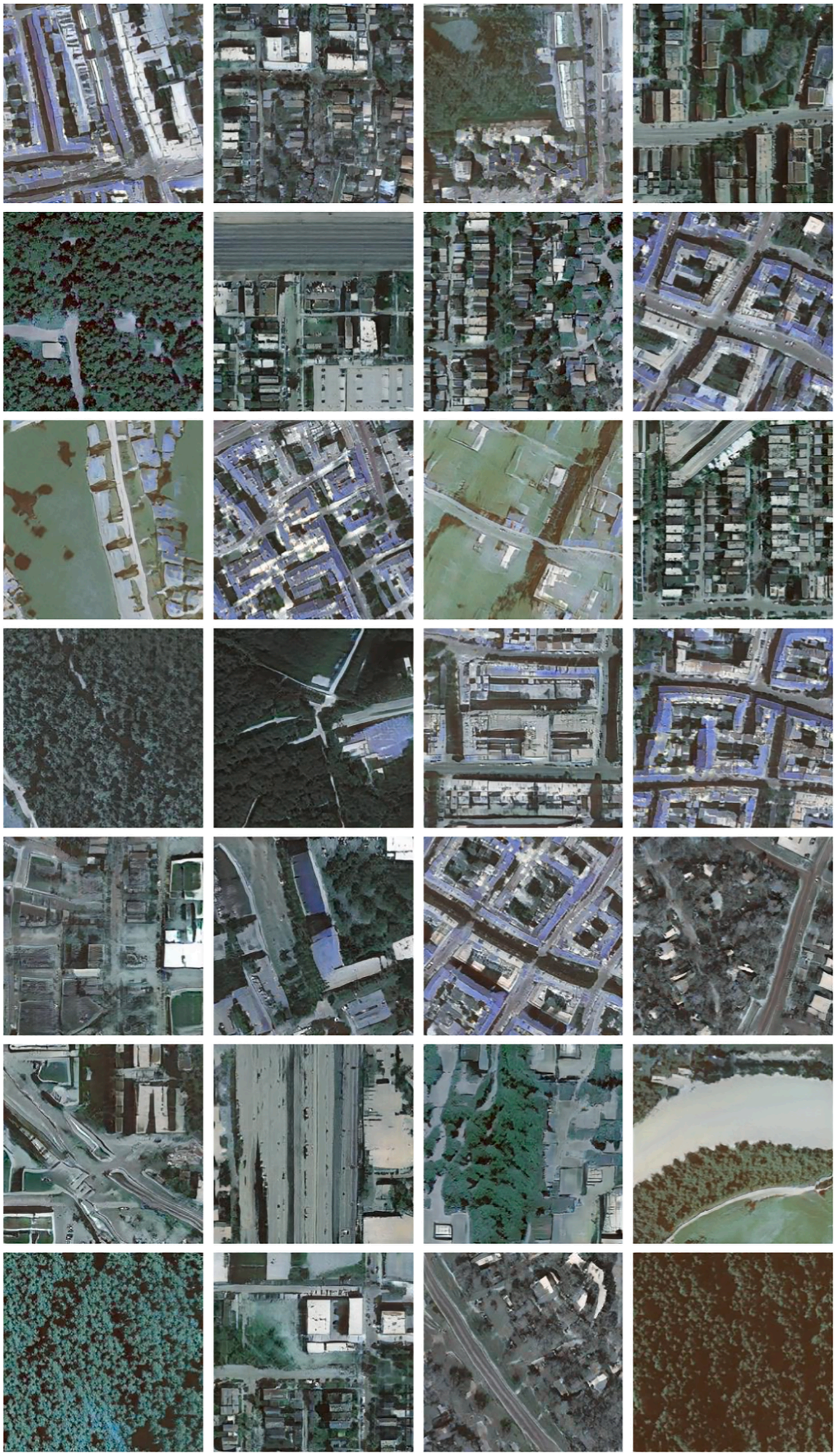


Fig. B.13. Randomly selected images generated from PGGAN (256 × 256).

