Building a World With Deepfake Content – Who Needs Real Data?

Graham Long
Thales UK
Crawley, West Sussex
Graham.Long@uk.thalesgroup.com

ABSTRACT

Synthetic content can be artificially produced, manipulated, or modified using artificial intelligence (AI) to create a wide range of synthesized data from text, prose, music, or images to videos. When used for a malign purpose to mislead or deceive, this synthetic content has come to be known as "deepfakes". But beyond the potentially sinister, headline-grabbing generation of fake videos of world leaders, deepfake technology can be applied to create content and data valuable to the construction of synthetic environments.

Data describing the physical world is now very abundant and provides a rich data source to support synthetic environment development. Still, there are many scenarios where this data may not be available or suitable. In these situations, artificially created data can provide data where none exists, augment existing data, improve data quality to provide accurate representations of real features, data and places, or even create entirely fictitious data and environments.

Generative artificial intelligence is one of the most promising advances in AI in the past decade. Generative models produce synthetic content by learning to mimic a data distribution and generate new, similar, credible content. For example, they can create entirely artificial satellite imagery, increase image resolution or remove artefacts with inpainting. They can also generate or manipulate other data types, such as land cover, point clouds, maps, 3D models, or even the design and style of the entire environment, all of which will appear authentic but are, in fact, entirely artificial

This paper will explore the current state and capabilities of generative models. It will identify those model types, such as Generative Adversarial Networks, that are most suited to generating synthesized content for synthetic environments. Finally, it will illustrate how these models can be applied to specific types of content and use cases and evaluate the currently achievable results.

ABOUT THE AUTHORS

Graham Long has worked in the training and simulation industry for over 30 years. He has extensive experience of synthetic environment development and is a Thales synthetic environment specialist. His roles and responsibilities have included managing engineering teams in the development and deployment of visual system solutions, synthetic environment production, development of synthetic environment generation tools and processes and interoperability standards. These activities have involved delivering solutions to over 20 civil and military simulation programmes, internal and external research and development projects as well the application of these techniques to Digital Twins.

Building a World With Deepfake Content – Who Needs Real Data?

Graham Long
Thales UK
Crawley, West Sussex
Graham.Long@uk.thalesgroup.com

INTRODUCTION

The application of Artificial Intelligence (AI) to the generation of entirely synthetic data, such as images and videos has given rise to the concept of "deepfake" technology that can create "fake" data that is so plausible it appears indistinguishable from real data. Deepfakes have gained prominence mainly by demonstrating their capability to generate and manipulate images of humans. ThisPersonDoesNotExist.com generates a new fake, but completely realistic facial image each time a user refreshes the site. A recent study (Nightingale, Farid, 2021) found that the photorealism of such AI-synthesized faces has now progressed to the point that the fake faces are indistinguishable - and more trustworthy - than real faces. When asked to distinguish fake photos from real photographs of people, study participants only achieved a slightly worse than chance accuracy rate of 48.2 percent.

The potential for sinister, malicious applications of deepfake technology to deceive and mislead is widespread. The first deepfake video emerged in 2017 when a Reddit user swapped faces of celebrities into pornographic videos using deep learning and posted the fake videos online. In 2018, BuzzFeed used the Reddit user's software (FakeApp) to generate and release a deepfake video of former president Barak Obama giving a talk about deepfakes (Mirsky, Lee, 2020). These fakes use face-swapping, lip-syncing and puppet-master techniques to superimpose face images, synchronize audio, and animate facial expressions to produce an authentic-looking but fake video (Nguyen et al, 2021).

The same AI technology may also be used to adversely impact tactical decisions or mission planning by deceiving AI-assisted image analysis with manipulated satellite data containing false content (Tucker, 2019). More generally, the increasingly widespread use and reliance on open-source images creates the possibility that applications such as Google Maps or autonomous vehicle operations are vulnerable to serious disruption if just a handful of AI manipulated data sets are maliciously entered into the open-source image supply line.

However, despite the ethical, legal, privacy, personal, and national security concerns arising from the use and abuse of deepfake technology, this same technology can be constructively applied to the AI creation of art, music, prose, visual effects, digital avatars, digital mapping, realistic video dubbing or virtually trying on clothes while shopping. There are also many potential opportunities and benefits to be realized from incorporating AI-generated synthetic data into synthetic environments (SE) - from providing a lower-cost alternative to real data, enhancing or augmenting its content or fidelity, or even producing entirely artificial, fictitious environments.

GENERATIVE MODELLING

The term deepfake is derived from the combination of 'deep learning' and 'fake' and primarily relates to content generated by artificial neural networks incorporated into Deep Learning generative model architectures. Deep Generative models (DGM) (Ruthotto, Haber, 2021) belong to a class of statistical models that can generate new data instances. They can automatically discover and learn the underlying hidden structure and patterns in input data and use this knowledge in such a way that the model can generate new, plausible data examples that appear indistinguishable from the original dataset.

Machine learning models are classified as generative or discriminative models. Unlike generative models, discriminative models cannot generate new data instances; they are conditional models that learn the boundaries

between classes or labels in a dataset so that one class can be separated from another. Discriminative models are mainly used for supervised machine learning, which aims to learn a mapping function between labelled input and output data. Supervised learning is the dominant form of deep learning today, and discriminative models are behind many common applications of deep learning to tasks such as object detection or image classification. In contrast, DGMs are a form of unsupervised deep learning and do not require labelled data. Instead, they possess the powerful capability to learn the inherent structure and patterns of unlabelled data and use this knowledge to generate new data instances based on these learned patterns.

