A SURVEY ON IMAGE GENERATION TECHNIQUES: PARADIGMS, EVOLUTION, DEEP LEARNING ADVANCEMENTS AND FUTURE DIRECTIONS

Nrupesh Shah[1], Dr. Sanjay Patel[2]

•

[1]Research Scholar, GTU, Ahmedabad.
nrupesh_shah@gecg28.ac.in
[2]Associate Professor, GEC, Gandhinagar.
sp_patel1@gtu.edu.in

Abstract

Image generation techniques have witnessed significant advancements in recent years. Classification of Image Generation Approaches is an important topic in today's rapidly advancing technological landscape. We will examine primary approaches in this field: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Stable Diffusion Processes (SDPs) and other similar approaches. GANs, VAEs and SDPs have shown remarkable performance in terms of image quality, as well as scalability and efficiency. This survey focuses onto seminal works that have shaped the field of image generation, spanning Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Autoregressive Models, Flow based image generation as well as diffusion-based approaches, and also compares various evaluation metrices in the field. We also provide a list of challenges in this field.

Keywords: Image generation, Image synthesis, text-to-image, image-to-image, GANs, VAEs, Stable diffusion, Autoregressive Models, Flow Based Image generation models, Generative Adversarial Networks, Variational Autoencoders.

I. Introduction

Image generation is the task of synthesizing new images from scratch. This is a challenging task, as images are high-dimensional data and it is difficult to generate realistic and diverse images. However, recent advances in machine learning have led to development of various powerful image generation techniques. Image generation has found applications in diverse fields such as art, design, entertainment, and medical imaging. Over the years, various strategies have been developed to create images, each with its own set of advantages and challenges.

Image generation is a rapidly evolving field, and new techniques are being developed all the time. This review paper will provide a comprehensive overview of the state-of-the-art in image generation techniques. The paper will discuss the different types of image generation techniques, their strengths and weaknesses, and their applications. There have already been some surveys/reviews carried out in the field of image generation. Each of them has contributed greatly to the field. For example, a paper by Trevisan de Souza et el. [1] specifically reviews Generative Adversarial Networks (GANs) for image generation. It explains the fundamental architecture of

DOI: https://doi.org/10.24412/1932-2321-2025-489-82-104

GANs, involving a generator and a discriminator in adversarial training to create realistic, novel samples. The paper focuses on the applications of GANs in image processing and synthesis. It discusses how challenges associated with GANs, such as mode collapse, stability, and convergence, have been addressed. The review covers the basics and notable GAN architectures, detailing techniques used in various image-based applications including image synthesis, image enhancement, image-to-image translation, image manipulation, style transfer, super resolution, and image repair. It also touches upon using GANs for text-to-image synthesis. The paper is intended as a guide for those new to studying GANs.

Another review by Singh et el.[2] focuses *specifically* on the applications of Generative Adversarial Networks (GANs) in the domain of medical imaging. It describes GANs as unsupervised deep learning tools that are particularly useful for analysing multimodal medical imaging data. The paper highlights the utility of GANs in generating realistic medical images and annotations for tasks like data augmentation, image registration, reconstruction, and image-to-image translation within the medical field. It reviews popular GAN variants, such as DCGAN, LAPGAN, pix2pix, CycleGAN, and UNIT, and how they are applied in medical imaging. The survey covers applications in medical image reconstruction and synthesis across different modalities. It notes the significant rise in GAN studies in medical imaging, particularly focusing on cross-modality synthesis and MRI image synthesis, and discusses future research directions.

Whereas, Elasri et el. [3] provides a comprehensive overview of existing image generation tasks, focusing on creating images from various data types beyond just other images. It discusses generating images from text, sketch, scene graphs, object layouts, and images themselves. The survey highlights how deep learning, particularly GANs, has made generating new, realistic images with high quality possible. It categorises methods based on the nature of adopted architectures, the type of input data used, and the main objective. The paper presents a taxonomy of approaches including image-to-image translation, sketch-to-image translation, conditional image generation, text-to-image generation, and more. It also details public datasets and evaluation metrics used in the field, offering a comparison of performance for significant methods. Current challenges facing image generation are also presented. The paper notes that GANs and conditional GANs (cGANs) are commonly used architectures for conditional image generation approaches.

There may be some overlap between our review and others work but our main and unique contributions are summarized as follows: Diverging from prior surveys, this review provides a focused examination of deep learning's role in image generation. We trace the developmental trajectory of these models, offering a comparative analysis of their architectures, core principles, and diverse applications. Additionally, we delve into available evaluation metrics, identify present limitations, and outline future directions, incorporating the latest advancements in the field.

I. Applications of Image Generation in the field of Reliability Engineering

While image generation techniques might seem distant from reliability engineering at first glance, there are several ways this paper (or the broader field it surveys) could be related or contribute:

1. **Synthetic Data Generation for Testing and Validation:** Many modern systems, especially autonomous vehicles, robotics, and industrial automation, heavily rely on AI for tasks like object recognition, anomaly detection, and predictive maintenance. Training and validating these AI models require vast amounts of data. Image generation techniques, particularly those using GANs (Generative Adversarial Networks) or VAEs (Variational Autoencoders) discussed in the paper, can generate synthetic images that mimic real-world scenarios.

   In reliability engineering, rare events (e.g., specific types of failures, unusual operating conditions) are often critical but difficult to capture sufficient real data for. Image generation can create synthetic data for these rare events, allowing for more robust testing and training of models designed to detect or predict them. This improves the reliability of the system's performance under diverse and challenging circumstances.

Synthetic images can be used for data augmentation, making AI models more robust to variations in input. This directly contributes to the reliability of AI systems, as they are less likely to fail when encountering slightly different or noisy data.

2. **Visual Inspection and Quality Control:** Image generation can help train systems for automated visual inspection in manufacturing and quality control. By generating images of "normal" products and various types of defects (even rare ones), the survey's discussed techniques can lead to more reliable automated inspection systems that reduce human error and improve product reliability.

   In some cases, visual cues (e.g., cracks, wear and tear patterns) are crucial for predictive maintenance. Image generation can create synthetic deterioration patterns, allowing models to be trained to reliably identify early signs of failure, thus improving equipment uptime and reliability.
3. **Digital Twins and Simulation:** In the context of digital twins, accurate and realistic visual representations of physical assets are important for monitoring, analysis, and simulation. Advanced image generation techniques can contribute to creating highly realistic digital twins, which in turn can be used for reliability analysis and what-if scenarios without impacting the physical asset. This can help in predicting failures and optimizing maintenance strategies.
4. **Security and Tampering Detection:** While less direct, the advancements in image generation also bring challenges related to deepfakes and manipulated images. Understanding these generation techniques (as surveyed in the paper) is crucial for developing robust and reliable methods to detect forged or tampered images, which can be critical for maintaining the integrity of data used in reliability assessments or incident investigations.

In essence, the paper on image generation techniques contributes to reliability engineering by enabling the creation of synthetic, yet realistic, data that can be used to:

- Improve the reliability of AI-systems with better training and validation.
- Enhance automated inspection and quality control processes.
- Support predictive maintenance by simulating various fault conditions.
- Facilitate more comprehensive digital twin environments for reliability analysis.

By providing a comprehensive overview of the state-of-the-art in image generation, the paper equips researchers and engineers in reliability with knowledge of powerful tools to address data scarcity, improve model robustness, and ultimately build more reliable systems.