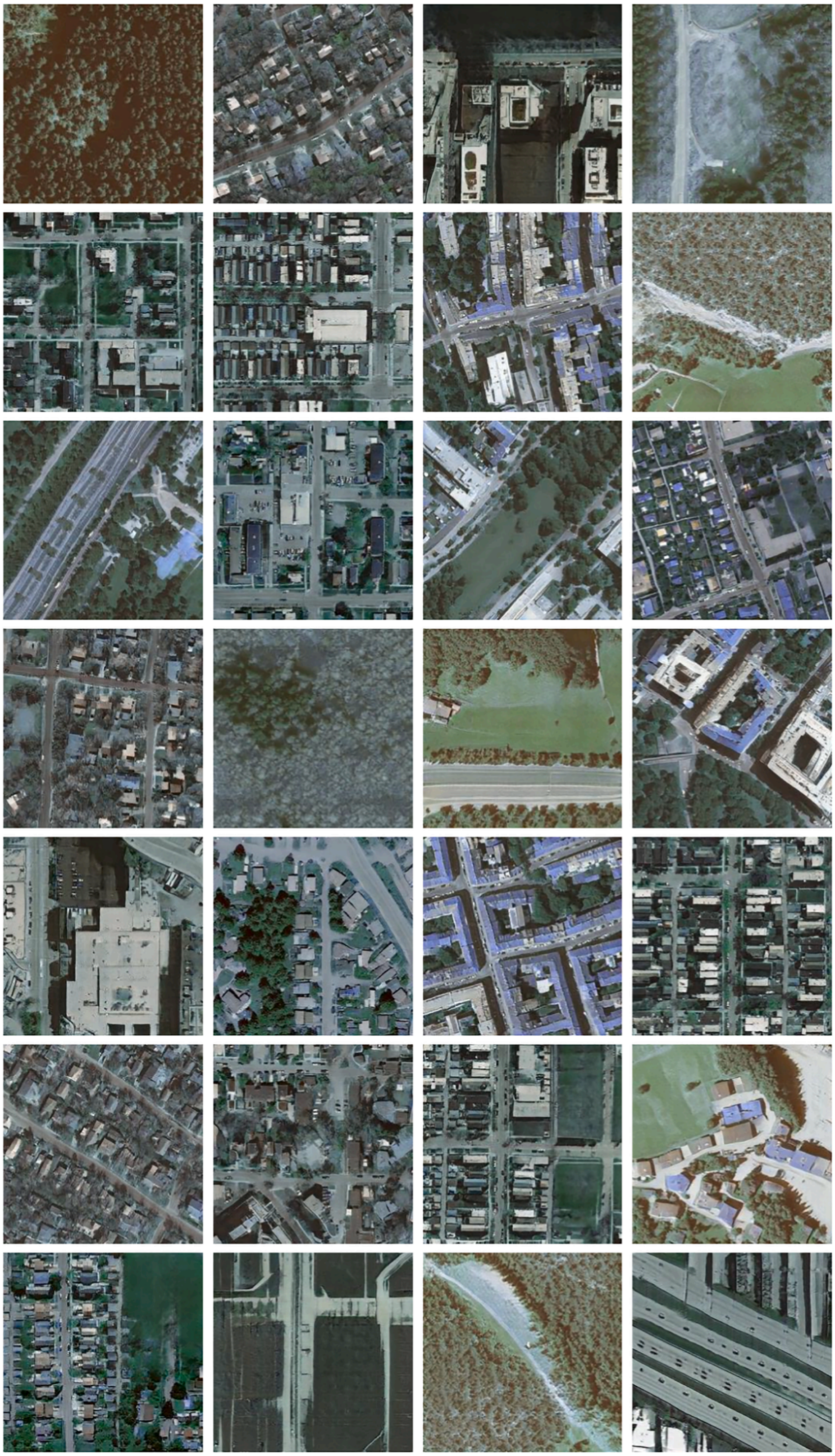


Fig. B.14. Randomly selected images generated from StyleGAN2 (256 × 256).

