Diffusion Models vs. GANs – A Comparative Study of Each Model and Applications in Medical Imaging

Maxwell Stonham
Dept. of Electrical and Computer Engineering
of the University of Nevada, Las Vegas
Las Vegas, United States of America
stonham@unlv.nevada.edu

Abstract — **Generative Adversarial Networks (GANs) have been prominent within the past decade since its development in 2014 and have been mainly used in the generation of training models, realistic image, video, and audio generation, and general applications that involve the generation and construction of datasets. This deep generative model (DGM) was developed to replace other previously used DGMs, such as the Markov Chain Monte Carlo (MCMC)-based algorithms that suffer from slow, computationally intensive, inaccurate, and unclear datasets when sampling in higher dimensional spaces. GANs were developed to counter these problems to generate faster and more accurate datasets by incorporating two adversarial models, the generative and discriminative models, to compare whether the data generated was real or not. However, GANs still suffer from issues such as mode collapse, training instability, and non-convergence that limits its potential. Diffusion models are a much more recent generative model that has been increasingly popular, mainly known for being used in Stable Diffusion and the DALL-E image models used to generate hyper-realistic images from text. This model works by adding gaussian noise to its training data, and learning to recover it by reversing this process, or denoising. This method has also been increasingly popular within the medical field for its robustness and lack of problems that GANs suffer from. This paper will focus on the comparison between the two prominent generative machine learning models of the past decade, the GANs and Diffusion models, how they differ, the tradeoffs, advantages of the two, and applications, primarily within medical imaging.**

Keywords — Deep Generative Models, Diffusion Models, Gaussian Noise, Generative Adversarial Networks, Generative Machine Learning Models, Image Generation, Medical Imaging

I. Introduction

Deep generative models are a class of machine learning models and techniques that generate new data by learning the previously trained data's probability distribution [1]. Generally, DGMs train a generative model which mimics the distribution of the data that is being trained on and tries to generate a similar distribution to predict how the data is produced. Generative models are primarily used in image synthesis (inpainting, text-to-image applications), video synthesis (pixel prediction for blurred frames), audio and music synthesis, computer graphics (3D rendering, texture generation, simulations), and medical applications (image generation and segmentation).

Early on, most of these methods used Energy-Based Models (EBMs) which were introduced in the 1980s and used energy functions to model data distribution. EBMs utilized Markov Chain Monte Carlo (MCMC) algorithms, which combine Markov chains and Monte Carlo sampling to predict samples over a given target distribution by calculating the gradient of log-likelihood [2]. However, this training process is slow and iterative and does not work well in higher dimensions due to the increased number of parameters needed. These limitations inspired the discovery of techniques that were computationally more efficient, which is when Variational Autoencoders (VAEs) and Generational Adversarial Networks (GANs) were invented. Though VAEs are touched on as a brief comparison to add context towards the history and development of generative models, GANs and Diffusion Models will be the main focus of this paper.

II. Variational Autoencoders

Variational Autoencoders (VAEs) are an example of generative models that use variational Bayesian interference to approximate the probability density using an encoder and decoder [3]. This model was inspired by the Helmholtz machine, which is a probabilistic model of pattern recognition trained by the wake-up sleep algorithm that was proposed in 1995 [4]. VAEs utilize an encoder to map a given data through a multitude of different layers to reduce a data's dimensionality to a latent space, almost like data compression. Once "compressed", the decoder then attempts to reproduce the original data through variational interference to sample the original data from the latent space.

This type of model is considered an "explicit" model due to it being a probabilistic model, which are models that define parameters over a distribution over a random variable and specify a log-likelihood function [3]. This is noted since GANs are considered implicit models due to it being a stochastic model that generates synthetic data directly, or implicitly. Fig 1 shows the general architecture that makes up a VAE. The major drawback to explicit models would be its nature of approximating data through probability distribution, which can be problematic. VAEs can have weak approximations of the posterior distribution and the tendency to oversimplify this distribution which can lead to problems with the quality of the samples being reconstructed. These drawbacks are where implicit models are desired.

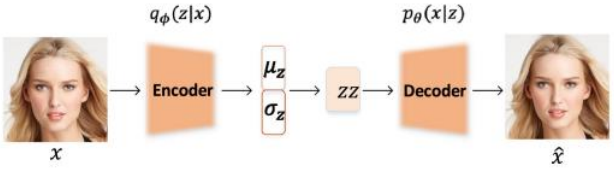


Fig. 1. General architecture of the VAE. The encoder adopts a recognition model to approximate the posterior distribution. The decoder adopts a generative model to map the latent variable z to the data x. The VAE trains its generative model to learn a distribution to be near the given data by maximizing the log-likelihood function. [5]

III. Generative Adversarial Networks

Generative Adversarial Networks (GANs) is a recent model that was introduced by Goodfellow et al [6] in 2014 that helped mitigate the issues that MCMC-based models had. GANs use two neural networks: a generative model (generator) and a discriminative model (discriminator), to train the overall model based on a minimax zero-sum game [7]. The Generator is trained to produce synthetic or fake data based off of the real data that it is trained on, while the Discriminator is trained to identify whether or not the data being produced is real or fake. The Generator continuously trains itself to produce data that is as real as possible, and the Discriminator trains itself to identify fake or real data until the Generator reaches a point where it produces data that is seemingly so real that the Discriminator cannot tell is fake. This iterative process eventually produces synthetic data to be as identical as the original data as possible. The convergence of this process (or lack thereof), however, can cause problems and instability since mode-collapse and non-converging gradients still occur [7], which will be discussed later in the paper.

Two types of GANs have been proposed, unconditional and conditional GANs. Unconditional GANs were the originally proposed GANs that utilize a multilayer perceptron to create a probabilistic model taking latent noise variables z and observed real data x as inputs. Studies have shown, however, that using convolutional neural networks (CNNs) has been shown to be more effective than MLPs for capturing image features [5]. The downside to unconditional GANs is that the user has no control on what to generate since the only input is the random noise vector z. Studies have shown that adding in a new conditional input y of additional information (image specifications, labels, text, etc.) can adjust the generated results to match a user's specification.

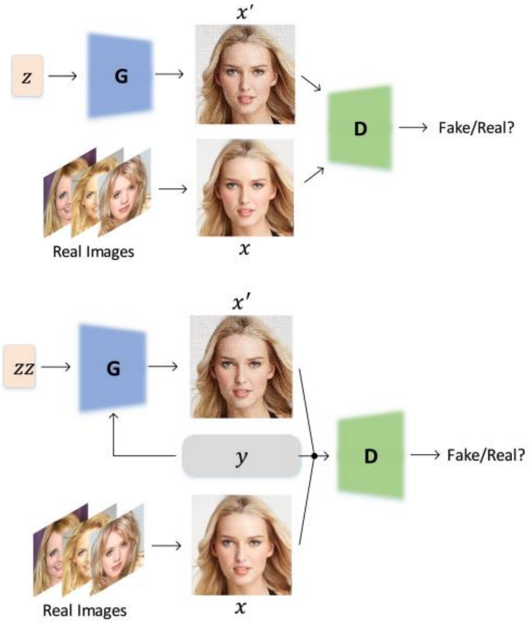


Fig. 2. Top: Unconditional GAN Structure. Bottom: Conditional GAN Structure. Z denotes the random noise, G represents the generator, and D as the discriminator. Y is the additional variable added in for new information to get more user-specified results. [5]

A. Issues

GANs suffer from training instability, which occurs when either generator or discriminator cannot reach a stable equilibrium and causes oscillations or non-convergence from simultaneous iterations during training and optimization [8]. This instability prevents the generator from generating samples that are similar to the images being fed into the model, which also causes a phenomenon called mode collapse.