DGMs aim to learn the training set's true data distribution to enable them to generate new data points. However, any training set or observed data is only a finite set of samples from an underlying distribution, and it is not always possible to learn the exact distribution of this data, either implicitly or explicitly. Therefore, the model aims to learn a representation of the distribution that is as similar as possible to the true data distribution, (Figure 1), then sample from this learned distribution to generate new data points that appear real and are variations from the images in the training data.

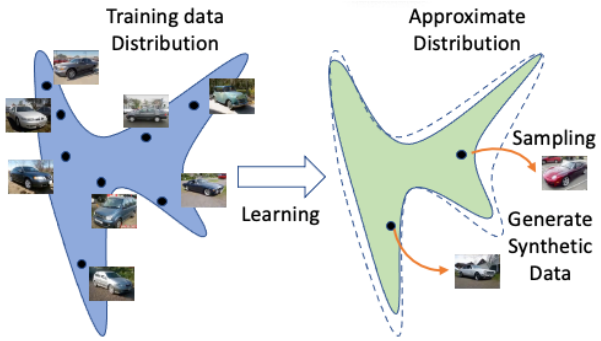

Figure 1 Data Distributions

Learning a function to approximate the model distribution to the true distribution is made difficult and complex by the very high dimensionality of data like imagery – it does not follow a simple normal distribution that can be predicted with a known function. However, applying neural networks to this task is very powerful because they can learn extraordinarily complex functional mappings and estimates of high dimensional data distributions.

In the case of computer vision tasks, to understand an image dataset, neural networks need to learn a representation of the underlying features of the images. These unobservable or latent features are captured as lower-dimensional latent variables in a latent space. In simple terms, the latent space is a representation of compressed data. This compression makes it easier to discover patterns while providing an improved understanding of the data's overall behaviour and enables the model to efficiently incorporate the intrinsic nature of the data as the basis for generating new synthetic data.

DGM Architectures

The family of generative models can be subdivided into two major branches – models that try to explicitly define a parameterised probability density function, and implicit models that attempt to learn to produce samples from a learnt data distribution without explicitly defining any probability density function (Goodfellow, 2017). Two of the most common DGM architectures are Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN).

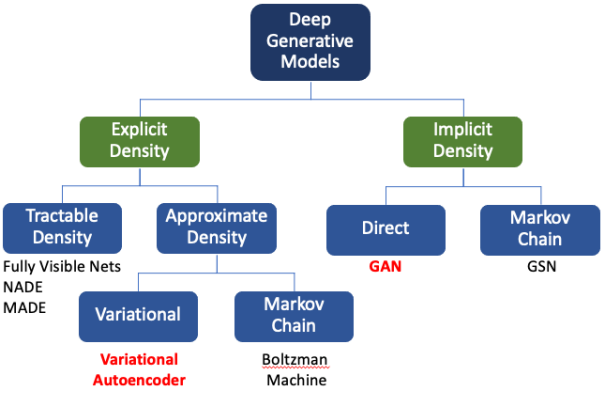

Figure 2 Generative Model Taxonomy

Standard autoencoder models essentially learn to output whatever they receive as input - they consist of an encoder network that captures a latent representation of the features of the input data distribution and a decoder network that effectively applies this process in reverse to reconstruct the input data. Variational autoencoders (Soleimany, 2022) are an evolution of this architecture designed to enable the model to generate new data. In a standard autoencoder, each attribute of the latent state is normally represented by a single, discrete variable. VAEs adopt a probabilistic approach and represent these latent vectors in a defined

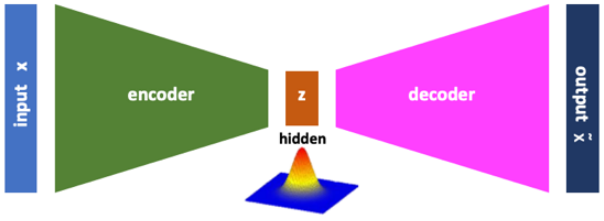

Figure 3 Variational Autoencoder

probability distribution. The decoder then randomly samples from these learned latent distributions to generate new data that is unique every time.

Generative Adversarial Networks (Goodfellow et al, 2014) fall into the category of implicit density models – they do not try and explicitly compute the density distribution like Variational Autoencoders but attempt to learn the approximation of the underlying distribution through the training process. They employ an adversarial, game theory approach that represents a zero-sum game between two competing neural networks, a generator, and a discriminator. The rules of this minimax game are straightforward – the generator aims to produce realistic data, and the discriminator aims to determine fake from real. The performance of both networks improves during training, but the objective is for the generator to become so good at producing realistic data that the discriminator cannot distinguish it from real data.

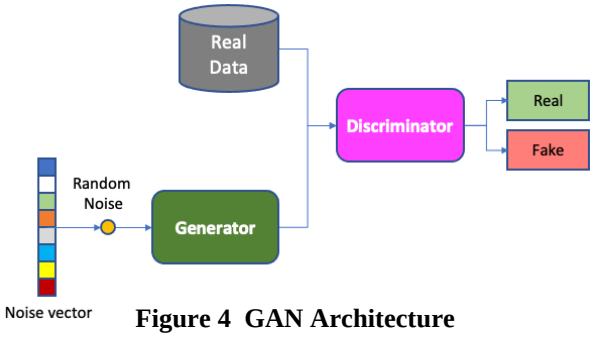

Figure 4 GAN Architecture

The generator model generates new data by receiving a random noise input drawn from a Gaussian distribution and transforms this noise into a data sample. The generator applies meaning to points in the latent space (Brownlee, 2019) and then draws new points from this latent space as input for creating new and different output data samples. The discriminator receives an example image input, either a real image from the training data or an image generated by the generator model. The discriminator is a normal classifier model and predicts if the input is real or fake. In effect, the discriminator determines if the generated images appear to be a realistic variation of the images from the training set. After training, the generator is kept, and the discriminator is discarded.