II. Historical Overview of Image Generation

In recent years, there has been a sudden increase of interest in image generation techniques. This is due to the development of new deep learning techniques, the availability of large datasets of images, and the increasing demand for realistic images in a variety of applications, such as entertainment, marketing, and education. Image generation has come a long way since its early days. In the 1970s, researchers began to explore the use of procedural algorithms to generate images. These algorithms were based on mathematical rules that could be used to create specific patterns or shapes. For example, one algorithm could be used to generate a checkerboard pattern, while another could be used to generate a spiral.

Procedural algorithms were a major breakthrough in image generation, but they had some limitations. They were often limited to generating simple patterns, and they could be difficult to control. In the 1980s, researchers began to explore the use of neural networks for image generation. Neural networks are inspired by the human brain, and they can be used to learn complex patterns from data. One of the earliest neural network-based image generation models was called the Neocognitron [4]. The Neocognitron was developed in the 1980s by Kunihiko Fukushima, and it was able to generate images of simple objects, such as lines and circles.

Some of the most recent advances in image generation include: Generative Adversarial Networks (GANs) [5], Variational Autoencoders (VAEs) [6], Diffusion Models [7], etc. We will review recent advancements in each of these approaches in this paper. Image generation is a rapidly evolving field, and there are sure to be even more advances in the years to come. These advances will enable us to generate even more realistic and creative images, and they will have a huge impact on a variety of industries.

III. Classification of Image Generation Task

Image generation is a fundamental task in computer vision and artificial intelligence that involves creating new images that resemble a given dataset or follow a specific style or distribution. The goal of image generation is to create visually plausible and diverse images that exhibit characteristics similar to those found in real images. This task finds applications in a wide range of fields, including art, entertainment, data augmentation, medical imaging, and more. Image generation can be classified based on several parameters that describe different aspects of the task. Here, we'll classify image generation based on various key parameters provided below:

I. Input Domain:

1. **Random Noise Input:** Here the model takes random noise as input and learn to generate images that resemble real data distribution. GAN and VAE are good examples of such approaches.
2. **Conditional Input (Text, Labels, or Images):** The models generate images conditioned on additional input, such as text descriptions, labels, or images. The conditional information guides the image generation process. Models like conditional GANs are used.
3. **Text-to-Image Generation:** In this we generate images from textual descriptions, translating the semantic information into visual content. Models like AttnGAN are useful for this task[8].
4. **Images as Input (Style Transfer):** Style transfer methods modify the style of an input image using the style of another image while preserving content [9].
5. **Incomplete Images (Image Inpainting):** Inpainting models fill in missing or corrupted regions in images while maintaining coherence and realism[10].
6. **Reference Images (Super-Resolution):** Single Image Super-Resolution: These models enhance the resolution of low-resolution images using high-resolution reference images during training [11].
7. **Noisy or Degraded Images (Denoising):** Denoising models remove noise from images, enhancing their quality and clarity [12].
8. **Medical Data (Medical Image Generation):** Models generate medical images for diagnosis, research, and data augmentation in the medical field [2].
9. **Artistic Style as Input (Artistic Image Generation):** Models create images in specific artistic styles based on input style images or characteristics [13].
10. **Multi-Modal Inputs (Cross-Modal Generation):** Here, Models generate images from different modalities (e.g., text to images or images to text) [14].

II. Generative Approach:

1. **Explicit Models:** These models define a direct mapping from a latent space to the image space, such as Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs).
2. **Implicit Models:** These models learn the distribution of the data through sampling, like autoregressive models and some flow-based models.

III. Architecture:

1. **Autoencoders:** Models that learn to encode and decode data to reconstruct input images, e.g., VAEs.
2. **GANs:** Models involving a generator and discriminator network that compete with each other to produce realistic images.
3. **Flow-Based Models:** Models that use series of invertible transformations to map a simple distribution like gaussian distribution to the complex data distribution.
4. **Autoregressive Modelling:** The models use values of previous pixels to compute the value of the next pixel. Images are generated pixel by pixel.
5. **Diffusers:** Models that learn by adding and removing noise in a gradual manner in the given input image.

IV. Latent Space:

1. **Continuous:** Models with continuous latent spaces (e.g., VAEs) that allow for smooth interpolation and exploration.
2. **Discrete:** Models with discrete latent spaces (e.g., some flow-based models) that capture categorical features.

V. Loss Function:

1. **Adversarial Loss:** Models using adversarial training, like GANs, to minimize the difference between real images and generated images.
2. **Reconstruction Loss:** Models minimizing the difference between input and reconstructed images, as seen in VAEs.

VI. Sampling Method:

1. **Direct Sampling:** Models that directly sample images from the learned distribution, common in GANs and some flow-based models.
2. **Autoregressive Sampling:** Models generating images sequentially by predicting pixels one at a time, like PixelCNN [15].

VII. Output Domain:

1. **Natural Images:** Models generating images that resemble natural scenes, objects, and textures.
2. **Artistic Images:** Models creating images with specific artistic styles or characteristics.

3. **Medical Images:** Models generating medical images for diagnosis and research.
4. **Scientific Data:** Models generating data visualizations for scientific exploration.

VIII. Use Cases:

1. **Data Augmentation:** Generating synthetic data to augment limited datasets This is especially useful in medical field machine learning as the data available is limited.
2. **Art Creation:** Generating artistic images, paintings, and visual effects.
3. **Style Transfer:** Transforming images into different artistic styles.

IV. Major Approaches in Image Generation

There are many major approaches in image generation, which includes variable autoencoders, generative adversarial networks, transformer-based architectures, autoregressive models, diffusion models, etc. Some of the most common approaches are explained below:

I. Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of machine learning models designed to produce novel data. They operate through an adversarial process between two neural networks: a generator tasked with creating new data, and a discriminator that evaluates its authenticity.

GANs have been used to generate a variety of different types of data, including images, text, and music. They have been shown to be able to generate realistic and diverse data that is indistinguishable from real data.

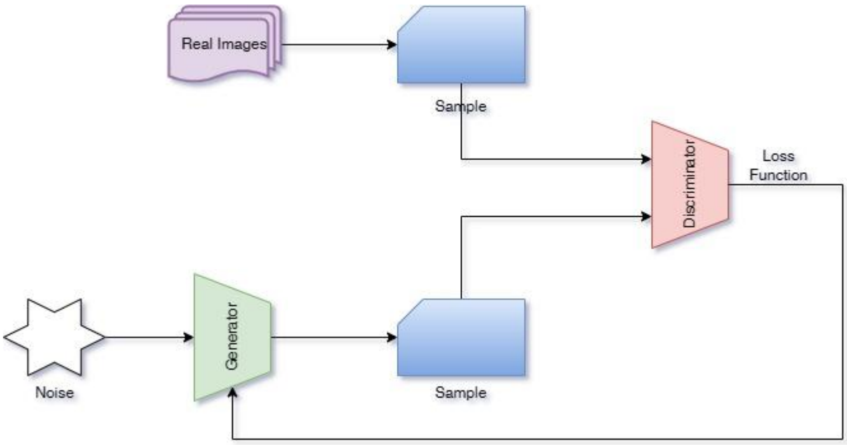

Figure 1: *GAN Architecture*

The architecture of a GAN is shown in the figure above. The generator is a neural network that takes a random noise as input and outputs a new data sample. The discriminator is also a neural network that gets a data sample as input and outputs a probability that the data sample is real. The generator and discriminator are trained together in an adversarial manner. The generator is trained to create data that the discriminator cannot distinguish from real data. The discriminator is trained to become better at detecting fake data.