Mode collapse (also known as the missing mode problem) is a prominent issue when it comes to training GANs. This occurs when the generator fails to capture important peaks present in the data distribution (known as modes) and struggles to produce a diverse range of outputs and starts to produce a very limited set of outputs that is repetitive and does not fully reflect the overall set of data that it was initially trained with. [9]. This results in the GANs having less variability than the training data that was sampled originally.

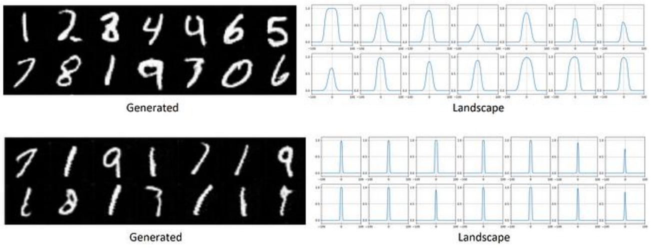


Fig. 3. Top: GAN Training with a MNIST dataset showing good convergence. Bottom: The same training done with discriminator overfitting, showing mode collapse [10].

Fig 3 shows GAN training for an MNIST dataset, which is a dataset of handwritten digits between 0-9, generally used for machine learning tasks such as classification and generative modeling (used in this case). The top graph shows that the function representing our desired output manages to reach a local maximum at k=0 for every real datapoint (good convergence). When a generated sample falls within the range of a real datapoint, gradient updates will move it closer towards the real datapoint, and these local maxima being present in different regions of the data space moves the generated samples in different directions to distribute them throughout the space and prevents mode collapse. The bottom graph, however, shows how discriminator overfitting can be problematic to training GANs and cause mode collapse (seen by the inaccuracy of the image being generated). Notice the graphs are much sharper here than they were previously, which is caused by discriminator overfitting on the real datapoints, which is problematic since now the scores of nearby datapoints approach zero. This causes flat regions which prevents the generated samples to move towards the real datapoints, which limits the space that the generated samples can effectively be distributed through, causing the diversity of generated samples to decrease which triggers mode collapse. The local maxima need to have a wide shape (as opposed to sharp edges) to allow the generated samples to move towards different directions to prevent mode collapse from occurring [10].

Methods to prevent these issues have been proposed in the past and is a current area of research. Possible methods to prevent mode collapse could be to use better metrics to improve training to reach better optima, or to use multiple generators to capture more modes within the data distribution [8]. However, alternative generative models have also been proposed as an alternative to having to solve the current problem of mode collapse by developing a completely new method of generating data, through diffusion models.

IV. Diffusion Models

Diffusion models are an even newer form of deep generative models that show to be even more promising than previous DGMs. They are a class of probabilistic generative models [11, 12] that is trained by adding and removing Gaussian Noise to a given data structure. This procedure is done through forward diffusion (by adding Gaussian Noise to a given data structure to destroy it) and then reversed through reverse diffusion (removing the noise added to reconstruct the data structure). The addition of noise is gradual per iteration, starting small, and adds more noise as the process goes on until the training data is pure Gaussian noise [11]. The same process is done when reverse diffusion occurs, removing noise gradually per iteration, until the data is reconstructed as similar as the original data was. Three main formulations of the diffusion models are discussed in this paper: Denoising Diffusion Probabilistic Models (DDPMs), Score-Based Generative Models (SGMs), and Stochastic Differential Equations (SDEs).

Diffusion models have recently become the new prominent generative model, overtaking GANs in terms of improving tasks and challenges involving generative applications since they don't suffer the problems that GANs are prone to having. As of the past few years, diffusion models have also shown to be extensively used in the medical field [13] and the later sections of the paper will also discuss the contributions made in this field.

Despite DDPMs being a having a more robust framework for image generation as well as reliably covering multi-modal data distributions, it comes at a high computational cost during training and interference in order to generate high-quality and diverse outputs [14]. Methods to improve diffusion models have been to enhance empirical performance and extend model capacity [12]. The three models discussed in this section use the same approach as previously mentioned, that is by iteratively adding noise to a given data structure and removing it through a similar process to try and generate synthetic data as close as it can to the real data it was trained on. Fig 4 shows a visual representation of a diffusion model.

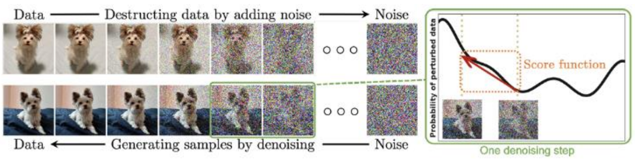

Data
Destructing data by adding noise
Noise
Score function
Probability of perturbed data
Data
Generating samples by denoising
Noise
One denoising step

Fig. 4. A visual representation of the process of training diffusion models. The sample image is fed into the model, and Gaussian noise is iteratively added until the image is deconstructed. The process is then reversed to attempt to replicate the same image. The graph to the right illustrates the denoising step in the reverse process, generally requiring estimating the noise function which is a gradient pointing to the directions of data with higher likelihood and less noise [12].

A. *Denoising Diffusion Probabilistic Models (DDPMs)*

Denoising Diffusion Probabilistic Models (DDPMs) are a class of latent variable models which were inspired by the non-equilibrium thermodynamics theory [11, 15]. Latent variable models (LVMs) are a type of statistical model that uses latent variables, which are hidden, unobserved variables, to define the structure of observed data, which also categorizes all diffusion models. DDPMs are considered probabilistic modelling since they use the distribution of the trained data to produce similar, identical versions of the real data [16]. During the training phase, DDPM uses a maximum likelihood estimation (MLE) to maximize how likely it is to produce clean images after noise has been added onto it during training [16].

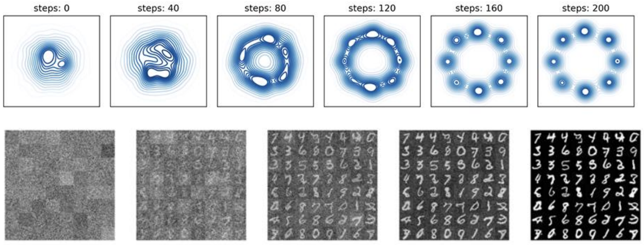


Fig. 5. Top: Reverse diffusion process for 8 Gaussian datasets. Bottom: Reverse diffusion process for an MNIST dataset. Note how the reverse process starts from noise and iteratively adds steps of denoising to generate samples that resemble the true distribution [17]

Fig 5 shows a visual example of the reverse diffusion process. How well the DDPM is trained, and its performance would also be heavily dependent on the initial random noise used for sampling. To capture all modes in the distribution, the noise should be sampled from a range that covers the boundaries of the distribution in the synthetic dataset. For image data, it is generally advised to normalize inputs to the range [-1,1] and sampling noise from a Gaussian distribution that covers this range is recommended. This helps the DDPM to effectively model the entire data distribution [17].

B. *Score-based Generative Models (SGMs)*

Score-based Generative Models (SGMs) differ from DDPMs since on top of adding and removing Gaussian noise to an image, the model also simultaneously estimates the score function for all noisy data distributions [12]. This estimation uses a model called Noise-Conditioned Score Networks (NCSN) [12], to estimate and sample a score based on the logarithmic data density (the gradient of the log-density function at the input data point) [18]. On top of score-based sampling, this model also uses the concepts of Langevin dynamics [15, 18] which in physics is a stochastic method used to predict the motion of particles in a system which are affected by the drag force of the given system. Similarly, in terms of machine learning, the gradient of our log density can be seen as the drag force which affects a random sample through the data space into regions with high data density [11]. The sampling method based off of this is called Annealed Langevin Dynamics (ALD), which is a process that starts with noting the scores of the highest noise levels, and gradually annealing (or reducing) the noise levels until it is small enough to be indistinguishable from the original data distribution [18].