APPLICATION OF GENERATIVE MODELLING

GANs and VAEs are the principle generative modelling architectures behind most of the deepfakes that spawned the term and gained widespread public attention. For example, deepfake videos employ VAEs to apply believable "face-swapping" techniques that capture lighting conditions and identify distinctive face features for retention and those that are replaceable. The face swap is then executed frame-by-frame by merging source data with the deepfake target face and using the VAE encoder to reconstruct the movements and emotional expressions of the source with the target face.

Generally, the definition of a generative model means that any technique can be used on any modality or task; however, some models are more suited for specific tasks. For example, although VAEs can generate fake images of human faces (or potentially images of any subject), they are known to produce blurry output images (Stewart, 2019). This stems from the simplified approximation of the input image that VAEs capture in the latent space, which prevents the complex distribution of the input image from being reconstructed, resulting in a blurry output. GANs do not suffer from the problem because the generator is forced to generate sample images that look real.

Both VAEs and GANs are the subject of extensive research and lead the way in generative modelling solutions. But since their inception in 2014 GANs have made extraordinary advancements, progressing from synthesizing MNIST digits and low-resolution greyscale faces to high-resolution, photo-realistic images of diverse subjects (Jahanian,

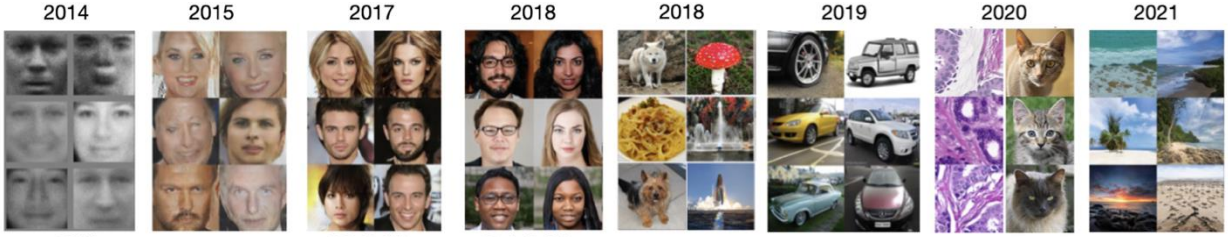

Figure 5 GAN Quality Improvement

2019) (Figure 5). Moreover, GANs are not confined to image generation; they can be used to generate sequential data such as audio and music or synthesize other time-series data like electrocardiograms and stock market trends.

Nevertheless, GANs excel at image synthesis, super-resolution, text and image-to-image conversion, inpainting, attribute manipulation, computer graphics rendering and texture generation – all of which are very relevant to the synthetic environment domain.

Application to Synthetic Environments

There is a very wide spectrum of synthetic environments requirements and solutions dictated by specific use cases and domains. This in turn, directly influences the fidelity, content, accuracy and authenticity of the synthetic environment itself, and the underlying data from which it is constructed. In general, high-quality data describing the real, physical world has never been more available and accessible. But despite this, there remain many challenges in using this data – it may not be available in the right timeframes, at the right cost, fidelity, quality, or coverage. Real data may contain unsuitable artefacts – clouds, shadows, vehicles on roads, seasonal appearance. Data usage rights may be prohibitive. In some cases, there may be no real data at all. Some applications may need the flexibility to represent changes or variations to the environment that cannot be easily introduced into real data. Whilst others may require the synthetic environment to represent an entirely artificial world that does not exist. In all these cases, the capability to generate plausible, synthetic data alternatives can offer significant benefits.

Moreover, generating synthetic data for SE can take advantage of the widespread use of data augmentation in machine learning. The training of deep learning models is reliant on large training datasets of suitable, and often labelled data samples, but building these datasets from real world data can be difficult and costly. Data augmentation aims to address this problem and increase the size of the dataset by changing a property of existing data, such as flipping, rotating, or randomly changing image hue, or by generating entirely new synthetic data. In many cases, GANs are being used to generate synthetic training data for domains that have a direct cross-over into SE. For example, remote sensing deep learning training requires GAN generated synthetic images with high quality representation of features and details that also take account of the spatial and spectral resolution, geometric and radiometric characteristics of the synthetic images. The use of GANs from these domains offers the opportunity to leverage their generative capabilities for the creation of similar synthetic data for use in synthetic environments.

Synthetic environments can incorporate a wide variety of data types depending on their use case, but most are composed of some combination of geospatial data layers (digital elevation models, multi-spectral imagery, point clouds, semantic data, cultural features, digital maps, topological data). The following sections will examine the application of DGMs to the generation of synthetic versions of these geospatial data layers.

Multi-Spectral Imagery

Multi-spectral imagery is typically composed of a tiled mosaic of aerial or satellite imagery of a specific spatial resolution and multi-spectral wavebands. GANs can be trained to generate images of any subject, including realistic satellite or aerial images. However, a standard GAN architecture (Figure 4) will produce images that may look realistic but whose detail, content, and arrangement of features - buildings, roads, trees, fields - is randomly selected and varied by the GAN. This is of limited value if images need to represent specific locations - applying GANs to real problems requires control over the output.