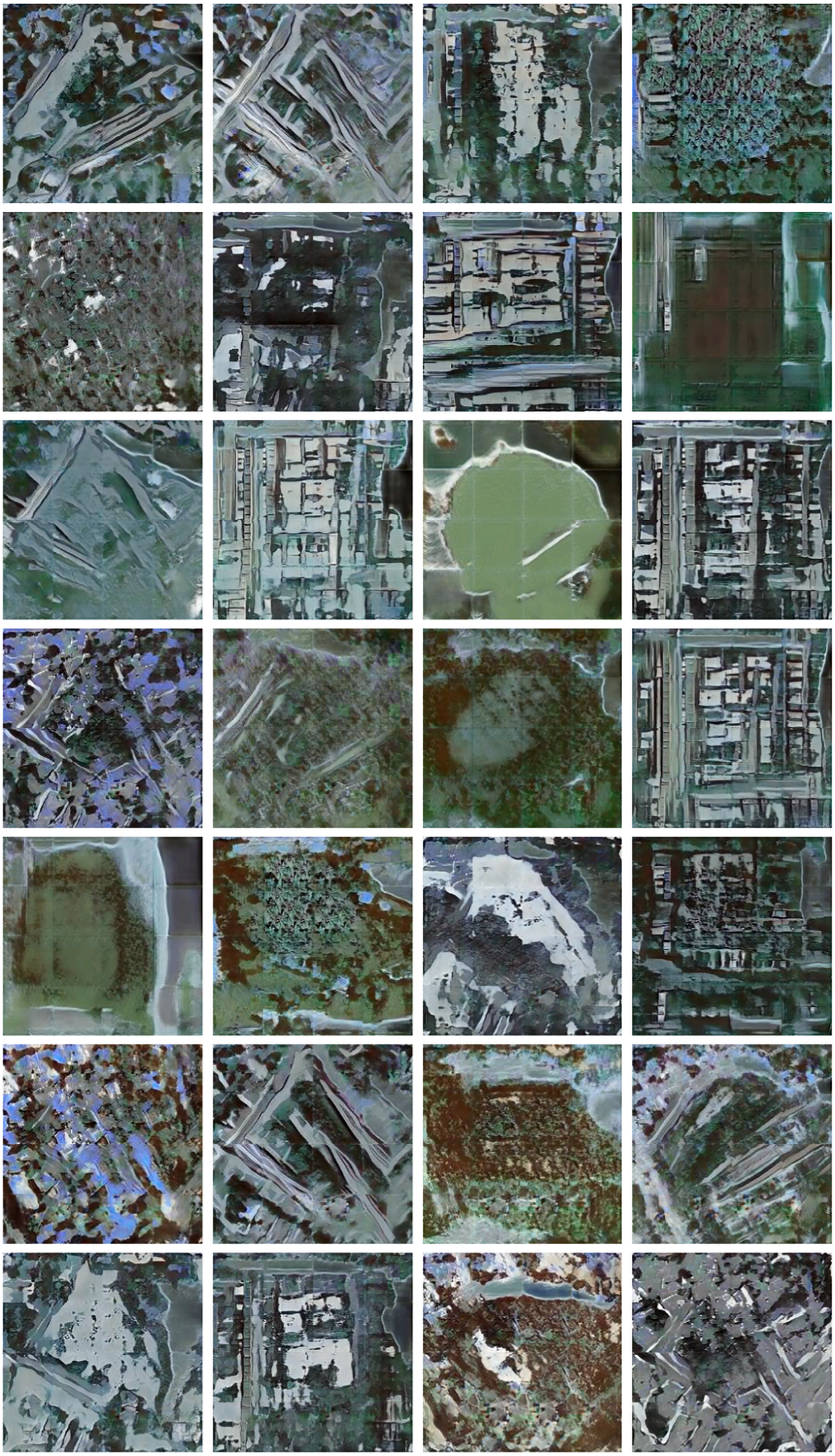


Fig. B.15. Randomly selected images generated from CoCoGAN (256 × 256).