C. Stochastic Differential Equations (SDEs)

Stochastic Differential Equations (SDEs) is a method of generalizing both DDPMs and SGMs since they are both continuous and are solutions to SDEs [11, 12]. The main difference between SDEs and DDPMs is that SDEs use a framework that is continuous, while DDPMs are discrete. This method offers high precision and flexibility but can be inefficient and impractical due to its high computational cost, which is why DDPMs are generally preferred.

V. Applications of Generative Models

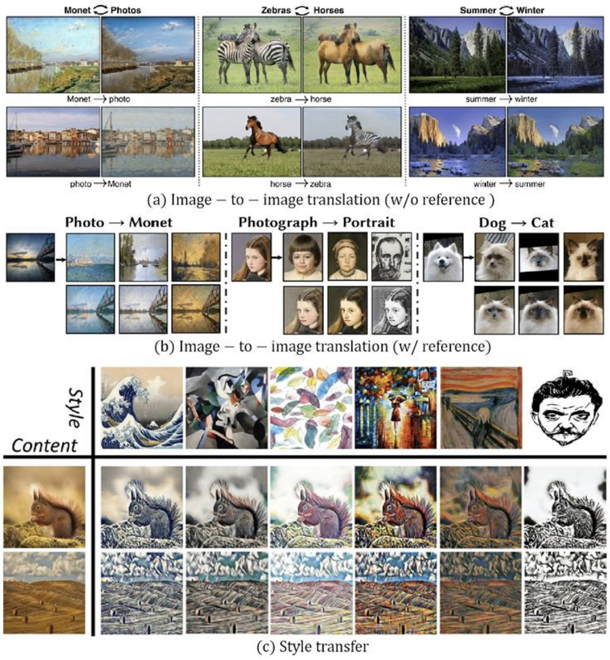
Generative models have been heavily researched on and optimized over the past few years to be able to generate high-quality synthetic images, human-like natural language, highly diverse human speech, and numerous other audio and video applications. They are popularly used primarily in imaging tasks, such as image-to-image translation, style/texture transfer, image super resolution, or image generation from text prompts. While numerous fields use generative models, such as in computer vision, entertainment, games, music production, natural language processing, healthcare, and more, this paper will focus on applications that relate to images as opposed to video or audio generation. This section will discuss some popular applications in entertainment that heavy research is being conducted in, while later sections will focus on its applications within the medical field.

As of 2021, GANs held the state-of-the-art on most image generation tasks as measured by sample quality metrics such as FID (Fréchet Inception Distance), Inception Score, and Precision [19], which are widely used metrics for evaluating the quality of generative models, such as image generation. Diffusion models then started becoming popular due to their similar ability to produce high-quality images while offering desirable properties such as distribution coverage, stationary training objective, and easy scalability. Diffusion models also held the state-of-the-art on CIFAR-10, but still lagged behind GANs for generating difficult datasets like LSUN and ImageNet [19]. However, heavy research is currently being done on diffusion models to see whether or not they will surpass GANs in terms of their image generation capabilities.

Image-to-image translation (I2I) is a popular application of generative models and refers to the conversion of different types of images and merging them into one to mesh the major structures and context of two or more images. This is done by mapping different image domains and generating a mix of the images that were analyzed within the model [5]. For example, an image of a dog might want to be converted to an image of a cat, so a user would then decide to merge two images together (one that is the subject image that outlines the dogs major characteristics such as face shape or color, and another that is an image of a cat) to try and make the initial image have the design and outlines of the second image.

Style/Texture Transfer is another popular use of generative models that is very similar to I2I but differs slightly. They can be compared by the following use case: I2I directly converts the input image from the source domain to the target domain, while style transfer generates a stylized image from the content image with the style image as reference [20]. An example would be to convert a realistic image of a person to a sketch-like image by feeding into the model two images, one of a real person, and another of a sketch of something. The model will then try and match the face of a person to make it look like it was sketched. This type of translation can even be utilized in the medical field by converting architectural drafts into lifelike images for different design and training applications [7]. Fig 6 shows a visual example of the comparison between I2I and style transfer applications.

Super resolution is another application that uses a model to enhance the resolution of an image. In recent years, super resolution has seen significant advancements in both GANs and DDPMs. Given a low resolution or blurry image, super resolution is a technique that increases the dimensions of an image, by upscaling the pixel density while preserving the characteristics and details of the original image to produce sharper details and tones when compared to the original image being fed into the model. Text-conditioned super resolution also differs from pixel-based natural image-focused super resolution by incorporating textual descriptions, such as a dedicated multi-modal large language model, to automatically caption input images with non-image super resolution specific captions at a noticeable performance boost [14]. With how developments have been made in training and scaling DDPMs, diffusion models are surpassing GANs for all these applications, especially due to the problems that GANs suffer from. The comparison between the two models in this application will be discussed in the next section of this paper.



(a) Image − to − image translation (w/o reference)

(b) Image − to − image translation (w/ reference)

(c) Style transfer

Fig. 6. (a) Shows I2I translation with reference. The left side of the image is the input image, and the right side is the generated image, given a text prompt. (b) Shows I2I translation without reference, but instead using a reference image. The first row is the reference image, the second row is the generated image. (c) Shows style transfer. The first column is the style that is referenced, the first column is the input image. Everything else is the generated image with the input and reference merged together [20].

Perhaps the most popular system in image synthesis and generation that has gained popularity within the past year or two would be the widely known Stable Diffusion (released in 2022 by Stability AI) and DALL-E 2 (released in 2022 by OpenAI) models. Both these models are relatively recent and are based on DDPMs to generate their images through text-to-image synthesis and have become popular to mainstream media due to its ease of access and extremely realistic outputs when generation images based off text prompts. As two highly successful generative models, Fig 7 shows a comparison between these models through the FID score when generating realistic human faces.

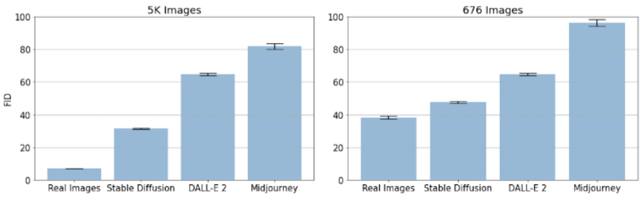


Fig. 7. FID score comparisons between real images, Stable Diffusion, DALL-E 2, and Midjourney (image generation model created by an independent research lab). Left: FID score of models over random sets of 5000 faces. Right: FID score of models over a random set of 676 faces. Results are averaged over 10 runs [21].

We can note from Fig 7 that these models produce realistic images, arguably comparable to real images. Due to these impressive results, ethical concerns have been addressed due to implications that might relate to discrimination, privacy, job displacement, and other unintended consequences [22]. Due to its recent inception, current development is still being made to implement effective regulation and government framework regarding the use of these models to ensure responsible AI practices are being practiced as we move towards the increasing growth and development of generative models.

VI. Model Comparisons

This section of the paper will primarily discuss the comparisons between GANs and DDPMs. General comparisons between image generation in entertainment will be discussed in this section, while a focus on biomedical applications is going to be discussed in the section after this. To begin are some figures for general comparisons between GANs and DDPMs. Note that Figs 7 and 8 showcase rather extreme comparisons between the two models since both these figures emphasize the problem that GANs suffer from, that is mode collapse.