Here is a comparison of various advancements proposed over time in the basic GAN architecture.

Table **1:** *Comparison of GAN Based Approaches*

GAN Approach	Advantages	Limitations	Applications
DCGAN (Deep Convolutio-nal GAN) [16]	Stable training, high-quality images	Limited to simple datasets, mode collapse	Art generation, data augmentation
WGAN (Wasserstein GAN) [17]	Improved stability, meaningful loss	Hyperparameter sensitivity, slow convergence	Image synthesis, style transfer
LSGAN (Least Squares GAN) [18]	More stable training, sharper images	Mode collapse, harder to optimize	Image generation, data augmentation
InfoGAN [19]	Interpretable latent factors, controlled generation	Limited to discrete latent variables, sensitive to hyperparameter	Image synthesis, style manipulation
StyleGAN [20]	High-resolution images, fine-grained control	Complex architecture, resource-intensive	Artistic image synthesis
BigGAN [21]	High-resolution images, large batch size	Computationally expensive, resource-intensive	High-quality image generation

DCGANs [16] propose a set of architectural constraints for Convolutional Neural Networks used in GANs to make them more stable to train. Key recommendations for these networks involve using strided convolutions instead of pooling layers in the discriminator, and fractional-strided convolutions in the generator. Additionally, widespread batch normalization in both networks and the elimination of fully connected hidden layers for deeper models are proposed. Specific activation functions are recommended: ReLU in the generator (except Tanh output) and LeakyReLU in the discriminator. These architectural decisions contribute to improved training stability compared to earlier attempts to scale GANs with CNNs. DCGANs are shown to learn a hierarchy of representations, and their learned features are effective for downstream tasks like image classification.

WGANs [17] introduce a fundamental shift in the GAN objective by proposing the use of the Earth Mover (Wasserstein) distance as the metric for comparing the real and generated data distributions. This distance is theoretically shown to have better properties than the JS or KL divergences used in traditional GANs, particularly when the distributions lie on low-dimensional manifolds, addressing issues like vanishing gradients. WGAN training is significantly more stable and less prone to mode collapse compared to standard GANs, without needing careful balancing of generator and discriminator capacities. A key practical benefit is that the critic's loss provides a meaningful learning curve that correlates with sample quality and convergence, which is invaluable for debugging and hyperparameter tuning.

LSGANs [18] propose an alternative objective function for GANs to address training stability issues. They replace the sigmoid cross-entropy loss used in regular GANs with a least squares loss function for the discriminator. This modification helps alleviate the vanishing gradients problem that can occur in regular GANs, especially when the discriminator becomes confident. The least square’s objective corresponds to minimizing the Pearson χ^2 divergence [22]. LSGANs demonstrate improved training stability and can converge to a good state even without batch normalization in some configurations. Empirically, LSGANs generate more realistic images and are less prone to mode collapse compared to regular GANs. Conditional LSGANs were also successfully applied to Chinese character generation.

InfoGAN [19] is an information-theoretic extension to the standard GAN framework. Its core contribution is modifying the GAN objective to maximize the mutual information between a small subset of the latent variables and the observed generated samples. This objective encourages the model to learn interpretable and disentangled representations in a completely unsupervised manner, without requiring labelled data for the disentangled factors. A simple and efficient algorithm is derived to optimize a lower bound of this mutual information objective. InfoGAN successfully disentangles various semantic features like writing styles, pose, lighting, hair styles, and emotions across different datasets, achieving results comparable to supervised methods for learning disentangled representations. The added computational cost compared to standard GAN training is negligible.

Karras et el. [20] builds upon the StyleGAN [23] architecture, addressing characteristic artifacts such as blob shapes. They propose architectural and training changes, notably redesigning the generator normalization by replacing instance normalization with weight demodulation to eliminate artifacts. The progressive growing method used in the original StyleGAN is replaced with an alternative design that shifts focus to higher resolutions without altering network topology during training. The improved model incorporates skip connections in the generator and a residual discriminator architecture for better performance. They also introduce path length regularization, which leads to more reliable models and makes projecting images back into the latent space much easier. These improvements significantly advance the state of the art in unconditional image synthesis.

The BigGAN[21] paper demonstrates that scaling up GAN training dramatically improves performance. They train models with significantly more parameters and larger batch sizes than prior art. Key architectural changes include using shared class embeddings and skip connections (skip-z) in the generator for efficiency and performance gains. A crucial side effect of their modifications, including orthogonal regularization, is the discovery of the truncation trick, which allows fine control over variety and fidelity. Their models claim state-of-the-art results on ImageNet and also perform well on the larger JFT-300M dataset. The paper also analyses instabilities specific to large-scale training, particularly concerning the singular values of the generator's weights.

II. Variational Autoencoders (VAEs) [3]

Variational Autoencoders (VAEs) are generative models that can learn a compressed, lower dimensional, feature-rich latent representation of input data. This allows them to subsequently generate novel data by sampling from this learned latent space and then decoding those samples.

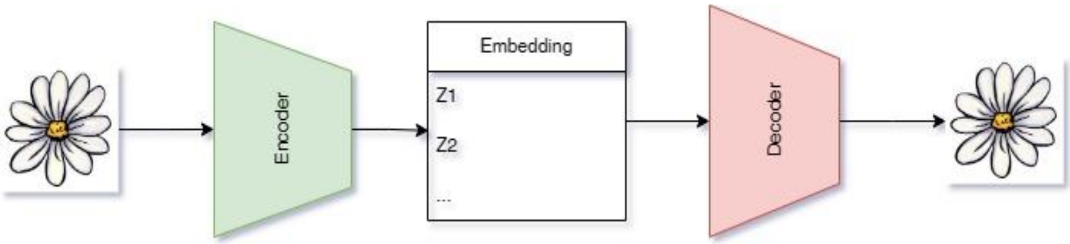

Figure 2: *VAE Architecture*

The basic architecture of a VAE is shown in the figure above. The encoder is a neural network that takes the observed data (image) as input and outputs a latent representation of the input, which is a compressed version of the input. The decoder is also a neural network that takes the latent representation as input and outputs the original data (image). VAEs are based on the idea of variational inference. This is a technique for estimating the posterior distribution of latent variable given observed data. In the case of VAEs, the latent variable is the latent representation of the data, and the observed data is the original data. We provide a tabular comparison of various Variational Autoencoder (VAE) approaches for image generation.:

Table 2: *Comparison of VAE Based Approaches*

VAE Approach	Advantages	Limitations	Applications
Vanilla VAE [24]	Simple architecture, easy to implement	Blurry images, difficulty in capturing complex data distribution	Data compression, image generation
Β-VAE [25]	Disentangled latent space, controlled generation	Mode collapse, trade-off in reconstruction quality	Interpretability, data compression
CVAE (Conditional VAE) [26]	Conditional image generation	Limited to paired data, sensitive to label noise	Image-to-image translation
AAE (Adversarial Autoencoder) [27]	Improved image quality, less blurriness	Vulnerable to mode collapse, harder to train	Image generation, data denoising
VQ-VAE [28]	Discrete latent space, clear structure	Limited diversity in generated samples	Speech synthesis, image generation
DR-VAE (Disentangled Representation VAE) [29]	Disentangled factors, improved interpretability	Complexity, challenging optimization	Data generation, representation learning
FactorVAE [30]	Independent across the dimensions Better trade-off between disentanglement and reconstruction quality	Complexity and Interpretability, Hyperparameter Tuning	Guide Learning in Highly Noisy Market Data, Risk Modelling

Vanilla Variational Autoencoder (VAE) [24] framework proposed by Kingma et el. is an efficient method for inference and learning in deep directed graphical models, with continuous latent variables. This is achieved using the stochastic gradient variational Bayes (SGVB) estimator and the reparameterization trick. The training objective is based on maximizing the variational lower bound of the log-likelihood. While the original VAE demonstrated potential for disentanglement on simple datasets, its performance did not scale to more complex ones.