The required level of control can be provided by using a Conditional GAN that extends the standard GAN architecture by conditionally generating the GAN output. The Pix2Pix conditional GAN (Isola et al, 2016) is designed for general purpose image-to-image translation. This converts an image from one domain to a corresponding image in another domain – for example, greyscale to colour images, day to night images - and can be used to generate synthetic aerial imagery that is conditioned on map data. In this case a google map tile is provided as input to the generator which then attempts to generate a realistic image

Figure 6 Map-to-Image / Image-to-Map

tile with corresponding detail (roads, buildings etc). The discriminator is provided with the fake image output of the generator along with the corresponding real google map tile, as well as paired images of the real image tile and the matching Google tile. This process can also be applied in reverse to generate a synthetic map image from an aerial image input. Figure 6 illustrates the output of the map to synthetic aerial image process (first two columns), as well as the results from the image to map process.

The Pix2Pix generated 512x512 synthetic aerial images possess very plausible colorization, and have reproduced areas of vegetation as well as representatively capturing building and road infrastructure, although their boundary edges can be indistinct. The synthetic maps appear to have omitted areas of parkland and vegetation, building and road infrastructure is well represented but roads tend to have uneven edges. In perceptual realism trials, the synthetic aerial images fared better (18.6%) at fooling participants than synthetic map images (6.1%). This may be because the minor structural errors present in the maps are more visible in the context of their clean geometry, compared to the more cluttered content of aerial photographs.

One of the drawbacks of the image-to-image translation employed by pix2pix is the requirement to create a training data set of corresponding pairs of aerial and map images. Cycle-Consistent Adversarial Network (CycleGAN) architectures overcome this problem by performing image-to-image translation using unpaired examples from the source and target domain. A CycleGAN architecture was chosen to demonstrate the potential dangers of falsified satellite imagery in a recent study (Zhao et al, 2021) warning of the emergence and proliferation of deep fakes in geography and the need for timely detections of deep fakes in geospatial data.

Figure 7 Image Style Transfer

This study demonstrated the use of style transfer to generate synthetic satellite images of cities that have their infrastructure (roads, buildings, vegetation, etc) derived from a basemap of one city, Tacoma, (Figure 7) and their visual style transferred from the imagery of another (Seattle or Beijing). The generated images capture the geospatial and style differences in the low-rise buildings and greener space of Seattle, compared to the high-rise compact buildings with large shadow areas in Beijing. This study concluded that the authenticity of the GAN generated simulated satellite imagery was good enough to make it very difficult to distinguish with the human eye. However, it was possible to identify fakes by applying analytical methods to the generated images.

This CycleGAN solution illustrates that GANs can create images based on a "predicted" ground truth that appears real but is, in fact, a modified version of reality. GANs have the architectural flexibility to incorporate various approaches to conditioning the output to create synthetic data results and offer the potential to generate synthesized variations of the real ground truth to satisfy different scenarios and applications. The SSSGan (Marín, Escalera, 2021) architecture utilizes OpenStreetMap (OSM) classes that semantically describe features present in satellite imagery as a way to help enrich synthetic generation with finer details and properties. The model incorporates layers that focus on the high-frequency spatial details important in aerial images and learns texture and colour style from different geographic regions. Figure 8 illustrates how the real ground truth can be modified by incrementing or diminishing the presence of different semantic classes (grass or forest) to adjust the content and appearance of the generated image.

Figure 8 Semantic Class Variation

The remote sensing field needs to generate realistic synthetic multi-spectral imagery as training data for deep learning image analysis in tasks such as change detection, as well as to create synthetic modified ground truth images to support impact analysis of events such as floods. Remotely sensed data poses synthetic data challenges because of its much greater dynamic range compared to regular photographs (Baier et al, 2020) – Digital Elevation Models (DEMs) range from below sea level to Mount Everest; a Synthetic Aperture Radar (SAR) backscatter coefficient can vary considerably over a few metres from buildings to roads - all of which must be accounted for by normalizing the output datasets.

Figure 9 Synthetic RGB & SAR

Careful selection of data types to control the generation process will increase the quality of the synthesised images. Figure 9 (Baier et al, 2020) illustrates the results of a GAN-based image synthesis method that merges semantic information from land cover maps and auxiliary DEM raster data to generate synthetic RGB or SAR images. For optical and SAR images, there is good correspondence between real and synthesized images with realistic shadows that are consistent for the entire image and match building heights. Vegetation colour is realistic but differs from the ground truth because of the seasonal variation of the training data set. Artificial structures like buildings, bridges or roads are more easily identified as synthetic than natural forest, grassland or water.

Super Resolution

Utilizing imagery of the appropriate resolution is essential to the effectiveness of synthetic environment solutions. Higher resolution imagery makes smaller objects detectable and distinguishable from other objects and provides more accurate representations of shapes and areas in remotely sensed imagery. However, accessing high resolution data is not always feasible. Super-resolution offers a potential source of higher resolution data - it is another form of image-to-image translation that increases the resolution of a given image as well as sharpening its content by predicting the high-frequency component and the missing information. ESRGAN increases the resolution of input images by 4x (Nyberg, 2021). Figure 11 presents two samples of results showing the low-resolution image, synthetic high- resolution result, and the ground truth image (Nyberg, 2021). The super resolution quality is good with only minor artifacts where small features occasionally blend into each other due to lack of information in the low-resolution image. These same super-resolution techniques can be applied to enhancing the resolution of other critical raster data sets such as DEMs. High-resolution DEM data can be difficult to obtain. Despite better measuring equipment such as synthetic aperture radar, which has become the primary source of DEMs at a global scale, the limitation of equipment precision can still result in systematic errors that reduce the resolutions of DEM products. Applying super-resolution techniques to recover high-resolution DEMs from easily obtained low-resolution DEMs using SRGAN (Zhang, Yu, 2022) has proven effective in terms of robustness and preservation of accuracy and features on the generated high-resolution DEMs when applying a 4x resolution increase (Zhang, Yu, 2022).