The top half of Fig 8 shows us a comparison of 8 Gaussians shaped as a circle (shown as our original dataset to the right). We can note that the red samples are consistently generating fairly close representations, whereas the green samples are shown to have trouble generating certain modes, regardless of how many iterations there are. This is a clear drawback to the GANs as this is a visual example of mode collapse. We can clearly see here that the model struggles to capture all the data distribution and fails to produce a diverse range of dataset, limiting its output.

Similarly with the bottom half, the GAN struggles to produce accurate representations of the given dataset. This showcases the issues that GANs can experience but note that not all GANs have a tendency to experience mode collapse. Factors including poor weight initialization, improper learning rates, or insufficient training times are some factors that can affect the tendency for a GAN to experience this phenomenon.

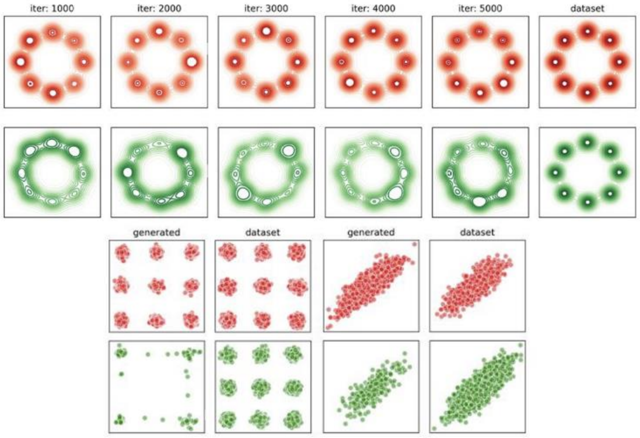


Fig. 8. Samples of DDPM (Red) and GAN (Green) [17]

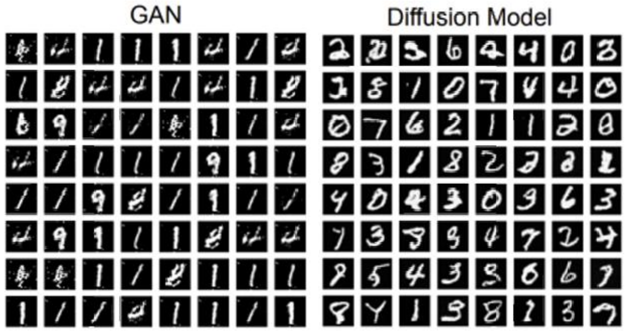


Fig. 9. MNIST Dataset Comparison between GANs and DDPMs [17]

Fig 9 shows another comparison between the models using the MNIST dataset. We can note that the repetitiveness due to mode collapse in GAN models is shown more clearly with the continuous generation of the numbers “1” and “9”, whereas the diffusion model is shown to produce a reasonably good representation of a variety of numbers within this range.

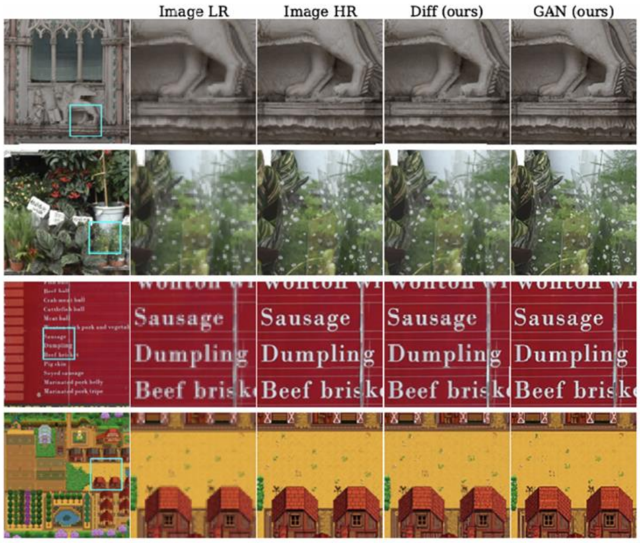
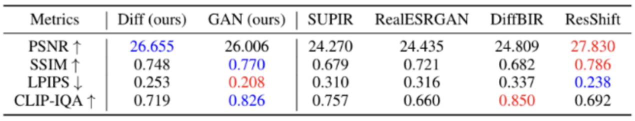


Fig. 10. Visual comparison between GAN and Diffusion Model in the application of super resolution [14].

The authors in Fig 10 [14] conducted a systematic comparison between GAN and Diffusion Model for image super resolution (ISR). The research was conducted as a fair comparison between the two models by matching the model size and training data available during their setup. [14] trained their models under a comparable setup using a large-scale dataset of 17 million text-image pairs consisting of 1024x1024 px images of exceptionally high-quality images and image-text relevance. They trained their models on image segments (shown in the figure as a zoomed in portion of their high-resolution image sample) to increase training efficiency and sample variety. Their goal was to produce low-resolution - high-resolution image pairs by extracting 256x256 px random crops of the original images which are then downscaled to 64x64 px. Both their GAN and Diffusion Model use the same architecture and are identical in terms of parameters. However, the slight difference is that their Diffusion Model has slightly more parameters due to the group of noisy inputs and image condition in the input layer and timestep embeddings [14].

Their setup also aimed to test the impact of adding text conditioning within their super resolution (SR) models. The results are shown in Fig 9. They observed that their GAN SR model converges very quickly. After several hundred iterations, their GAN models achieved equilibrium between the generator and discriminator and continued to improve steadily until convergence. The Diffusion Model, however, showed slower convergence and requiring up to 620,000 iterations to fully converge. Table 1. Shows a qualitative comparison between the models. The metrics used, PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index Measure), LPIPS (Learned Perceptual Image Patch Similarity), and CLIP-IQA (CLIP-Based Image Quality Assessment), are widely used in evaluating the quality of images.

Table 1. Quantitative comparison between GAN, Diffusion SR, and current state-of-the-art models on Image Super Resolution. Red shows the best result, Blue shows the second-best results.



Metrics	Diff (ours)	GAN (ours)	SUPIR	RealESRGAN	DiffBIR	ResShift
PSNR ↑	26.655	26.006	24.270	24.435	24.809	27.830
SSIM ↑	0.748	0.770	0.679	0.721	0.682	0.786
LPIPS ↓	0.253	0.208	0.310	0.316	0.337	0.238
CLIP-IQA ↑	0.719	0.826	0.757	0.660	0.850	0.692

The results of this research concluded that both GAN and Diffusion Models yielded approximately the same quality upsampled images with sharp edges, good textures, and small details, which is seen in Fig 9. Further experiments were conducted, but similar results were seen. [14] expected that mode collapse was going to be an issue for GAN-based models, but did not encounter any difficulties with optimization. They concluded that GANs can achieve the quality level of modern diffusion models if trained under the same protocol and under similar setup conditions, but with the advantage that it has faster training times and single-step inference as opposed to the computational-heavy, iterative denoising procedure.

VII. Generative Models in the Medical Field

The field of healthcare has been shown to have a growing interest in generative models such as diffusion models, primarily in medical imaging. The medical imaging community has seen exponential growth within a number of diffusion-based techniques [13, 23]. Modern biomedical image analysis using deep learning often encounters the challenges of limited annotated data, which is why deep generative models have been explored to synthesize realistic biomedical images [23]. A big portion of medical image analysis often includes image segmentation, also requiring a large amount of annotated training data which is time-consuming and costly [28]. This difficulty of data collection procedure, privacy concerns, lack of experts, and requirement for authorization from patients creates major delays within the annotation process [13]. This section will discuss the applications of generative models within the medical field (as well as their importance), some which have been touched on in previous sections, and some that will be introduced in this section. These applications include image reconstruction, segmentation, 2/3D generation, anomaly detection, I2I, denoising, and more.