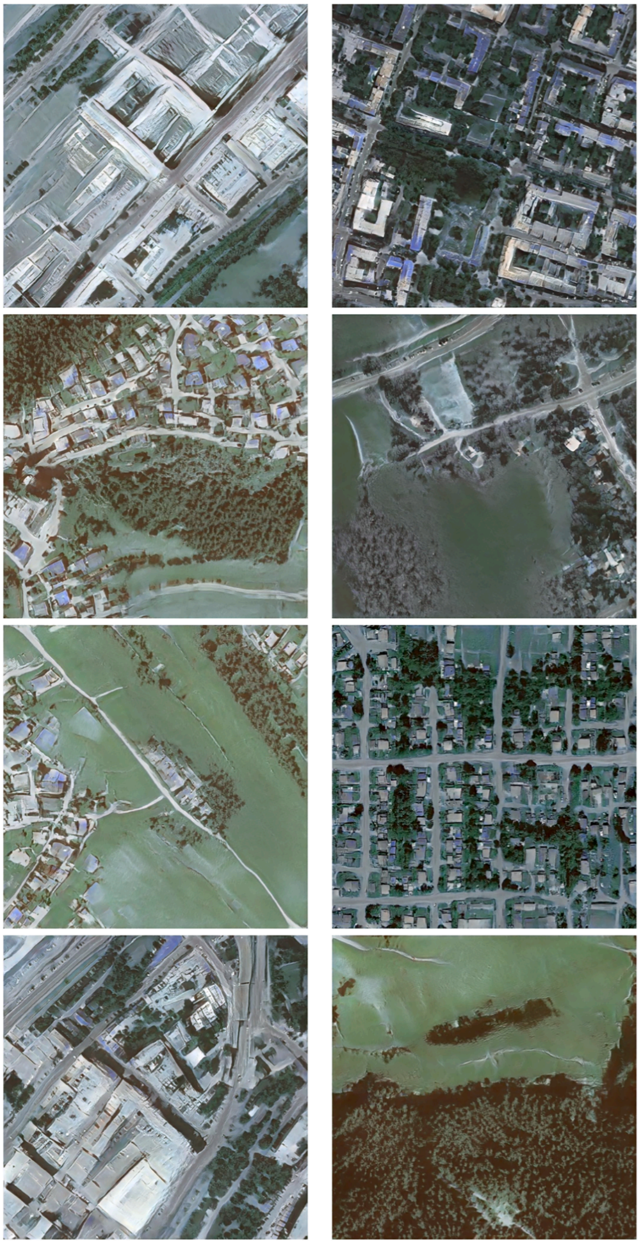


Fig. B.16. Randomly selected images generated from StyleGAN2 (1024×1024).