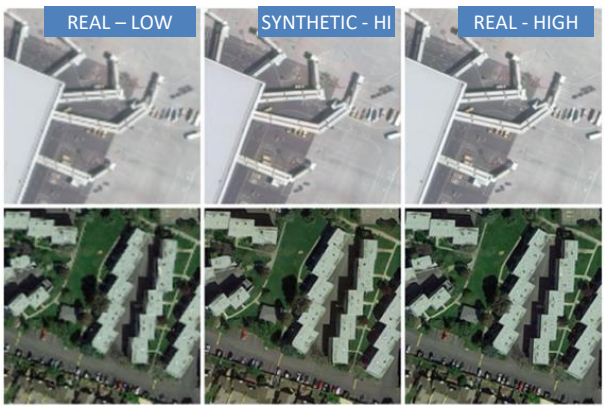

Figure 10 Super Resolution

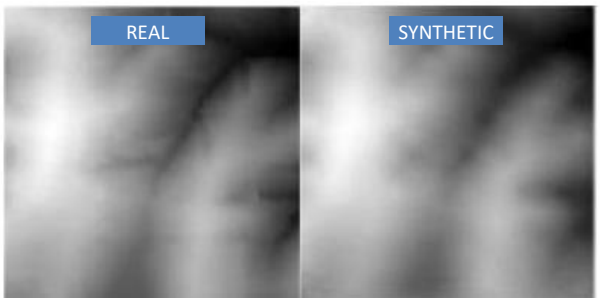

Figure 11 Super Resolution DEM

Cultural Features

Cultural feature data representing natural and man-made features is an integral data layer for synthetic environment applications. Some land-use datasets, urban form and street networks have an almost global-coverage but at low resolution. Others, such as Openstreetmap are widely used but have inconsistent coverage – road data is very complete, but even in areas of fully mapped street networks, building data is missing or incomplete for large areas of the world. GANmapper (Wu, Biljecki, 2022) addresses this sparse building data issue with a geographical content translation of street network data into synthetic building footprint data that is visually and morphologically similar to the ground truth. Using an image-to-image conditional GAN and coloured road hierarchy diagrams based on OSM road networks, the model learns appropriate building sizes and shapes to generate stitched tiled images of building footprint details and density represented at four different zoom levels. The model currently has problems generating plausible data in areas of very spare roads, large variations in building size, or similar street networks with different building typologies. Nevertheless, the model can generate realistic and morphologically correct urban patterns in previously unseen city areas or another city with a similar urban form. It can also transfer its learned patterns to areas of street networks with missing footprints or supplement existing footprint data. Notably, the GAN does not see the ground truth during training; therefore, synthetic building footprint locations may not exactly match the ground truth location.

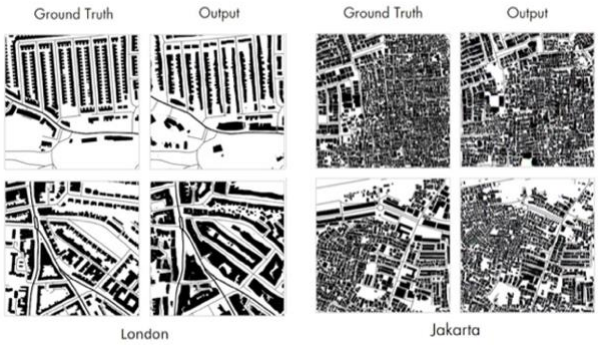

Figure 12 Synthetic Building Footprints

P**oint Cloud**

Point clouds represent spatial data as a collection of x,y,z coordinates, and are captured by LiDAR scans or photogrammetry to describe very fine, millimetre detail, or objects as large as buildings, cities or areas of terrain elevation. Synthetic point clouds are used for data augmentation, shape completion, enhancing low-resolution data and 3D reconstructions of objects or terrain. However, point clouds captured by LiDAR or depth cameras can be low-resolution and usually non-uniform, sparse and noisy, negatively affecting their use for tasks such as 3D reconstruction. PUFA-GAN (Liu et al, 2022) aims to address these problems with a GAN based point cloud super-resolution approach that generates a high-resolution point cloud from a real low-resolution point cloud input. The model generates an up sampled, noise-free, dense, uniform, high-resolution point cloud with rich geometric details consistent with the geometry distribution of the corresponding low resolution point cloud.

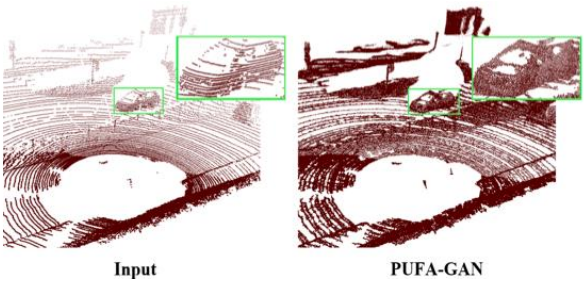

Figure 13 Synthetic Point Cloud

Synthetic Data Assessment

These geospatial synthetic data generation examples reveal that it is possible to create plausible, suitable synthetic data using generative models with sufficient control of the generation process to achieve the required output. Furthermore, this control can provide the capability to generate a synthetic version of real-world data that accurately represents a specific ground truth or to produce modified versions of the real ground truth.