β-VAE[25] introduces a modification of the original VAE framework. The key contribution is the introduction of an adjustable hyperparameter β. This β coefficient modulates learning constraints, specifically balancing the latent channel capacity and independence constraints against reconstruction accuracy. For β > 1, β-VAE is pushed to learn a more efficient and disentangled latent representation. The paper demonstrates that β-VAE qualitatively and quantitatively outperforms the original VAE (where β=1) and other state-of-the-art methods like InfoGAN and DC-IGN in achieving disentanglement. β-VAE is also noted for being stable to train, requiring few assumptions, and depending only on tuning β. It has the ability to discover more latent factors and learn better disentangled representations over a wider range of factor values than baselines.

Sohn et el. [26] proposes the Conditional Variational Autoencoder (CVAE) for structured output prediction problems, such as image segmentation. The core idea is to model the conditional distribution of output variables given the input, which can be multi-modal. The CVAE architecture builds recognition, prior, and generation networks on top of a baseline CNN and incorporates recurrent connections. Compared to deterministic neural networks, CVAE and its variants show significant improvements in structured output prediction accuracy and conditional log-likelihoods. They also generate more realistic and diverse output samples.

Adversarial Autoencoder (AAE) [27] framework transforms any VAE into a generative model by imposing an arbitrary prior distribution on the latent codes. This is done using a GAN to match the aggregated posterior distribution of the autoencoder's latent codes to the desired prior. AAE is shown to successfully impose specified prior distributions and fill the latent space more effectively than VAE. It can be extended for semi-supervised learning and to disentangle label information from image style by providing the label to the decoder. AAE achieved state-of-the-art likelihoods on tested datasets.

Vector Quantised-Variational AutoEncoder (VQ-VAE) [28] modifies the VAE framework by using discrete latent variables instead of continuous ones, achieved through vector quantization. A key improvement is that this approach helps to avoid posterior collapse, a common issue in VAEs with powerful decoders. VQ-VAE learns a learnt prior distribution and uses an embedding table indexed by discrete latent codes. It demonstrates performance comparable to continuous VAEs in terms of likelihood and can generate high-quality samples when paired with an autoregressive prior. VQ-VAE is effective at modelling long-range dependencies.

Wang et el. [29] categorizes various disentangled representational learning approaches, including. It notes that the vanilla VAE shows poor disentanglement on complex datasets despite having potential on simple ones. The author highlights how many subsequent VAE-based methods improve disentanglement by adding extra regularizers or inductive biases to the VAE objective, citing β-VAE, DIP-VAE, and β-TCVAE as examples. It explains the trade-off in β-VAE between reconstruction quality and independence. The paper also discusses hierarchical VAE structures for learning factors at different abstraction levels.

Kim et el. introduces FactorVAE [30], another method based on the VAE framework. Its primary contribution is addressing the trade-off between reconstruction quality and disentanglement seen in β-VAE. FactorVAE achieves this by augmenting the VAE objective with a specific penalty on the Total Correlation of the latent codes. This encourages independence in the latent space. Empirically, FactorVAE achieves higher disentanglement scores than β-VAE while maintaining similar or better reconstruction quality on tested datasets. The paper also proposes an improved disentanglement metric.

III. Autoregressive Models

Autoregressive models for generate images by modeling the conditional distribution of each pixel or group of pixels given their preceding pixels. These models generate images one pixel at a time, sequentially predicting the value of next pixel based on the values of previously generated pixels. Autoregressive models have been successful in producing high-quality images with fine-grained control over the generation process.

In autoregressive models, an image is conceptualized as a sequence of pixels. The overall probability of an image is expressed as the product of the conditional probabilities of individual pixels, where the intensity of each pixel is determined by the values of all preceding pixels, as represented in the following equation.

$$p(x) = \prod_{i=1}^{n^2} p(x_i | x_1, \ldots\ldots, x_{i-1}) \tag{1}$$

In other words, to generate the value of pixel x_i in below figure, we need the intensity values of all previously generated pixels shown in blue.

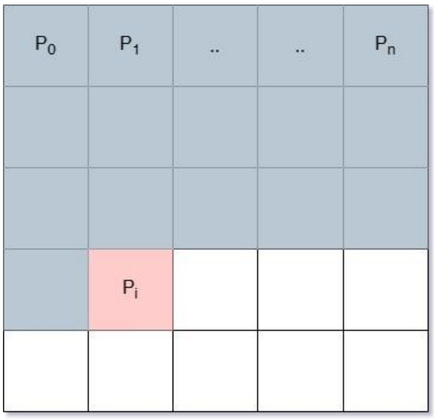

Figure 3: *Image Pixel value Calculation in An Autoregressive Model*

Now, we compare some prominent autoregressive approaches in the following table.

Table 3: *Comparison of Autoregressive Approaches*

Autoregressive Approach	Advantages	Limitations	Applications
PixelRNN [31]	High-quality images, fine-grained control	Slow generation speed, not scalable	High-resolution image generation
PixelCNN [32]	Efficient generation, controllable generation	Slow generation speed, receptive field limitations	Image generation, style transfer
DRAW (Deep Recurrent Attentive Writer)[33]	Attention-based generation, variable sequence length	Complex architecture, challenging training	Image generation, handwriting synthesis

PixelRNN [31] sequentially predict pixels in an image into two spatial dimensions. Pixel-level autoregressive models, such as PixelRNN and PixelCNN, offer a tractable way to model the complete dependencies within an image by factoring the joint distribution of pixels into conditional probabilities. Pixel RNN introduces unique 2D LSTM layers (Row LSTM and Diagonal BiLSTM) that leverage convolutions for spatial state computation. Meanwhile, PixelCNN is a fully convolutional network that uses masked convolutions to preserve spatial resolution and respect the autoregressive order. This approach has led to significantly better log-likelihood scores on datasets like CIFAR-10 and ImageNet, where PixelCNN models pixels as discrete variables via a simple yet effective multinomial distribution with a softmax layer. The inclusion of residual connections around LSTM layers further aids in training deeper networks and accelerating convergence.

Van Den et el. Introduces Conditional Image Generation with PixelCNN Decoders [32], which explores conditional image synthesis by adapting and improving PixelCNN [31]. It introduces a gated variant of PixelCNN (Gated PixelCNN) that matches the log-likelihood of PixelRNN on datasets like CIFAR and ImageNet while significantly reducing training time. It addresses the blind spot in the receptive field of the original masked convolutional architecture using two stacks of CNNs and introduces Conditional PixelCNN that allows modelling complex conditional distributions of natural images when a latent vector embedding is provided. According to the authors, the single Conditional PixelCNN model demonstrates versatility: it can generate images across diverse classes by being conditioned on a one-hot encoding (e.g., ImageNet classes), and it can produce varied new portraits of a single individual when conditioned on a face embedding from a convolutional network.