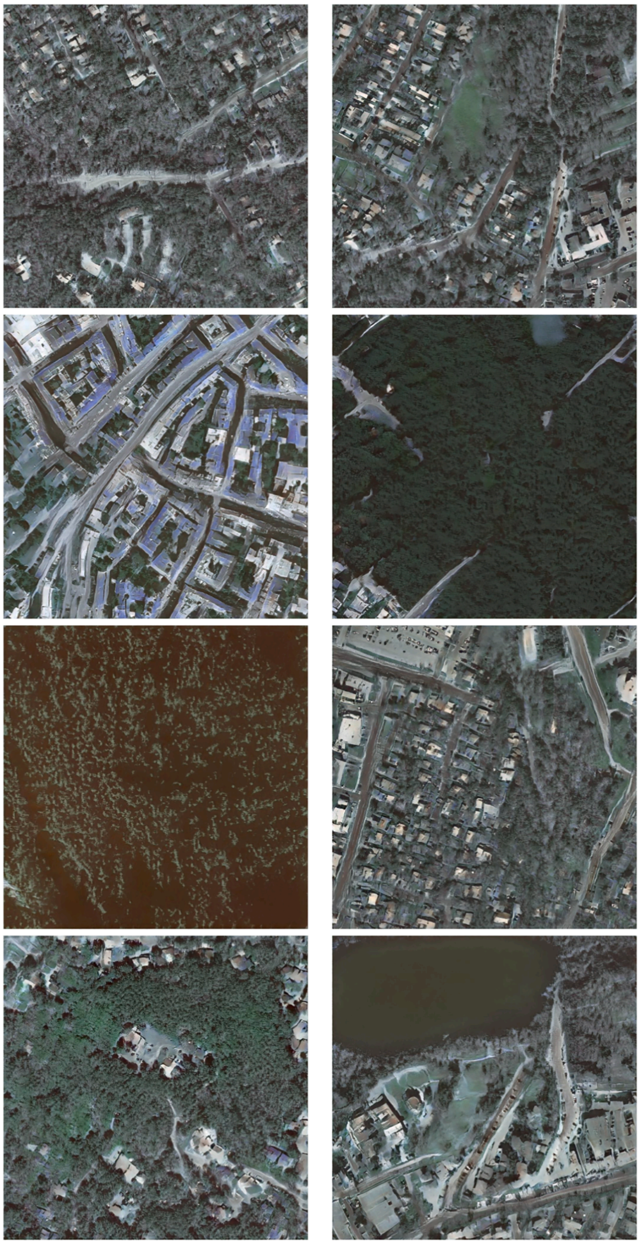


Fig. B.17. Randomly selected images generated from StyleGAN2 (1024×1024).

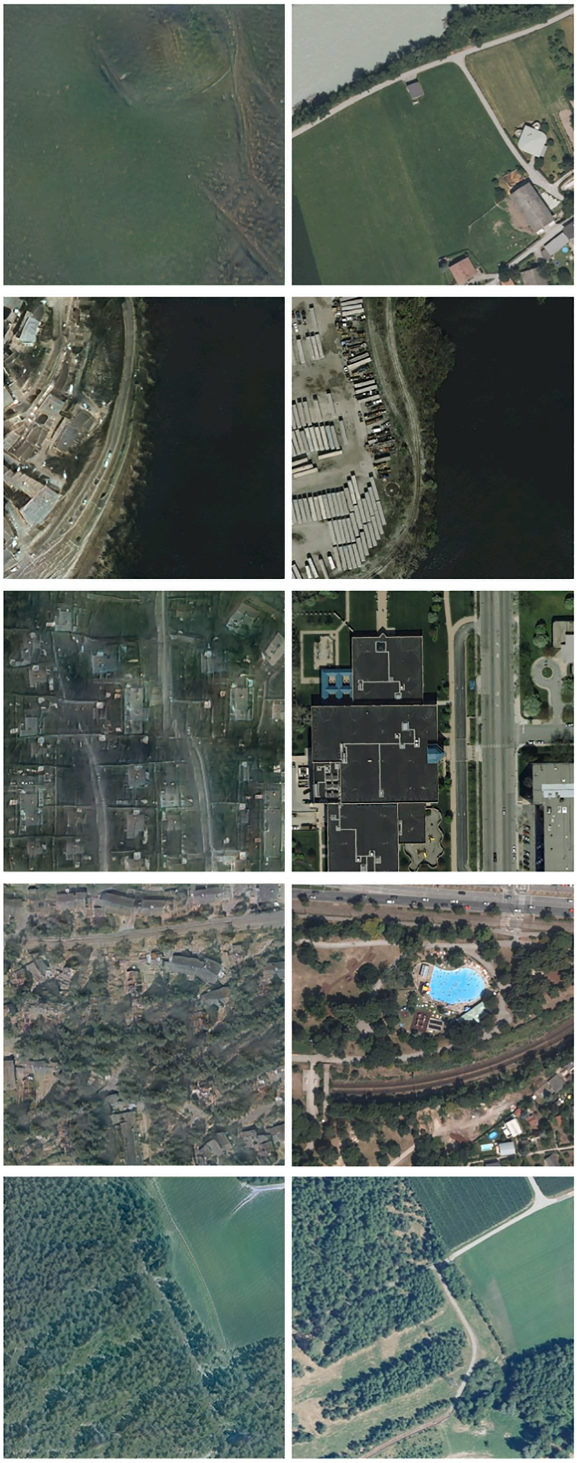


Fig. B.18. StyleGAN2 latent space representations (left) of target images from the Inria training dataset (right).