Evaluation of generated synthetic data currently lacks metrics and objective methods and relies heavily on a qualitative assessment. GANs can capture realistic representations of both man-made infrastructure and natural features in synthetic images and display good correspondence between real and fake images. Man-made infrastructure such as roads and buildings are less precisely represented than the corresponding features in a real aerial image, with edges and boundaries that are not as geometrically sharp and distinct. This may indicate that buildings are the most difficult class to synthesize for a generator. Natural features such as forest, water or grassland are less easily identifiable as

synthetic. The quality, content and accuracy of the synthetic data is closely related to the DGM architecture, the control methods applied to the generation process and the features of the data being synthesised. For the moment at least, these synthetic geospatial outputs are very realistic, plausible representations of real data but are not an exact reproduction of the real ground truth. For example, although synthetic images capture natural and man-made features, buildings are not a perfect geometric or stylistic representation but will be located correctly if ground truth data like maps have conditioned the output. On the other hand, where the conditioning ground truth data is sparse and the goal is to synthetically add detail, as in building footprint synthesis, the morphology of the synthetic data will be realistic, but location will be less precise because it is being inferred by the generator. All the synthesized data examples contain small artefacts and ambiguities, but as found in the generation of synthetic remote sensing data, careful selection of the right combination of network architecture and conditioning data will not only minimise these issues but also increase the quality of the synthesized output. The comparison of different network architectures' performance on DEM super-resolution tasks concluded that network structure and loss function are also the most concerning points of the methods based on neural networks. Therefore, specifically designing the network structure and appropriate loss function for spatial data is necessary to integrate geospatial tasks and artificial intelligence (Zhang, Yu, 2022).

A variety of generative network architectures have been employed in these examples, which reflects the high level of architectural research and evolution in this field. This is progressively improving the synthetic data output in key areas such as image resolution, an important consideration for synthetic aerial imagery data. The architectural improvements in models such as Pix2pixHD, which incorporates a novel adversarial loss, as well as new multi-scale generator and discriminator architectures enable it to generate high-resolution 2k images (compared to the 512x512 output of pix2pix).

The current quality and fidelity of synthetic data will not be suitable for all use cases, but overall, these examples demonstrate the potential to generate and incorporate this data into synthetic environments under the right conditions. Generative capabilities and synthetic data requirements will need to be considered in these choices, along with the numerous potential benefits that synthetic data can offer – a low-cost alternative to real data, the ability to generate as much data as required on demand, synthetic versions of the ground truth, modified ground truth or a hypothetical ground truth, controllable synthetic data variations, augmenting and mixing of real with synthetic data. Synthetic data can now offer greater data flexibility and freedom and new opportunities to use and apply synthetic environments.

GENERATIVE MODELLING CHALLENGES

Deep generative models have made significant and impressive progress, but their successful development and application has several challenges. This paper has already described some of the difficulties with VAEs and their tendency to generate blurry images.

GANs are characteristically problematic to train. Simultaneously training the generator and discriminator models in a zero-sum game means that improvements to one model are at the expense of the other, making GAN training inherently unstable. Training aims to find a point of equilibrium between the two competing models, but if a balance is not reached, generator training can fail due to mode collapse, vanishing gradients, or non-convergence. Existing approaches seek to alleviate these issues by designing efficient model and network architectures, introducing suitable objective loss functions, or selecting alternative optimization algorithms. These have led to the development of many diverse GAN variants, which have improved but not yet fully resolved these training issues (Saxena, Cao, 2020). Model training also requires large training datasets that can be difficult and costly to develop. Generative modelling can help alleviate this problem by creating synthetic data to augment the training datasets. It can also help mitigate the bias found to exist in most training datasets that can inhibit the successful generalization of GANs (Jahanian, 2019). Furthermore, methods to meaningfully measure and evaluate the quality of GAN output are lacking, as are robust or consistent metrics to evaluate their performance (Saxena D, Cao J, 2020). This hampers comparison and selection of appropriate GAN architectures leading to a reliance on a qualitative assessment of outputs.

Finally, there is a need to capitalise on the remarkable advancement in generative modelling with tools that transfer this capability from research to users so that they can effectively exploit synthetic data opportunities.

GENERATIVE MODELLING OPPORTUNITIES

There has been a significant evolution and improvement in deep generative modelling capabilities in a relatively short time, which is evident in the progression of GAN performance and output quality since their inception in 2014. But research continues to open new opportunities to exploit the inherent characteristics of neural networks in manipulating and creating synthetic data.

Understanding and exploiting the latent space is a key area of research. Synthetic data created by a generative model can be considered augmented with extra functionality. This "generative data" (Isola, 2021) is composed of the synthetic data and the generative process that created it. Generative data behaves differently from regular data because its latent representation can be manipulated by performing operations, such as interpolation, to edit and change the generated data. The latent space arranges similar data features into groupings, and new data variations can be created by interpolating across the latent space between these features. For example, aerial imagery latent space may contain latent representations of urban and rural features. As illustrated in Figure 14, interpolating and sampling points in latent space from rural to urban will produce images with progressively more urban content (Jean et al 2019). Latent space editing and manipulation are unique to generative data. It opens a new dimension to generating synthetic data variation that is not possible with regular data. And while some latent space operations may appear superficially similar to other existing processes, such as standard image editing, in practice, the methods are very different. One potential application of these generative data properties is to explore counterfactual alternatives, or what ifs. For example, to generate multiple alternative versions of an aerial image to reflect and explore different conditions or scenarios. The degree of variation in the output data that is achievable by manipulating the latent space will be governed by the features captured in the latent space. But importantly, these variations are being created by manipulating the latent space alone.