In May 2023, a survey paper was conducted by [13] to gather papers published on deep generative models for medical imaging to showcase the increasing number of articles being published towards this area of research for use within the medical field. Fig 11 showcases this increase through graphical representation. We can note which application diffusion-based models are most commonly used within this field, as well as the imaging modalities, and popularity in the number of papers that were published within the past 3 years. From there being practically nothing in the third quarter of 2021, to there being over 40 published papers towards the first quarter of 2023, showcasing which application is used the most. Note that general image generation and segmentation are the most commonly used applications. This recent popularity is also most probably due to the fact that diffusion models have only recently been gaining traction in terms of popularity and efficiency within the past few years.

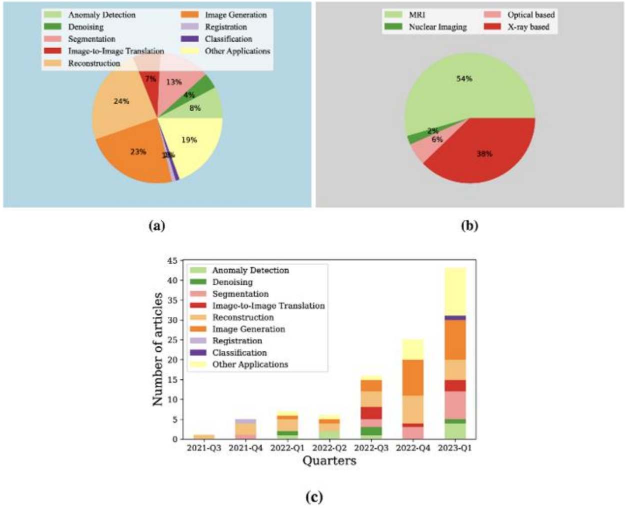


Fig. 11. The diagrams showcase the published papers within the medical field according to their (a) applications, (b) imaging modalities, and (c) number of papers published incremented by their year and quarter [13].

The current problem within the medical field is that many datasets suffer from severe class imbalance due to the rare nature of some pathologies causing the overall datasets in medical imaging to be small compared to natural image datasets that are more commonly accessible and easily used to train popular models (such as Stable Diffusion and DALL-

E2). To top this off, national and international patient privacy laws such as the Health Insurance Portability and Accountability Act (HIPAA) and European legislation pose barriers to data sharing across institutions [28], making it difficult to have a large pool of data from multiple institutions. Earlier studies showcase the use of generative models to create synthetic images to expand already existing medical datasets using GANs, but the problem lies with its inability to capture the full distribution of data causing mode collapse which limits the outputs diversity, as well as training instability, which was addressed earlier in this paper. Medical images are complex and high-dimensional data, so the diffusion model's reliability in generating these images make it very appealing. This is a reason why diffusion models have been dominating in popularity since they have the ability to not only generate high-resolution images without the risks of mode collapse, but also denoise images which is extremely useful in the field of medical imaging. The synthetic images generated also prevents data security concerns that arise when using patient data publicly, which is why diffusion models have such a promising outlook when utilized in a clinical setting to address imaging challenges.

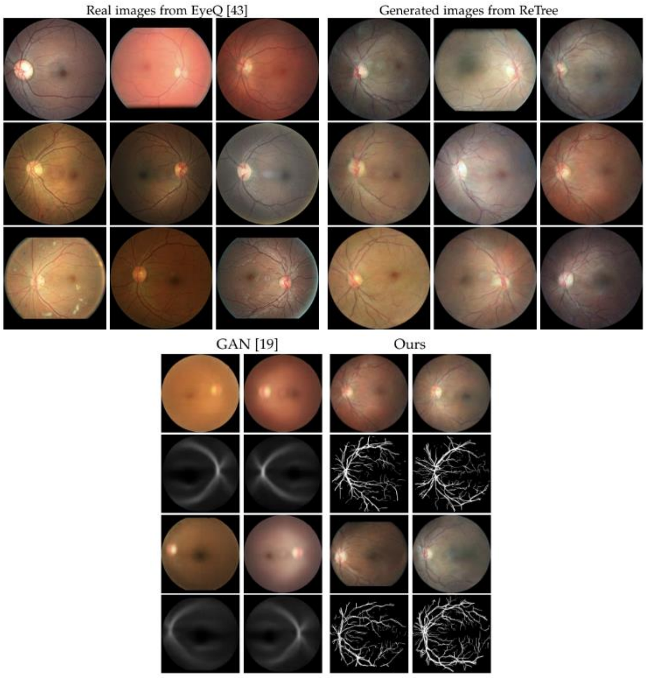


Fig. 12. Top: Comparison between the EyeQ dataset given and the images produced by the ReTree model. Bottom: A comparison between the authors DDPM model (right) to another author's GAN model (left) [24]

An example of DDPMs being used in the medical setting is through the use of retinal image generation and segmentation. Image segmentation simplifies the complexity of an image by decomposing into meaningful different segments [13]. [24] conducted research using GANs and DDPMs to generate retinal images by developing their own dataset (ReTree) containing retinal images, vessel trees, and a segmentation network based on DDPM trained with images from this dataset. The DDPM framework proposed used a lightweight architecture and training technique that is theorized to improve the performance of DDPMs. The motivation of this study was due to the difficulty encountered using previous learning-based models that resulted in false-positives and there being a limited difference between vessel tress and background. The authors also implemented GANs to generate retinal images, but it proved to be problematic when training since their models could not generate diverse data due to mode collapse, non-convergence, and experiencing vanishing gradient due to adversarial training. They proposed a DDPM that would mitigate these problems and generate more promising results.

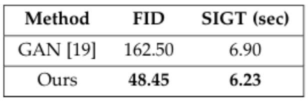
The top row of Fig 12 shows the generation of retinal images that the author of this paper uses based on DDPM. This comparison between the dataset and the generated images shows that DDPM's have promising potential when used for image generation within the medical field since the images are quite identical. The author then compares their DDPM with a previously used GAN model in the bottom of Fig 12. Notice the comparison between the two models is quite significant. Once again, the GAN suffers from mode collapse which prevents the model from converging, leading to low quality images when compared to the DDPM, which is evidently seen from the results by noticing the lack of color and overall structure.

Quantitatively, Table 2 shows a comparison between the two models through the FID and SIGT (Single Image Generation Time) in seconds. We can notice here that their DDPM outperformed the GAN model that was used, both in generation time and FID metric, which compares the distributions of features extracted from the real and generated images (lower is better). We can also note that Table 3 compares how well each model is able to be trained based on the three publicly available datasets. In all datasets, we can see that the DDPM (ReTree) once again outperforms the GAN in all the tests.

However, the drawback to the proposed DDPM framework is that it relies on random noise to generate samples and can possibly result in unrealistic images being generated. DDPMs simplify assumptions made about the probability distribution of the input data, which could involve assuming a Gaussian distribution or a certain level of smoothness. While these assumptions can help with the modeling process, they can also limit the model's ability to accurately capture complex data distributions which can lead to the generation of unrealistic images. Failed images showed two retinal cups, no retinal cups, color distortion, and overly bright regions.