Appendix C. Kruskal-Wallis test between experience groups

Fig. C.19

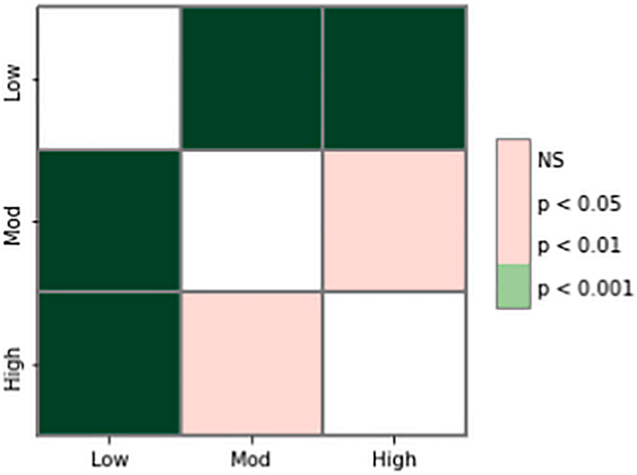


Fig. C.19. Pairwise comparison of experience levels. (H Statistic: 20.086, p : 0.001) Different distributions (reject H0) ε^2 : 0.223.

References

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Generative adversarial networks for realistic synthesis of hyperspectral samples. In: IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 4359–4362.

Barratt, S., Sharma, R., 2018. A note on the inception score. arXiv preprint arXiv: 1801.01973.

Bastos, M.T., Mercea, D., 2019. The Brexit botnet and user-generated hyperpartisan news. Soc. Sci. Comput. Rev. 37 (1), 38–54.

Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A., 2018. Demystifying MMD GANs. In: 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings.

Borji, A., 2018. Pros and Cons of GAN Evaluation Measures.

Boyd, D.S., Jackson, B., Wardlaw, J., Foody, G.M., Marsh, S., Bales, K., 2018. Slavery from space: Demonstrating the role for satellite remote sensing to inform evidence-based action related to UN SDG number 8. ISPRS J. Photogramm. Remote Sens. 142, 380–388.

Bredemeyer, S., Ulmer, F.-G., Hansteen, T.H., Walter, T.R., 2018. Radar path delay effects in volcanic gas plumes: the case of Láscar Volcano, Northern Chile. Remote Sens. 10 (10), 1514.

Brock, A., Lim, T., Ritchie, J.M., Weston, N., 2016. Neural photo editing with introspective adversarial networks. arXiv preprint arXiv:1609.07093.

Brock, A., Donahue, J., Simonyan, K., 2018. Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096.

Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J., Li, J., 2020. Exploring the Role of Visual Content in Fake News Detection. arXiv preprint arXiv:2003.05096.

Chai, L., Bau, D., Lim, S.-N., Isola, P., 2020. What makes fake images detectable? understanding properties that generalize. In: European Conference on Computer Vision. Springer, pp. 103–120.

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

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I., 2020. Generative pretraining from pixels. In: International Conference on Machine Learning. PMLR, pp. 1691–1703.

Chesney, B., Citron, D., 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. Calif. L. Rev. 107, 1753.

Chuvieco, E., Aguado, I., Yebra, M., Nieto, H., Salas, J., Martín, M.P., Vilar, L., Martínez, J., Martín, S., Ibarra, P., et al., 2010. Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. Ecol. Model. 221 (1), 46–58.

Collins, E., Bala, R., Price, B., Süsstrunk, S., 2020. Editing in Style: Uncovering the Local Semantics of GANs. arXiv preprint arXiv:2004.14367.

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

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

Dhariwal, P., Nichol, A., 2021. Diffusion models beat gans on image synthesis. arXiv preprint arXiv:2105.05233.

Do, Q.-T., Shapiro, J.N., Elvidge, C.D., Abdel-Jelil, M., Ahn, D.P., Baugh, K., Hansen-Lewis, J., Zhizhin, M., Bazilian, M.D., 2018. Terrorism, geopolitics, and oil security: Using remote sensing to estimate oil production of the Islamic State. Energy Res. Soc. Sci. 44, 411–418.

Donahue, J., Simonyan, K., 2019. Large Scale Adversarial Representation Learning. In: Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 32. Curran Associates Inc.

Donahue, C., McAuley, J., Puckette, M., 2018. Adversarial audio synthesis. arXiv preprint arXiv:1802.04208.

Donahue, J., Simonyan, K., 2019. Large scale adversarial representation learning. In: Advances in Neural Information Processing Systems, pp. 10541–10551.

Eklund, A., 2019. Feeding the zombies: Synthesizing brain volumes using a 3D progressive growing GAN. arXiv preprint arXiv:1912.05357.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., Sohl-Dickstein, J., 2018. Adversarial examples that fool both computer vision and time-limited humans. Adv. Neural Inform. Process. Syst. 31.