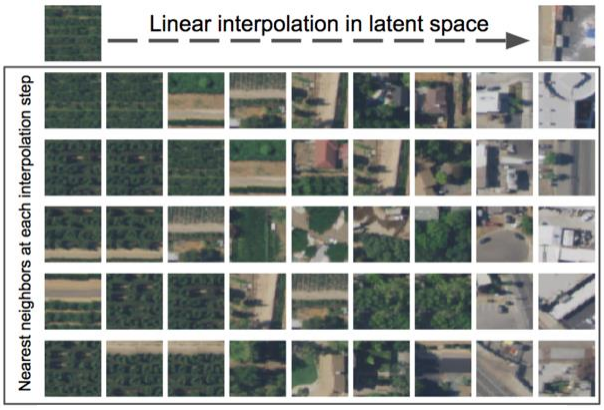

Figure 14 Latent Space Interpolation

This generative data concept is very compelling and underlines the unique capabilities of generative models. These are now being extended to the new and emerging field of neural rendering which aims to fuse classic computer graphics with generative machine learning to synthesise photo-realistic images and video. Classical computer graphics generate synthetic images of a scene using rasterization or ray tracing rendering algorithms which use scene representations composed of 3D representations of features built from triangle meshes, materials, textures or point clouds (SIGGRAPH, 2021). In contrast neural rendering works in reverse - it aims to learn a 3D scene representation from 2D images and render this as a synthesised image. This requires adding scene control by augmenting networks such as GANs with classical computer graphics rendering properties – controllable illumination, camera, pose, geometry, appearance, and semantic structure. The 3D scene representation can be learnt by several methods - Neural Radiance Fields (NeRF), multiplane images, voxel grids or scene representation networks – and then rendered using differential rendering schemes from computer graphics (Tewari et al, 2021).

Neural rendering brings a new perspective to how synthetic and virtual environments are generated and rendered, and the data and tools required to construct them. In theory, the traditional synthetic environment generation pipeline is replaced with a neural network that learns the 3D scene representation from 2D images and encapsulates it as a neural representation. In practice, Neural rendering is immature and far from providing a fully-formed replacement for existing graphics pipelines. It lacks well-developed methods and tools for controlling and modifying the learned scene parameterization (Tewari et al (2021). It is also currently focused on single objects and simple composite scenes and does not operate at real-time frame rates. Nonetheless, neural rendering solutions are being researched and implemented into real systems. Currently, these focus on rendering rather than the entire neural-based end-to-end synthetic environment generation and rendering pipeline.

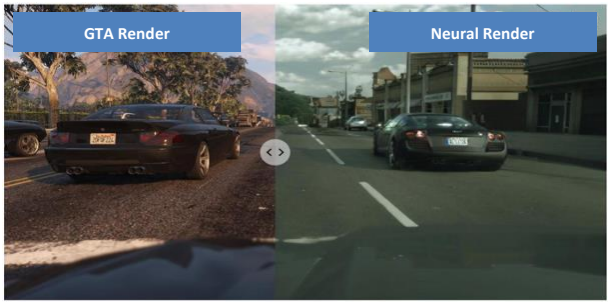

Figure 15 GTA Neural Rendering

Intel (Richter, AlHaija, Koltun, 2021) has recently demonstrated the integration of an adversarial learning-based approach with a conventional real-time rendering pipeline to enhance the photorealism of a rendered scene significantly (Figure 15). The system takes the rendered image and G-buffers from the Grand Theft Auto V graphics pipeline as input. It automatically enhances the images through an image synthesis algorithm, using real-world imagery from the Mapillary dataset and swapping out the less realistic lighting and texturing of the GTA game engine. This approach has many possible implications for the existing synthetic environment generation and rendering pipeline. First, it creates the possibility of transferring computationally expensive texturing and lighting from the game engine to the neural renderer, leaving the game engine to only generate base geometry and physics simulations. Secondly, this approach may require a simpler representation of a 3D environment to support the neural rendering of the synthesized output. Furthermore, it is possible to change the appearance of an existing 3D environment at the rendering stage without modifying the environment itself. The GTA example can be rendered with three different target datasets and reproduce the characteristic appearance of these datasets while keeping the structure of the original GTA images.

Others, such as Tesla (Tesla, 2021) have begun to implement a neural rendering solution incorporating a neural representation and neural rendering of the environment. In general, Tesla and other autonomous vehicle solutions require their computer vision systems to understand the 3D environment. Tesla discovered that using an image-based approach to train and execute their deep learning computer vision systems did not capture the full complexity of the real 3D world. Instead, Tesla has implemented a neural representation of the world, enabling them to capture a much more sophisticated 3D representation in neural vector space. This vector space version provides a much richer representation of the environment, better scene understanding, and ultimately better autonomous driving operations. The neural renderer generates a high-quality, photo-realistic rendered view of this vector space to provide a rich synthetic environment to the autopilot simulation in support of computer vision development and high fidelity simulation of the on-board sensors.

CONCLUSION

This paper has demonstrated that Generative modelling, particularly GANs, can create plausible, synthetic content and, by selecting the appropriate network architecture and conditioning data, can provide the necessary output control to produce realistic representations of a real or modified ground truth. The design and training of generative models are not trivial; nonetheless, their performance and the quality of synthetic data have improved through rapid architectural progress and their increasingly widespread application to data augmentation. Furthermore, the unique nature of generative data makes it possible to manipulate its inherent properties to create new data in ways that are not possible with regular data. However, synthetic environments have yet to embrace the data understanding and creative capabilities of generative methods or recognize their potential to transform existing approaches to data, synthetic environment development, and representation. In the short term, implementing GAN architectures and techniques like image-to-image translation offer practical data generation solutions. Leveraging this capability requires developing and training models to generate suitable output and creating the tools and methods that will enable data and content developers to exploit generative data. In the longer term, emerging concepts such as latent space manipulation, neural scene representation, and neural rendering offer the prospect of entirely new approaches to content generation, scene representation, and graphics rendering. However, these novel techniques are immature. To become viable alternatives to existing methods, they require further research and development to expand their capacity to support large, complex, synthetic environments and neural representations. Nevertheless, hybrid versions of these techniques, such as combining conventional graphic pipelines with neural rendering, provide practical first steps toward more comprehensive implementations. The examples presented in this paper underline the tremendous potential of generative modelling. It has rapidly evolved and will only continue to improve and broaden its capabilities. Generative techniques are widely regarded as the future of AI and have the potential to bring similar significant change to the future of synthetic environments.