A significant innovation in generative models, the Deep Recurrent Attentive Writer (DRAW) network [33] integrates a novel spatial attention mechanism that mimics human vision with a sequential variational auto-encoder. What sets DRAW apart from other VAEs is its use of recurrent networks for both the encoder and decoder, which iteratively refine images by exchanging latent codes. Its dynamically updated attention mechanism enables the model to selectively read from the input and write to the output at each timestep. This fully differentiable attention makes DRAW trainable with standard backpropagation and leads to substantial improvements in generative modeling for MNIST and realistic SVHN image generation, with the authors also highlighting its advantages for image classification on cluttered MNIST.

VI. Flow Based Approaches

Unlike other generative models, flow-based models generate images through a series of invertible transformations, which also allows them to directly learn the exact probability distribution of the data. While Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have excelled at learning complex data distributions, they both struggle with exact probability evaluation and inference. This limitation can lead to lower quality results in VAEs and present training challenges for GANs, such as mode collapse and vanishing gradients. Normalizing Flows were introduced to address these issues. By leveraging invertible transformation functions, they eliminate the need for adding noise during image generation and offer a significantly more stable training process compared to GANs.

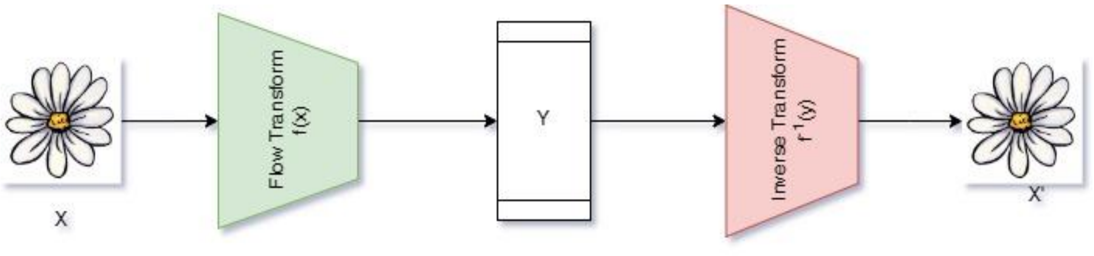

Figure 4: *Flow based model architecture*

Flow-based generative models leverage invertible transformation functions to establish a mapping from the data space (x) to the latent space (z). A crucial aspect is that the dimensionality of z must match that of x. This section presents a brief comparative overview of different flow-based methodologies.

Table 4: *Comparison of Flow Based Approaches*

Flow-Based Approach	Advantages	Limitations	Applications
Real NVP (Real Non-Volume Preserving) [34]	Exact likelihood, impressive sampling	Computational complexity	Density estimation, image synthesis
Glow [35]	High-quality images, efficient sampling	Computational complexity, resource-intensive	High-resolution image generation
MAF (Masked Autoregressive Flow) [36]	Flexible modeling, controllable generation	Slow generation speed, less efficient sampling	Image synthesis, density estimation
FFJORD (Continuous Normalizing Flows) [37]	Faster training, reversible dynamics	Trade-off between speed and quality	Image synthesis, density estimation

For unsupervised learning, Dinh et al. [34] developed real-valued non-volume preserving (real NVP) transformations. These are robust, stably invertible, and learnable, offering key advantages such as exact log-likelihood calculation, as well as exact and efficient sampling and efficient latent variable inference. It addresses the challenge of building powerful yet trainable models for high-dimensional data. Real NVP extends previous work like NICE by defining a more flexible class of bijective functions. A key contribution is the use of a deep multi-scale architecture leveraging batch normalization and residual networks. Sampling is efficient and parallelized over input dimensions, and the model learns a semantically meaningful latent space. Real NVP generates sharper samples by not relying on fixed form reconstruction costs like L2 norm. It incorporates a novel variant of batch normalization more robust for small minibatches. The coupling layer that is used in Real NVP is identified as a special case of the autoregressive transformation found in MAF and IAF.

Kingma proposed Glow [35], a generative flow model that builds upon the NICE and RealNVP architectures. Each step of the Glow flow is composed of three key components: actnorm, an invertible 1×1 convolution, and an affine coupling layer. Actnorm is a novel activation normalization layer with data-dependent initialization that is more robust to small batch sizes than batch normalization. The invertible 1× 1 convolution replaces fixed permutations with a learned, invertible linear transformation across channels, which improves performance and convergence speed. Glow incorporates affine coupling layers similar to RealNVP and uses a multi-scale architecture. It achieves a notable improvement in log-likelihood when tested on standard benchmarks compared to RealNVP. When trained on high-resolution data, Glow demonstrates the ability to synthesize realistic images and allows for efficient, parallelizable synthesis and latent space manipulation. The model also benefits from potential memory savings due to its reversible architecture.

Masked Autoregressive Flow (MAF) [36] is a type of normalizing flow specifically designed for density estimation. MAF is constructed by stacking multiple autoregressive models, where each layer models the random numbers of the next layer. A key innovation is the use of the

Masked Autoencoder for Distribution Estimation (MADE) as the building block. This allows density evaluations to be computed efficiently in a single forward pass through the entire flow on parallel hardware like GPUs, a significant advantage for density estimation compared to IAF which requires D passes. The paper shows that MAF is a generalization of Real NVP, with the coupling layer being a special case of the autoregressive layer used in MAF. Stacking MADE layers increases the model's flexibility, allowing it to learn multimodal conditionals. MAF achieves comparable performance on general-purpose density estimation tasks and outperforms Real NVP in all experiments presented. Batch normalization is used to improve training stability and performance.

FFJORD [37] is a continuous-time invertible generative model defined by ordinary differential equations (ODE). A major contribution is enabling unrestricted neural network architectures for the dynamics function, overcoming the architectural constraints imposed by previous flow-based models like Real NVP and Glow to ensure tractable Jacobian determinants. FFJORD achieves this by using Hutchinson’s trace estimator to provide a scalable unbiased estimate of the log-density with an efficient O(D) time cost, significantly improving upon the O(D^3) or O(D^2) costs of previous methods. The model allows one-pass sampling. Backpropagation for training is handled using the adjoint method, which solves another ODE backwards in time. FFJORD demonstrates competitive performance on density estimation and improved results in variational inference compared to other normalizing flows. It also uses significantly fewer parameters than Glow on image datasets.

V. Diffusion Models

To generate images, diffusion models begin with a noise image and incrementally introduce more detail through a guided diffusion process. This new approach in generative modeling allows for the creation of realistic images and offers improved efficiency over GANs and VAEs.

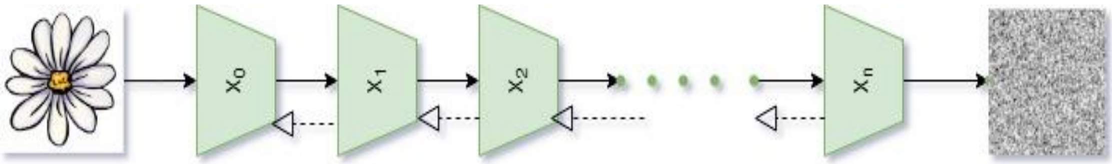

Figure 5: *Diffusion Model*

Stable Diffusion-based approaches focus on modeling data as the result of a diffusion process, often based on partial differential equations. These methods leverage the concept of diffusion to generate images in a stable and controlled manner. Here is the comparison of various approaches based on Stable Diffusion.