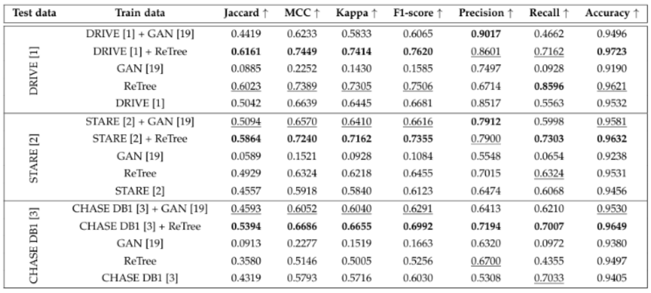
The author solved this problem by training a discriminative model using real images from a dataset to classify real and generated images. This helped filter out the unrealistic images that were generated and instead only keep the realistic ones within this dataset. Despite this limitation that they eventually solved, the overall consensus of this research was that the DDPM architecture proposed significantly improved overall performance in terms of efficiency, computational cost, quantitative and qualitative results, also showcasing its superiority when compared to GAN models. They succeeded in image generation, synthesis, segmentation, and super resolution of retinal vessels, showcasing a good model to produce an original dataset that bypasses the limitations of GANs and any ethical concerns.

Table 2. FID and SIGT Comparison between the two models [24].



Method	FID	SIGT (sec)
GAN [19]	162.50	6.90
Ours	**48.45**	**6.23**

Table 3. Quantitative comparison between 3 publicly available datasets (DRIVE, STARE, and CHASE DB1) used to train both GAN and DDPM. [24].

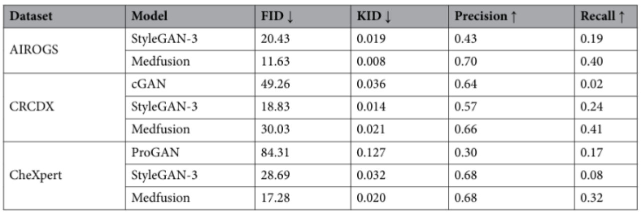


Test data	Train data	Jaccard ↑	MCC ↑	Kappa ↑	F1-score ↑	Precision ↑	Recall ↑	Accuracy ↑
DRIVE [1]	DRIVE [1] + GAN [19]	0.4419	0.6233	0.5833	0.6065	**0.9017**	0.4662	0.9496
	DRIVE [1] + ReTree	**0.6161**	**0.7449**	**0.7414**	**0.7620**	0.8601	0.7162	**0.9723**
	GAN [19]	0.0885	0.2252	0.1430	0.1585	0.7497	0.0928	0.9190
	ReTree	0.6023	0.7389	0.7305	0.7506	0.6714	**0.8596**	0.9621
	DRIVE [1]	0.5042	0.6639	0.6445	0.6681	0.8517	0.5563	0.9532
STARE [2]	STARE [2] + GAN [19]	0.5094	0.6570	0.6410	0.6616	**0.7912**	0.5998	0.9581
	STARE [2] + ReTree	**0.5864**	**0.7240**	**0.7162**	**0.7355**	0.7900	**0.7303**	**0.9632**
	GAN [19]	0.0589	0.1521	0.0928	0.1084	0.5548	0.0654	0.9238
	ReTree	0.4929	0.6324	0.6218	0.6455	0.7015	0.6324	0.9531
	STARE [2]	0.4557	0.5918	0.5840	0.6123	0.6474	0.6068	0.9456
CHASE DB1 [3]	CHASE DB1 [3] + GAN [19]	0.4593	0.6052	0.6040	0.6291	0.6413	0.6210	0.9530
	CHASE DB1 [3] + ReTree	**0.5394**	**0.6686**	**0.6655**	**0.6992**	**0.7194**	**0.7007**	**0.9649**
	GAN [19]	0.0913	0.2277	0.1519	0.1663	0.6320	0.0972	0.9380
	ReTree	0.3580	0.5146	0.5005	0.5256	0.6700	0.4355	0.9497
	CHASE DB1 [3]	0.4319	0.5793	0.5716	0.6030	0.5308	0.7033	0.9405

Another research conducted by [25] aims to introduce their own DDPM (Medfusion) to evaluate its performance against GANs. The paper also emphasizes the drawbacks of GANs with their limited diversity and risk of non-convergence. Their DDPM model is compared to StyleGAN-3 (as well as other GAN models). They propose a conditional latent DDPM for medical image generation that was trained from the CRDX dataset. Latent DDPM is a variation of conventional DDPM's, but instead of operating directly on high-dimensional pixel data, it compresses the data into a latent space first before proceeding with the diffusion process. This study focuses on three types of medical data: ophthalmologic data (fundoscopic images), radiological data (chest X-rays), and histological data (images of stained tissue). The Medfusion architecture proposed for this use case showed that the images generated from their DDPM was far better than the baseline GAN-generated images, proving to exceed the results through the metrics such as the FID, KID (Kernel Inception Distance), Precision, and Recall.

Through the results shown on Table 4, we can see how the Medfusion model outperforms the StyleGAN-3 model in every metric. Something to note is that the Recall results (which is a measure of diversity of generated images) are extremely low in the GAN models, which indicates that mode collapse occurred during training. This is seen when training the CRCDX dataset through the cGAN model (with a Recall score of 0.02) and the CheXpert dataset through the StyleGAN-3 model (with a Recall score of 0.08).

Table 4. Comparison of the different GANs and DDPM given different datasets, through the FID, KID, Precision, and Recall metrics [25].



Dataset	Model	FID ↓	KID ↓	Precision ↑	Recall ↑
AIROGS	StyleGAN-3	20.43	0.019	0.43	0.19
	Medfusion	11.63	0.008	0.70	0.40
CRCDX	cGAN	49.26	0.036	0.64	0.02
	StyleGAN-3	18.83	0.014	0.57	0.24
	Medfusion	30.03	0.021	0.66	0.41
CheXpert	ProGAN	84.31	0.127	0.30	0.17
	StyleGAN-3	28.69	0.032	0.68	0.08
	Medfusion	17.28	0.020	0.68	0.32

Qualitatively comparing the results, we can see in Fig 13 that the comparison between the StyleGAN-3 and Medfusion is quite obvious in some cases. Similar to the previous case study done by [24], retinal images in Fig 13 (a) were also generated for those with glaucoma, and those without. We can notice that StyleGAN-3 has an obvious difference in color distribution and lack of vessels compared to the real image as well as the one generated by Medfusion. Results from Fig 13 (b) and (c) might be more difficult to qualitatively compare for those who are unfamiliar with these scans, but Table 4 still indicates the quantitative results in a clearer presentation.

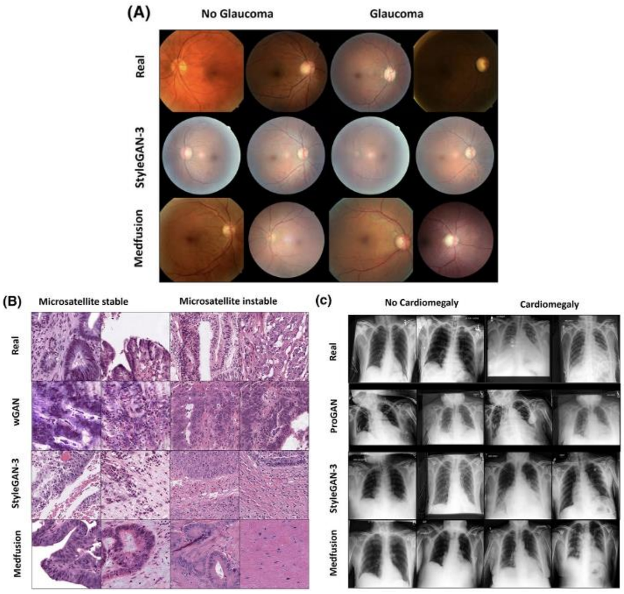


Fig. 13. (a) Retinal images, (b) Tissue samples, (c) Chest X-Rays [25]