REFERENCES

Baier G, Deschemps A, Schmitt M, Yokoya N, (2020). Building a Parallel Universe Image Synthesis from Land Cover Maps and Auxiliary Raster Data. Retrieved From: https://doi.org/10.48550/arXiv.2011.11314

Brownlee J, (2019). A Gentle Introduction to Generative Adversarial Networks (GANs). Retrieved From: https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y, (2014). Generative Adversarial Networks. Retrieved From: https://doi.org/10.48550/arXiv.1406.2661

Goodfellow I, (2017). NIPS 2016 Tutorial: Generative Adversarial Networks Retrieved From: https://doi.org/10.48550/arXiv.1701.00160

Isola P, Zhu J, Zhou T, Efros A, (2016). Image-to-Image Translation with Conditional Adversarial Networks. Retrieved From: https://doi.org/10.48550/arXiv.1611.07004

Isola P, (2021). Data++: Exploring Model-Based Data for AI. Retrieved From: https://youtu.be/YRbFDThaAmo

Jahanian A, Chai L, Isola P, (2019). On the "steerability" of generative adversarial networks. Retrieved From: https://doi.org/10.48550/arXiv.1907.07171

Jean N, Wang S, Samar A, Azzari G, Lobell D, Ermon S, (2019). Tile2Vec: Unsupervised Representation Learning for Spatially Distributed Data. Retrieved From: https://doi.org/10.48550/arXiv.1805.02855

Li Z, Li L, Ma Z, Zhang P, Chen J, Zhu J, (2022). READ: Large-Scale Neural Scene Rendering for Autonomous Driving. Retrieved From: https://doi.org/10.48550/arXiv.2205.05509

Liu H, Yuan H, Hou J, Hamzaoui R, Gao W (2022). PUFA-GAN: A Frequency-Aware Generative Adversarial Network for 3D Point Cloud Upsampling Retrieved From: https://doi.org/10.48550/arXiv.2203.00914

Marín J, Escalera S, (2021). SSSGAN: Satellite Style and Structure Generative Adversarial Networks Retrieved From: http://doi.org/10.3390/rs13193984

Mirsky Y, Lee W, (2020). The Creation and Detection of Deepfakes: A Survey Retrieved From: https://doi.org/10.48550/arXiv.2004.11138

Nightingale S J, Farid H (2021). AI-synthesized faces are indistinguishable from real faces and more trustworthy Retrieved From: https://www.pnas.org/doi/pdf/10.1073/pnas.2120481119

Nguyen T T, Nguyen Q V H, Nguyen C M, Nguyen D, Nguyen D T (2021) Deep Learning for Deepfakes Creation and Detection: A Survey. Retrieved From: https://doi.org/10.48550/arXiv.1909.11573

Nyberg, D. (2021). Exploring the Capabilities of Generative Adversarial Networks in Remote Sensing Applications Retrieved From: http://liu.diva-portal.org/smash/get/diva2:1573013/ATTACHMENT01.pdf

Richter S, Alhaija H, Koltun V, (2021). Enhancing photorealism enhancement. Retrieved From: https://doi.org/10.48550/arXiv.2105.04619

Ruthotto L, Haber E, (2021). An Introduction to Deep Generative Modeling. Retrieved From: https://doi.org/10.48550/arXiv.2103.05180

Saxena D, Cao J (2020). (Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions Retrieved From: https://doi.org/10.48550/arXiv.2005.00065

Soleimany A (2021). Deep Generative Modelling. Retrieved From: http://introtodeeplearning.com

Stewart M (2019). GANs vs. Autoencoders: Comparison of Deep Generative Models Retrieved from: https://towardsdatascience.com/gans-vs-autoencoders-comparison-of-deep-generative-models-985cf15936ea

SIGGRAPH (2021). Advances in Neural Rendering. Retrieved From: https://youtu.be/otly9jcZ0Jg

Tesla AI Day – The Presentation. (2021). Retrieved From: https://youtu.be/j0z4FweCy4M

Tewari A, Thies J, Mildenhall B, Srinivasan P, Tretschk E, Yifan W, Lassner C, Sitzmann V, Martin-Brualla R, Lombardi S, Simon T, Theobalt C, Nießner M, Barron J, Wetzstein G, Zollhöfer M, Golyanik V (2022). Advances in Neural Rendering. Retrieved From: https://doi.org/10.48550/arXiv.2111.05849

Tucker P (2019). The Newest AI-Enabled Weapon: 'Deep-Faking' Photos of the Earth. Retrieved From: https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/

Wang T, Liu M, Zhu J, Tao A, Kautz J, Catanzaro B (2018). High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. Retrieved From: https://doi.org/10.48550/arXiv.1711.11585

Wu A N, Biljecki F (2022). GANmapper: geographical data translation Retrieved From: https://doi.org/10.48550/arXiv.2108.04232

Zhang Y, Yu W, (2022). Comparison of DEM Super-Resolution Methods Based on Interpolation and Neural Networks. Retrieved From: https://doi.org/10.3390/s22030745

Zhao B, Zhang S, Xu C, Sun Y, Deng C (2021). Deep fake geography? When geospatial data encounter Artificial Intelligence. Retrieved From: https://doi.org/10.1080/15230406.2021.1910075