Table 5: *Comparison of Diffusion Based Approaches*

Diffusion Approach	Advantages	Limitations	Applications
Latent Diffusion [38]	High-quality images with fine details.	Slow to generate especially for large images.	Image generation, image inpainting, image style transfer, image upscaling

Fast Sampling of Score-Based Models with Cyclical Diffusion Sampling [39]	Enhances effectiveness of SGLD in learning complex multimodal distributions. Decreases inference time. Easy integration to existing models.	Approximation based accumulates truncation errors over time. Require right balance between the number of function evaluations (NFE) and image quality.	Image generation on diverse domains
Consistent Diffusion Models [40]	Enhance the quality of generated samples for both conditional and unconditional generation tasks.	The regularization process leads to longer training times. The method assumes that the learned vector field is conservative, which is not directly tested or enforced.	Applied across various modalities, including images, videos, audio, 3D structures, proteins, and medical applications.
simple diffusion [41]	Generate high-resolution images while maintaining simplicity. Supports a single end-to-end training setup without the need for cascades or mixtures of experts.	Training diffusion models on high-resolution images is computationally expensive. Require more training data.	Image generation
DuDGAN [42]	Dual diffusion-based noise injection process helps to prevent overfitting. Fast Convergence.	Struggles with classes with less than 50 images. Independent noising schedules for the discriminator and classifier can complicate optimization.	Diffusion-based 3D scene generation. GAN inversion involving multi-network training
Versatile Diffusion [43]	Handle multiple modalities and tasks. More efficient than training separate for each task.	Training and inference are compute-intensive. Complex to fine-tune. Limitations in cross-modal blending.	Text to Image generation. Image to Text generation.

Latent Diffusion Models (LDMs), proposed by Rombach et al. [38], revolutionize high-resolution image synthesis by moving the computationally intensive diffusion process from pixel space to a compressed latent space. This significantly reduces training expense and speeds up inference while maintaining high synthesis quality. LDMs are built upon a sequential application of denoising autoencoders and employ a reweighted variational lower bound. They exhibit enhanced sample quality when trained in VQ-regularized latent spaces compared to pixel-based diffusion models. A key innovation is the integration of cross-attention conditioning, which enables flexible control, including powerful text-to-image generation. Consequently, LDMs achieve state-of-the-art performance and strong competitiveness across a range of generative tasks and metrics.

While diffusion models show great promise in generative tasks, their slow sampling is a significant hurdle. Makhtidi et al. [39] offer a solution with Cyclical Diffusion Sampling, which aims to speed up this process. Their method integrates cyclical stochastic gradient Langevin dynamics (SGLD)—a technique proven to enhance stability when learning complex distributions—with the elucidated diffusion models (EDM) sampler. By strategically using cyclical step-size noise scheduling and Heun's second-order method within this stochastic sampler, the research achieves remarkable improvements in image quality and a substantial reduction in inference time. For example, it outperformed prior methods on CIFAR-10, needing only 35 function evaluations (NFE) instead of 1,000, all without requiring any network architecture modifications.

Daras et el. [40] proposed Consistent Diffusion Models: Mitigating Sampling Drift by Learning to be Consistent. Diffusion models generate data by denoising a noisy sample from a distribution corrupted with noise. This paper presents Consistent Diffusion Models (CDM), which aim to mitigate sampling drift introduced by imperfect learning of the score function. CDM relies on the architecture and hyperparameters of the EDM model. The proposed method involves training with a Constant Prediction (CP) regularization alongside the standard Denoising Score Matching (DSM) objective. The CDM models are trained using a weighted objective combining both DSM and CP regularization. CDM aims to enforce a property that on expectation, predictions should not change for points with the same origin. Best results were obtained by combining DSM and CP regularization. While a concurrent work (Consistency Models) also enforces a similar property, CDM's primary motivation is to improve generation quality, whereas Consistency Models aim to accelerate sampling.

Hoogeboom et el. [41] introduced simple diffusion: End-to-end diffusion for high resolution images. Applying standard diffusion models directly to high-resolution images in pixel space is challenging and typically requires complex approaches like latent diffusion or cascades. This paper aims to improve standard denoising diffusion for high resolutions while keeping the model simple. The method introduces several simple modifications to the original formulation. Key findings include: 1) Adjusting the noise schedule is crucial for high-resolution images. 2) Scaling only specific parts of the architecture is sufficient. 3) Strategic use of dropout improves performance. 4) Down-sampling is an effective strategy for higher resolutions. Combining these techniques allows for training a single denoising diffusion model on resolutions up to 512x512. Simple diffusion claims state-of-the-art performance on ImageNet among diffusion models without sampling modifiers. It is presented as the first single-stage text-to-image model capable of generating such high visual quality images. The model can also be distilled for faster sampling.

DuDGAN [42]proposes a novel approach for class-conditioned image generation using GANs. It addresses challenges faced by traditional class-conditional GANs such as instability, mode collapse, and low-quality output on datasets with significant intra-class variation. DuDGAN proposes a dual diffusion-based noise injection process into two neural networks: the discriminator and a classifier. Gaussian-mixture noises are injected in distinct ways, helping to prevent overfitting and introducing more challenging tasks. The method investigates the impact of using an additional classifier trained with this diffusion-based noise injection. The dual-diffusion training approach signifies the collaboration between the discriminator and the classifier. DuDGAN achieves fast convergence within a limited number of iterations and iteration-efficient training. It claims to outperform traditional conditional GAN models on various datasets (AFHQ, Food-101, CIFAR-10, BAAT) across metrics like FID, KID, Precision, and Recall.

Versatile Diffusion [43] is a multi-task multimodal network designed to handle multiple flows (text-to-image, image-variation, image-to-text) within one unified model, instead of existing diffusion models that often require separate models for different tasks (e.g., text-to-image, image-variation). This novel multi-flow diffusion framework extends single-flow diffusion pipelines. It categorizes diffuser layers as global, data, or context, activating them selectively based on input and output modalities. This structure promotes significant parameter sharing across tasks. VD uses a UNet with cross-attentions for the diffuser, VAEs (Autoencoder-KL and Optimus) for latent representations, and CLIP encoders for context. VD performs well on its base tasks, outperforming baselines and better capturing context semantics. Its multi-flow nature enables novel derivative tasks such as semantic-style disentanglement and dual-/multi-context blending. VD represents a step toward universal AI by addressing multiple modalities and tasks in a single framework.

V. Comparison of Different Image Synthesis Techniques

We have seen many different image synthesis techniques, each with its own strengths and weaknesses. Here is a comparison of some of the most common techniques:

1. **Generative Adversarial Networks (GANs):** GANs are one of the most powerful image synthesis techniques available. They can generate very realistic images, and they can be used to generate images of a wide variety of objects and scenes. However, GANs can also be difficult to train, and they can be prone to generating images that are blurry or distorted.
2. **Variational Autoencoders (VAEs):** VAEs are less powerful than GANs, but they are easier to train and they are less prone to generating blurry or distorted images. VAEs are also better at generating images that are consistent with the original image.
3. **Autoregressive Models:** They are able to generate high-quality images with good detail. They can be slow to generate images, especially for large images compared to other models.
4. **Flow Based Models:** They are very efficient to train and generate compared to GANs and Stable diffusion. But they can sometimes generate images that are blurry or unrealistic.
5. **Diffusion Models:** Diffusion models can generate realistic images, and they are also more efficient than GANs and VAEs. Diffusion models are a relatively new technique, but they have shown promising results. They are able to generate realistic images that are better than GANs, and they are much faster to train.
6. **Procedural Generation:** Procedural generation is a technique for creating images by following a set of rules. This can be done by using mathematical formulas, or by using algorithms that generate images based on a set of parameters. Procedural generation is often used to create stylized images, such as those used in video games. Procedural generation is a versatile technique that can be used to create a wide variety of images. It is not as realistic as GANs or VAEs, but it can be used to create images that are unique and visually appealing.
7. **Neural Style Transfer:** It is a technique for transferring the style of one image to another image. This is done by using a neural network to learn the style of first image, and then using that knowledge to apply the style to the second image. Neural style transfer can be used to create images with a variety of different styles, such as impressionism, cubism, and abstract art. Neural style transfer is a powerful technique that can be used to create visually appealing images. However, it is not as realistic as GANs or VAEs.