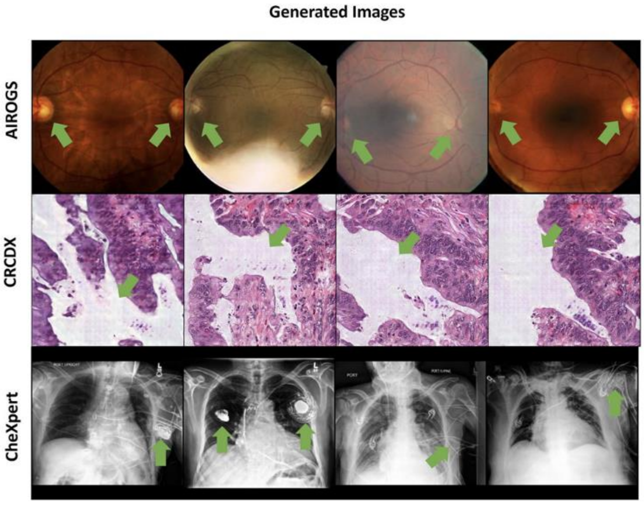


Fig. 14. Images generated by the GAN models show how mode-collapse has affected the results of training. Errors seen are indicated by the green arrows [25].

Some limitations were present during this study, namely the lower resolution images that were used to train the models. This was chosen to stay consistent with the GAN results from previous studies since this study aimed to compare their results to previous studies to showcase the benefits DDPMs have over GANs. All of which were proven to show similar conclusions, that is that the DDPMs have an advantage over GANs. Another detail to note is that the evaluation metrics used to compare the images were not made for medical images, but instead for natural images. While this may be something to consider, we can still note that all the images were compared the same way, and also compared qualitatively, while the scores were used to better visualize a numerical estimate on the comparison between the model. However, it is still important to keep that in mind for future studies since some metrics might eventually be developed to compare medical images in a fairer experiment that might lead to a more accurate comparison between medical images.

Overall, this research conducted by [25] proved that their latent DDPM model outperforms the GAN models once more. Medfusion resulted in lower FID scores, better recall scores, and did not suffer from any problems regarding convergence or training stability, which did occur in some of the GANs used as seen in Fig 14. [25] also hypothesized that due to most medical images being black and white (or limited in color), as opposed to natural images being full of color, it is difficult to generate images with higher diversity since most of the details lie in small changes in detail and textures. When compared to natural images, the medical images still score lower than natural images and did not give an equal result in terms of the metrics used.

Comparing one more case study that conducted similar comparisons within the medical field, is one done by [26] for brain image generation using latent diffusion models, similar to what was done in [25] but with a different dataset and slightly different model comparisons. All the case studies discussed have the same motivation, that is the lack of large medical datasets available to the public for training, costly data, difficult to collect and requires expert skills, and ethical considerations/privacy concerns. The paper also discusses a different method of image generation used in the past, that is combining VAEs with GANs to generate various modalities of full brain volumes from a small training set to generate better results and performances when compared to baseline models. The problem with this, however, is that the images needed to be resized to a small volume and so the images generated did not produce the fine details that were necessary for a good comparison. The computational power needed was also limiting the results and performance [26]. These problems were later reduced when another research conducted by [27] proposed a 3D high resolution GAN to produce 3D images of thorax CT's and brain MRI's at up to 4 times the resolution of previous methods. However, the GAN limitations are still prominent and is why the research that is heavily being conducted nowadays is within diffusion models.

This study used a dataset from the UK Biobank of 31,740 sample images used for training their model. Similar to the previous case study, latent diffusion occurs by encoding the brain images to a latent space, where the diffusion process occurs in (both forward and reverse) and is reconstructed using a decoder to the original data space. Qualitatively, this model proved to be superior to the baseline GAN models and was observed to have comparable images generated at high resolution and sharp detail, shown in Fig 15. However, the GANs required extra fine tuning when designing the model due to the high-resolution nature of MRI images (in 3D as well), and still presented issues regarding instability and mode collapse.

Table 5. Comparison between different baseline models with the Latent Diffusion Model (LDM) proposed in the paper [26].

	FID ↓	MS-SSIM ↓	4-G-R-SSIM ↓
LSGAN	0.0231	0.9997	0.9969
VAE-GAN	0.1576	0.9671	0.8719
LDM	**0.0076**	**0.6555**	**0.3883**
LDM + DDIM	0.0080	0.6704	0.3957
Real images	0.0005	0.6536	0.3909

Table 5 shows a quantitative comparison between the models. Similar quantitative metrics were also used to compare, that is using the FID, but with the added MS-SSIM (Multi-Scale Structural Similarity Metric) and the 4-G-R-SSIM scores (lower is better). We can note that the LDM proposed outperforms all the other models during this test and comparison, which is to be expected considering the previous case studies have shown similar results. We can note that the low FID scores indicate that the generated images through the LDM was low to indicate realism, and the remaining metrics were low to indicate good diversity between the samples generated. We can note that a DDIM sampler was combined with the proposed LDM. DDIM is Denoising Diffusion Implicit Model, which allows for better mapping and efficient sampling with fewer iterations. While the DDIM mix with LDM showed slightly worse scores (practically negligible), the DDIM sampler did reduce the number of timesteps from 1000 steps to 50, which greatly improves the sampling time at almost no cost in the output results.

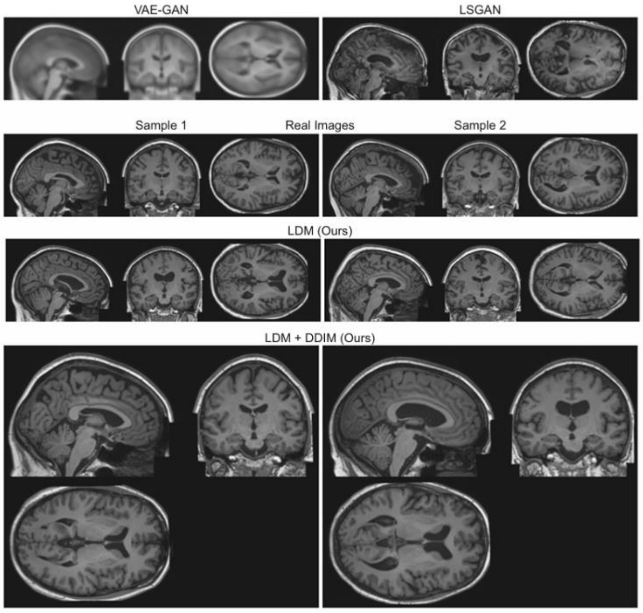


Fig. 15. Comparison between the baseline models (top row) with the real images and proposed LDM (labelled) [26].

The study concluded with the success of their proposed LDM for brain image generation of MRI's. The images generated also corresponded to the expected age, sex, and brain structure volumes to produce images that were within the parameters that were applied.

VIII. Conclusion

The studies explored in this paper showcased many things pertaining to generative models. A brief summary was made to begin, with the development of earlier generative models, their advantages, how they changed the AI scene within their time of discovery, as well as their drawbacks. Their drawbacks, however, prompted newer, more efficient and effective models to be proposed and made to improve upon the drawbacks of previously limited models. Namely, the GANs, which were increasingly popular throughout the past few years and gained lots of traction since its initial proposal in 2014 by [6]. The limitations to this model were then noted and addressed which resulted in the development of diffusion models, with no prominent limitations with even better results.