Esser, P., Rombach, R., Ommer, B., 2021. Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12873–12883.

Fan, S., Ng, T.-T., Koenig, B.L., Herberg, J.S., Jiang, M., Shen, Z., Zhao, Q., 2017. Image visual realism: From human perception to machine computation. IEEE Trans. Pattern Anal. Mach. Intell. 40 (9), 2180–2193.

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

Global Facility for Disaster Reduction and Recovery (GFDRR) Labs, 2020. Open Cities AI Challenge Dataset. https://doi.org/10.34911/RDNT.F94CXB. URL: https://registry.mlhub.earth/10.34911/rdnt.f94cxb.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets.

Gui, J., Sun, Z., Wen, Y., Tao, D., Ye, J., 2020. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. arXiv preprint arXiv:2001.06937.

Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S., 2020. GANSpace: Discovering Interpretable GAN Controls. arXiv preprint arXiv:2004.02546.

Haslett, S.K., Wong, B.R., 2019. Reconnaissance survey of coastal boulders in the Moro Gulf (Philippines) using Google Earth imagery: Initial insights into Celebes Sea tsunami events. Bull. Geol. Soc. Malaysia 68, 37–44.

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

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

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

Jiang, K., Wang, Z., Yi, P., Wang, G., Lu, T., Jiang, J., 2019. Edge-Enhanced GAN for Remote Sensing Image Superresolution. IEEE Trans. Geosci. Remote Sens. 57 (8), 5799–5812.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. 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, 2018.

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

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

Kim, J., Hong, S.-A., Kim, H., 2019. A StyleGAN Image Detection Model Based on Convolutional Neural Network. J. Korea Multimedia Soc. 22 (12), 1447–1456.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114.

Kolchinski, Y.A., Zhou, S., Zhao, S., Gordon, M., Ermon, S., 2019. Approximating Human Judgment of Generated Image Quality. arXiv preprint arXiv:1912.12121.

Krizhevsky, A., Nair, V., Hinton, G., 2014. The cifar-10 dataset, online: http://www.cs.toronto.edu/kriz/cifar.html 55.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., Lehtinen, J., 2022. The Role of ImageNet Classes in Fr\echet Inception Distance, arXiv preprint arXiv:2203.06026.

Lee, H., Ra, M., Kim, W.-Y., 2020. Nighttime data augmentation using GAN for improving blind-spot detection. IEEE Access 8, 48049–48059.

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

Liu, Z., Luo, P., Wang, X., Tang, X., 2015. Deep Learning Face Attributes in the Wild. In: Proceedings of International Conference on Computer Vision (ICCV).

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

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. 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, pp. 3226–3229.

Malarvizhi, K., Kumar, S.V., Porchelvan, P., 2016. Use of high resolution google earth satellite imagery in landuse map preparation for urban related applications. Procedia Technol. 24, 1835–1842.

Manderson, T., Wapnick, S., Meger, D., Dudek, G., 2020. Learning to Drive Off Road on Smooth Terrain in Unstructured Environments Using an On-Board Camera and Sparse Aerial Images. arXiv preprint arXiv:2004.04697.

Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., Malossi, C., 2018. Bagan: Data augmentation with balancing gan. arXiv preprint arXiv:1803.09655.

Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L., 2018. Detection of GAN-Generated Fake Images over Social Networks. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). IEEE, pp. 384–389.

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

Mirsky, Y., Mahler, T., Shelef, I., Elovici, Y., 2019. CT-GAN: Malicious tampering of 3D medical imagery using deep learning. In: 28th {USENIX} Security Symposium ({USENIX} Security 19), pp. 461–478.

Oubrahim, Y., Lbazri, S., Ounacer, S., Rachik, A., Moulouki, R., Azzouazi, M., 2018. A new architecture for monitoring land use and land cover change based on remote sensing and GIS: A data mining approach. Periodicals Eng. Nat. Sci. 6 (2), 406–414.

Pan, Z., Yu, W., Yi, X., Khan, A., Yuan, F., Zheng, Y., 2019. Recent progress on generative adversarial networks (GANs): A survey. IEEE Access 7, 36322–36333.