The ideal image synthesis technique varies with the application. For realism, consider GANs; for consistency, VAEs are a strong choice. If speed is critical, diffusion models excel, while procedural generation is great for stylized outputs. When a specific artistic flair is needed, neural style transfer is the way to go. Bear in mind that image synthesis is a swiftly developing domain, with new advancements constantly emerging. Future techniques are expected to be significantly more powerful and adaptable than those available today.

VI. EVALUATION METRICS FOR IMAGE GENERATION

Evaluation metrics for image generation are methods to measure the quality and diversity of the generated images. Some of the common metrics are:

1. **Inception Score (IS):** This metric calculates the KL divergence between the conditional and marginal distributions of the class labels predicted by an Inception network on the generated images. A high IS indicates that the images are both realistic and diverse. IS is a quantitative metric that was proposed by Salimans. (2016) [44]. To use this metric, you need to feed your generated images to a pre-trained Inception network and compute the statistics of the predicted class labels.
2. **Fréchet Inception Distance (FID):** This metric computes the Wasserstein-2 distance between the feature distributions of real and generated images, extracted by an Inception network. A low FID indicates that the generated images are similar to the real ones in terms of visual quality and diversity. FID is a quantitative metric that was introduced by Heusel. (2017) [45]. To use this metric, you need to feed both real and generated images to a pre-trained Inception network and compute the mean and covariance of the features in a hidden layer.
3. **Precision and Recall:** These metrics are based on the nearest-neighbor distance between real and generated images in the feature space. Precision measures how many generated images are close to some real images, while recall measures how many real images are close to some generated images. A high precision and recall indicate that the generated images are both realistic and diverse. These metrics are quantitative and were developed by Sajjadi (2018) [46]. To use these metrics, you need to feed both real and generated images to a pre-trained Inception network and find the nearest neighbors of each image in the feature space using a distance metric such as cosine similarity or Euclidean distance.
4. **Kernel Inception Distance (KID):** This metric estimates the Maximum Mean Discrepancy (MMD) between the feature distributions of real and generated images, extracted by an Inception network. KID is similar to FID, but it does not require fitting a Gaussian distribution to the features, and it is unbiased and consistent. KID is a quantitative metric that was suggested by Binkowski. (2018) [47]. To use this metric, you need to feed both real and generated images to a pre-trained Inception network and compute the MMD between the features using a kernel function such as radial basis function or polynomial kernel.
5. **Learned Perceptual Image Patch Similarity (LPIPS):** This metric measures the perceptual similarity between real and generated images using a deep network that is trained to predict human judgments of image similarity. LPIPS is a quantitative metric that can capture fine-grained differences in image quality and style. LPIPS was proposed by Zhang. (2018) [48]. To use this metric, you need to feed both real and generated images to a pre-trained network that is fine-tuned on human similarity ratings and compute the average distance between the images.

VII. Limitations and Challenges in Image Generation

Image generation is a rapidly evolving field, but there are still some limitations and challenges that need to be addressed. Here are a few of the most common limitations and challenges:

1. **Image quality:** Image generation techniques are still not perfect, and they can sometimes generate images that are blurry, distorted, or unrealistic. This is especially true for complex images, such as images of people or animals.
2. **Controllability:** Image generation techniques are often not very controllable. This means that it can be difficult to generate images with specific features, such as a certain color, a certain style, or a certain object.
3. **Efficiency:** Image generation techniques can be computationally expensive, especially for complex images. This can make it difficult to generate images in real time, which is necessary for some applications, such as virtual reality and augmented reality.
4. **Bias:** Image generation techniques can be biased, which means that they may generate images that are more likely to represent certain groups of people or objects. This is a concern because it can lead to the spread of harmful stereotypes.
5. **Safety:** Image generation techniques can be used to create fake images, which can be used to deceive people. This is a concern because it can be used for malicious purposes, such as spreading misinformation or propaganda.
6. **Training Instability:** Models like GANs and VAEs can be comparatively challenging to train due to instability.
7. **Scalability**: Complexity of the approaches can hinder the scalability of some models.

These are just a few of the limitations and challenges that need to be addressed in image generation. As the field continues to evolve, it is likely that these limitations and challenges will be overcome, and that image generation techniques will become even more powerful and versatile.

VIII. Future Trends in Image Synthesis

Image generation is a rapidly evolving field, and there are many exciting future trends that are likely to emerge in the coming years. We can hope that all the challenges that we have listed earlier can be solved in the near future. Here is what it can look like:

1. **Improved Image Quality:** Image generation techniques are constantly improving, and it is likely that we will see even more realistic and detailed images in the future. This will be made possible by the development of more powerful neural networks and by the availability of larger datasets of images. Improved image quality will be used in a variety of applications, such as virtual reality, augmented reality, and film production. In virtual reality, improved image quality will allow users to experience more realistic and immersive environments. In augmented reality, improved image quality will allow users to interact with virtual objects in a more realistic way. In film production, improved image quality will allow filmmakers to create more realistic and visually appealing movies.
2. **More Controllable Image Generation:** Image generation techniques are becoming more controllable, which means that users will be able to generate images with specific features, such as a certain color, a certain style, or a certain object. More controllable image generation will be used in variety of applications, such as product design, marketing, and art. In product design, more controllable image generation will allow designers to create more realistic prototypes of products. In marketing, more controllable image generation will allow marketers to create more visually appealing ads. In art, more controllable image generation will allow artists to create more original and creative works of art.
3. **Real-Time Image Generation:** Image generation techniques are becoming more efficient, which means that they will be able to generate images in real time. Real-time image generation will be used in a variety of applications, such as video games, social media, and

live streaming. In video games, real-time image generation will allow players to experience more realistic and immersive worlds. In social media, real-time image generation will allow users to create more visually appealing content. In live streaming, real-time image generation will allow viewers to experience more realistic and engaging events.

4. **Less Biased Image Generation:** Image generation techniques are becoming less biased, which means that they will generate images that are representative of all groups of people and objects. This will be made possible by the development of new techniques, such as debiasing and fairness training. Less biased image generation will be used in a variety of applications, such as education, healthcare, and law enforcement. In education, less biased image generation will allow students to learn about different cultures and groups of people. In healthcare, less biased image generation will allow doctors to diagnose diseases more accurately. In law enforcement, less biased image generation will allow police officers to make more informed decisions.
5. **Safer Image Generation:** Image generation techniques are becoming safer, which means that they will be less likely to be used to create fake images. This will be made possible by the development of new techniques, such as watermarking and authentication. Safer image generation will be used in a variety of applications, such as social media, news, and advertising. In social media, safer image generation will help to prevent the spread of fake news and misinformation. In news, safer image generation will help to ensure that the public is exposed to accurate information. In advertising, safer image generation will help to prevent consumers from being misled by false advertising.