The focus of this paper was also to emphasize the importance of generative models in the biomedical field. It seems that the rising popularity of diffusion models in entertainment might cloud the benefits of other fields that are overlooked, which is why this paper was written to gather some case studies showing how diffusion models can greatly impact the medical field and healthcare. Inspired by the survey done in [13] which showcases numerous different articles on how diffusion models have been popular within this field, this paper was written to showcase the benefits and drawbacks of two prominent generative models and compare results based on reliable metrics, as well as the performance during this comparison between GANs and Diffusion Models.

The overall results from the three case studies discussed showcased that diffusion models are far superior to GANs when generating images within the biomedical field. The case studies discussed in this paper used datasets that included retinal images, brain MRI's, tissue samples, and chest X-rays. All the models trained using the diffusion process outclassed GANs in terms of quality and reliability. All the models trained using GANs for the datasets discussed also showed to have experienced mode collapse and training instability. Different methods and variations to the GAN were proposed before the discovery of diffusion models, which did improve performance of GANs but did not fully remove the risk of mode collapse. The proposed GAN variations also required extremely fine turning within the design of both the generative and discriminative models, as well as hyper tuning parameters to optimize the models as best as it can without risking mode collapse. This issue was resolved and outperformed by diffusion models, which also has numerous different major variations, the main one discussed in the paper being the latent diffusion model, which takes some similar concepts and methods as the VAEs, but with the diffusion process added into it.

In terms of future works, this paper aims to bring awareness to the limitations of the medical field when it comes to the lack of large datasets that is publicly available, the high cost of acquiring labeled data, the expertise needed for annotating this data, the ethical and privacy concerns that arise when sharing patient data, and the lack of data that is available for rare diseases for analysis. This paper has showcased how diffusion models can be used to combat these problems, and the comparison with previously used models, such as VAEs and GANs. Improvements made in this paper could have been to add more case studies and technical jargon (such as more theoretical detail or math involved in each of the models discussed), but to cohesively compare the two models using strictly experimental data (both qualitatively and quantitatively), this level of detail is sufficient for the scope of this paper.

References

[1] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep Generative Modelling: a Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 1–1, 2021, doi: https://doi.org/10.1109/TPAMI.2021.3116668.

[2] D. Saxena and J. Cao, "Generative Adversarial Networks (gans)," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–42, May 2021. doi:10.1145/3446374

[3] M. Sami and I. Mobin, "A comparative study on variational autoencoders and generative Adversarial Networks," *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, pp. 1–5, Mar. 2019. doi:10.1109/icaiit.2019.8834544

[4] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, "The helmholtz machine," *Unsupervised Learning*, pp. 277–292, May 1999. doi:10.7551/mitpress/7011.003.0017

[5] Y. Pang, J. Lin, T. Qin, and Z. Chen, "Image-to-image translation: Methods and applications," *IEEE Transactions on Multimedia*, vol. 24, pp. 3859–3881, 2022. doi:10.1109/tmm.2021.3109419

[6] I. J. Goodfellow *et al.*, *Generative Adversarial Nets*, Jun. 2014. doi:https://doi.org/10.48550/arXiv.1406.2661

[7] A. Popuri and J. Miller, "Generative adversarial networks in Image Generation and recognition," *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 1294–1297, Dec. 2023. doi:10.1109/csci62032.2023.00212

[8] Bhagyashree, V. Kushwaha, and G. C. Nandi, "Study of prevention of mode collapse in Generative Adversarial Network (GAN)," *2020 IEEE 4th Conference on Information &amp; Communication Technology (CICT)*, pp. 1–6, Dec. 2020. doi:10.1109/cict51604.2020.9312049

[9] Y. Kossale, M. Airaj, and A. Darouichi, "Mode collapse in Generative Adversarial Networks: An overview," *2022 8th International Conference on Optimization and Applications (ICOA)*, pp. 1–6, Oct. 2022. doi:10.1109/icoa55659.2022.9934291

[10] A. Gainetdinov, "Gan mode collapse explanation," Medium, https://pub.towardsai.net/gan-mode-collapse-explanation-fa5f9124ee73 (accessed Dec. 4, 2024).

[11] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023. doi:10.1109/tpami.2023.3261988

[12] L. Yang *et al.*, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, Nov. 2023. doi:10.1145/3626235

[13] A. Kazerouni *et al.*, "Diffusion models in medical imaging: A comprehensive survey," *Medical Image Analysis*, vol. 88, p. 102846, Aug. 2023. doi:10.1016/j.media.2023.102846

[14] D. Kuznedelev, V. Startsev, D. Shlenskii, and S. Kastryulin, "Does Diffusion Beat GAN in Image Super Resolution?," *arXiv (Cornell University)*, May 2024, doi: https://doi.org/10.48550/arxiv.2405.17261.

[15] J. Ho, A. Jain, and P. Abbeel, *Denoising Diffusion Probabilistic Models*, Jun. 2020. doi:https://doi.org/10.48550/arXiv.2006.11239

[16] Z. Liu, C. Ma, W. She, and M. Xie, "Biomedical Image Segmentation Using Denoising Diffusion Probabilistic Models: A Comprehensive Review and Analysis," *Applied Sciences*, vol. 14, no. 2, p. 632, Jan. 2024, doi: https://doi.org/10.3390/app14020632.

[17] R. Bayat, "A Study on Sample Diversity in Generative Models: GANs vs. Diffusion Models." *Tiny Papers at the International Conference on Learning Representations* , 2023.

[18] Y. Song and Stefano Ermon, "Generative Modeling by Estimating Gradients of the Data Distribution," *arXiv (Cornell University)*, Jul. 2019, doi: https://doi.org/10.48550/arxiv.1907.05600.

[19] P. Dhariwal and A. Nichol, *Diffusion Models Beat GANs on Image Synthesis*, Jun. 2021. doi:https://doi.org/10.48550/arXiv.2105.05233

[20] X. Yu, J. Tian, and Z. Hu, *An Analysis for Image-to-Image Translation and Style Transfer*, Aug. 2024. doi:https://doi.org/10.48550/arXiv.2408.06000

[21] A. Borji, Generated Faces in the Wild: Quantitative Comparison of Stable Diffusion, Midjourney and DALL-E 2, Jun. 2023. doi:https://doi.org/10.48550/arXiv.2210.00586

[22] K.-Q. Zhou and H. Nabus, "The ethical implications of dall-E: Opportunities and challenges," *Mesopotamian Journal of Computer Science*, pp. 17–23, Jan. 2023. doi:10.58496/mjcsc/2023/003

[23] Y. Wu *et al.*, "Retinal OCT synthesis with denoising diffusion probabilistic models for layer segmentation," *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, May 2024. doi:10.1109/isbi56570.2024.10635836

[24] A. Alimanov and M. B. Islam, "Denoising diffusion probabilistic model for retinal image generation and segmentation," *2023 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–12, Jul. 2023. doi:10.1109/iccp56744.2023.10233841

[25] G. Müller-Franzes *et al.*, "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial

networks for Medical Image Synthesis," *Scientific Reports*, vol. 13, no. 1, Jul. 2023. doi:10.1038/s41598-023-39278-0

[26] W. H. Pinaya *et al.*, "Brain imaging generation with latent diffusion models," *Lecture Notes in Computer Science*, pp. 117–126, 2022. doi:10.1007/978-3-031-18576-2_12

[27] L. Sun *et al.*, "Hierarchical amortized gan for 3D high resolution medical image synthesis," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3966–3975, Aug. 2022. doi:10.1109/jbhi.2022.3172976

[28] B. Khosravi *et al.*, "Few-shot biomedical image segmentation using diffusion models: Beyond image generation," *Computer Methods and Programs in Biomedicine*, vol. 242, pp. 107832–107832, Sep. 2023, doi: https://doi.org/10.1016/j.cmpb.2023.107832.