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

Peirce, J., MacAskill, M., 2018. Building experiments in PsychoPy. Sage.

Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J.K., 2019. PsychoPy2: Experiments in behavior made easy. Behav. Res. Methods 51 (1), 195–203.

Perez, L., Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. In: Advances in neural information processing systems, pp. 2234–2242.

Sandfort, V., Yan, K., Pickhardt, P.J., Summers, R.M., 2019. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. Sci. Rep. 9 (1), 1–9.

Sanyal, A., Torr, P.H., Dokania, P.K., 2019. Stable Rank Normalization for Improved Generalization in Neural Networks and GANs. arXiv preprint arXiv:1906.04659.

Scheufele, D.A., Krause, N.M., 2019. Science audiences, misinformation, and fake news. Proc. Nat. Acad. Sci. 116 (16), 7662–7669.

Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H., 2017. Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter 19 (1), 22–36.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

Singh, P., Komodakis, N., 2018. 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, pp. 1772–1775.

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

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, PMLR, pp. 2256–2265.

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

Tanaka, F.H.K.D.S., Aranha, C., 2019. Data augmentation using GANs. arXiv preprint arXiv:1904.09135.

Taylor, L., Nitschke, G., 2017. Improving deep learning using generic data augmentation. arXiv preprint arXiv:1708.06020.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J., 2020. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. arXiv preprint arXiv:2001.00179.

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

Viazovetskyi, Y., Ivashkin, V., Kashin, E., 2020. StyleGAN2 Distillation for Feed-forward Image Manipulation. arXiv preprint arXiv:2003.03581.

Vincent, J., 2017. All of these faces are fake celebrities spawned by AI.

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., Pinheiro, P.R., 2020. Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. IEEE Access 8, 91916–91923.

Wang, X., Yu, K., Dong, C., Change Loy, C., 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 606–615.

Wang, R., Ma, L., Juefei-Xu, F., Xie, X., Wang, J., Liu, Y., 2019. Fakespotter: A simple baseline for spotting ai-synthesized fake faces. arXiv preprint arXiv:1909.06122.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., Efros, A.A., 2019. CNN-generated images are surprisingly easy to spot... for now. arXiv preprint arXiv:1912.11035.

Wang, L., Chen, W., Yang, W., Bi, F., Yu, F.R., 2020. A State-of-the-Art Review on Image Synthesis With Generative Adversarial Networks. IEEE Access 8, 63514–63537.

Wu, X., Xu, K., Hall, P., 2017. A survey of image synthesis and editing with generative adversarial networks. Tsinghua Sci. Technol. 22 (6), 660–674.

Xian, Y., Sharma, S., Schiele, B., Akata, Z., 2019. f-vaegan-d2: A feature generating framework for any-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10275–10284.

Xiong, W., Luo, W., Ma, L., Liu, W., Luo, J., 2018. Learning to Generate Time-Lapse Videos Using Multi-Stage Dynamic Generative Adversarial Networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Yildirim, G., Jetchev, N., Vollgraf, R., Bergmann, U., 2019. Generating High-Resolution Fashion Model Images Wearing Custom Outfits. In: Proceedings of the IEEE International Conference on Computer Vision Workshops.

Yin, W., Fu, Y., Sigal, L., Xue, X., 2017. Semi-latent gan: Learning to generate and modify facial images from attributes. arXiv preprint arXiv:1704.02166.

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

Yu, N., Davis, L., Fritz, M., 2019. Learning GAN fingerprints towards Image Attribution. arXiv preprint arXiv:1811.08180.

Zhang, X., Wang, Z., Liu, D., Ling, Q., 2019. Dada: Deep adversarial data augmentation for extremely low data regime classification. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2807–2811.

Zhang, Z., Pan, X., Jiang, S., Zhao, P., 2020. High-quality face image generation based on generative adversarial networks. J. Vis. Commun. Image Represent. 71, 102719.

Zhao, B., Zhang, S., Xu, C., Sun, Y., Deng, C., 2021. Deep fake geography? When geospatial data encounter Artificial Intelligence. Cartogr. Geogr. Inform. Sci. 48 (4), 338–352.

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

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232.