These are just a few of the ways that the future trends in image generation will be used in the real world. As the field continues to evolve, we can expect to see even more innovative and exciting applications for image generation.

IX. Conclusion: The Evolution and Potential of Image Generation Techniques

In this review paper, we have discussed the state-of-the-art in image generation techniques, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), diffusion models, autoregressive models, etc. While each strategy has its own merits and limitations, the rapid progress of deep learning has ushered in a new era of image generation capabilities. Researchers and practitioners can use this comparative analysis to select the most suitable strategy for their specific needs, taking into account factors like task requirements, dataset size, and computational resources. We have also discussed the different metrics that can be used to evaluate the quality of generated images. As image generation continues to evolve, we anticipate the emergence of novel strategies that combine the strengths of existing approaches and push the boundaries of visual creativity. The future of image generation research is bright. With the continued development of new techniques, it is likely that we will see even more realistic and diverse images being generated in the future. These images could be used for a variety of applications, such as entertainment, art, and education.

References

[1] V. L. Trevisan de Souza, B. A. D. Marques, H. C. Batagelo, and J. P. Gois, “A review on Generative Adversarial Networks for image generation,” Computers and Graphics (Pergamon), vol. 114, pp. 13–25, Aug. 2023, doi: 10.1016/j.cag.2023.05.010.

[2] N. K. Singh and K. Raza, “Medical Image Generation using Generative Adversarial Networks Medical Image Generation Using Generative Adversarial Networks: A Review”, doi: 10.48550/arXiv.2005.10687.

[3] M. Elasri, O. Elharrouss, S. Al-Maadeed, and H. Tairi, “Image Generation: A Review,” Oct.

01, 2022, Springer. doi: 10.1007/s11063-022-10777-x.

[4] K. Fukushima, "Neocognitron," Scholarpedia, vol. 2, no. 1, p. 1717, 2007, doi: 10.4249/scholarpedia.1717.

[5] I. Goodfellow et al., "Generative adversarial nets," Adv Neural Inf Process Syst, vol. 27, 2014.

[6] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," Jun. 2019. doi: 10.1561/2200000056.

[7] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," Adv Neural Inf Process Syst, vol. 34, pp. 8780–8794, 2021.

[8] T. Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks." [Online]. Available: https://github.com/taoxugit/AttnGAN.

[9] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks."

[10] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context Encoders: Feature Learning by Inpainting," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536–2544.

[11] C. Ledig et al., "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4681–4690.

[12] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders," in Proceedings of the 25th international conference on Machine learning, 2008, pp. 1096–1103.

[13] L. A. Gatys, A. S. Ecker, and M. Bethge, "Texture and art with deep neural networks," Curr Opin Neurobiol, vol. 46, pp. 178–186, Oct. 2017, doi: 10.1016/j.conb.2017.08.019.

[14] H. Zhang et al., "Cross-Modal Contrastive Learning for Text-to-Image Generation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 833–842.

[15] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "PIXELCNN++: IMPROVING THE PIXELCNN WITH DISCRETIZED LOGISTIC MIXTURE LIKELIHOOD AND OTHER MODIFICATIONS," in International Conference on Learning Representations, 2017.

[16] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," arXiv e-prints, p. arXiv-1511, Nov. 2015,

[17] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in International conference on machine learning, 2017, pp. 214–223.

[18] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least Squares Generative Adversarial Networks."

[19] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets."

[20] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119.

[21] A. Brock, J. Donahue, and K. Simonyan, "LARGE SCALE GAN TRAINING FOR HIGH FIDELITY NATURAL IMAGE SYNTHESIS," in International Conference on Learning Representations, 2018.

[22] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[23] D. P. Kingma, M. Welling, and Others, "Auto-Encoding Variational Bayes," Dec. 20, 2013, Banff, Canada.

[24] I. Higgins et al., "β-VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK," in International conference on learning representations, 2017.

[25] K. Sohn, X. Yan, and H. Lee, "Learning Structured Output Representation using Deep Conditional Generative Models."

[26] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and G. Brain, "ADVERSARIAL AUTOENCODERS", 2015.

[27] A. van den Oord DeepMind, O. Vinyals DeepMind, and K. Kavukcuoglu DeepMind, "Neural Discrete Representation Learning," Adv Neural Inf Process Syst, vol. 30, 2017.

[28] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, "Disentangled Representation Learning," IEEE Trans Pattern Anal Mach Intell, pp. 1–20, Jul. 2024, doi: 10.1109/tpami.2024.3420937.

[29] H. Kim and A. Mnih, "Disentangling by Factorising," in International conference on machine learning, 2018, pp. 2649–2658.

[30] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel Recurrent Neural Networks," in International conference on machine learning, 2016, pp. 1747–1756.

[31] A. Van Den et al., "Conditional Image Generation with PixelCNN Decoders," Adv Neural Inf Process Syst, vol. 29, 2016.

[32] K. Gregor, D. Com, D. J. Rezende, and D. Wierstra, "DRAW: A Recurrent Neural Network For Image Generation Ivo Danihelka," in International conference on machine learning, 2015, pp. 1462–1471.

[33] L. Dinh, J. Sohl-Dickstein, and B. Samy, "DENSITY ESTIMATION USING REAL NVP," in International Conference on Learning Representations, 2017.

[34] D. P. Kingma and P. Dhariwal, "Glow: Generative Flow with Invertible 1×1 Convolutions," Adv Neural Inf Process Syst, vol. 31, 2018.

[35] G. Papamakarios, T. Pavlakou, and I. Murray, "Masked Autoregressive Flow for Density Estimation."

[36] W. Grathwohl, R. T. Q Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, "FFJORD: FREE-FORM CONTINUOUS DYNAMICS FOR SCALABLE REVERSIBLE GENERATIVE MODELS," in International Conference on Learning Representations, 2018.

[37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models."

[38] K. Makhtidi, A. Bustamam, R. Adnan, H. A. Robbani, W. Mangunwardoyo, and M. A. Khan, "Fast Sampling of Score-Based Models with Cyclical Diffusion Sampling," IEEE Access, vol. 12, pp. 49578–49589, 2024, doi: 10.1109/ACCESS.2024.3365146.

[39] G. Daras, Y. Dagan, A. G. Dimakis, and C. Daskalakis, "Consistent Diffusion Models: Mitigating Sampling Drift by Learning to be Consistent," Adv Neural Inf Process Syst, vol. 36, pp. 42038–42063, 2023.

[40] E. Hoogeboom, J. Heek, and T. Salimans, "simple diffusion: End-to-end diffusion for high resolution images," in International Conference on Machine Learning, 2023, pp. 13213–13232.

[41] T. Yeom, C. Gu, and M. Lee, "DuDGAN: Improving Class-Conditional GANs via Dual-Diffusion," IEEE Access, vol. 12, pp. 39651–39661, 2024, doi: 10.1109/ACCESS.2024.3372996.

[42] X. Xu, Z. Wang, E. Zhang, K. Wang, and H. Shi, "Versatile Diffusion: Text, Images and Variations All in One Diffusion Model," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7754–7765.

[43] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs."

[44] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," Adv Neural Inf Process Syst, vol. 30, 2017.

[45] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing Generative Models via Precision and Recall for Learning Systems," Adv Neural Inf Process Syst, vol. 31, 2018.

[46] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in International Conference on Learning Representations, Jan. 2018.

[47] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.