International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Survey on Generative Models for Image Synthesis: Gans, Diffusion Models, and Beyond

Rashmi V[1], Radhika S.K[2]

[1,2]Assistant Professor Department of CSE JNN college of Engineering

Abstract

Generative models have changed computer vision by enabling realistic and diverse image synthesis. This survey provides a comprehensive review of popular generative models; Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models. We examine their architectures, training methodologies, strengths, limitations, and applications. Key datasets and evaluation metrics are discussed, in addition to insights obtained from comparative analysis. Additionally, ethical considerations and research challenges such as training instability, mode collapse, and computational complexity are addressed and potential future directions are explored to guide research in efficient, interpretable, and safe generative modeling.

Keywords: Generative Models, Image Synthesis, GANs, Diffusion Models, Variational Autoencoders, Deep Learning, Computer Vision, Ethical AI

1. INTRODUCTION

Image synthesis has emerged as a pivotal step in computer vision that spans entertainment and gaming, digital art, medical imaging, and scientific research. The goal of image synthesis is to develop novel images that resemble real-world data by learning probability distribution of existing datasets. High-quality synthetic images can enhance training datasets, enable creative content generation, and support research in domains where collecting real data is difficult or costly.

Over the past decade, several generative modeling paradigms have demonstrated remarkable capabilities in image synthesis:

- **Generative Adversarial Networks (GANs):** GANs employ a generator–discriminator framework, where the generator produces synthetic images and the discriminator checks their authenticity. The adversarial training process pushes the generator to create increasingly realistic images, making GANs highly effective for producing **photorealistic outputs**.
- **Variational Autoencoders (VAEs):** Proposed by Kingma & Welling (2013), VAEs encode images into a probabilistic latent space and reconstruct them via a decoder. They offer stable training and interpretable latent representations, which are essential for tasks such as data augmentation, anomaly detection, and controlled image manipulation. While VAEs may produce slightly blurred outputs compared to GANs, their continuous latent space provides significant advantages for exploring and generating structured variations of images.
- **Diffusion Models:** Latest advancements in generative modeling comprises Denoising Diffusion Probabilistic Models (DDPMs), which generate high-fidelity images through an iterative denoising

process. By refining random noise into coherent images, diffusion models achieve superior visual quality compared to traditional GANs, capturing fine textures and details with remarkable realism.

This survey provides an overview of the generative models- architectural comparisons, practical applications, commonly used datasets, and evaluation metrics. Additionally, it identifies the current challenges in generative modeling and explores future research **directions**, particularly in hybrid and emerging models that sums up the strengths of multiple strategies for improved performance, stability, and interpretability.

2. Background and Preliminaries

Generative Modeling Overview

Generative models aim to understand the underlying data distribution p(x) from a given dataset and generate novel samples that resemble real data. Here, which focus on predicting labels or categories for input data, generative models create new data instances that are statistically alike training sets. This property to synthesize realistic data makes generative modeling essential for applications such as image synthesis, data augmentation, anomaly detection, and creative content generation.

A few key concepts underpin generative modeling:

- **Latent Space:** A compressed, low-dimensional form of the input data that captures essential features and structure. Latent spaces allow generative models to encode meaningful variations in the data, enabling controlled manipulation and interpolation between samples.
- **Noise Vector:** A random input vector used by models - GANs and Diffusion Models to create diverse outputs. By sampling different noise vectors, models can produce a wide variety of images that still conform to the learned data distribution.
- **Training Objectives:** Different generative models optimize distinct objectives to learn the distribution of data effectively:
- **GANs:** Optimize a minimax game between the generator, which attempts to produce realistic samples, and the discriminator, which seeks to seggregate real from generated samples.
- **VAEs:** Maximize the **Evidence Lower Bound (ELBO)**, balancing reconstruction accuracy and regularization of the latent space to match a prior distribution.
- **Diffusion Models:** Learn to reverse a forward noise-adding process, progressively denoising images to create high-fidelity outputs that match the original data distribution.

Understanding these fundamental concepts aids in exploring the architectures, applications, and comparative strengths of GANs, VAEs, and Diffusion Models, which are covered in the following sections.

3. Generative Models

3.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) rely on adversarial training across two neural networks: a generator (G) and a discriminator (D). The generator creates realistic images from random noise, whereas the discriminator aims to segregate real images sampled from the dataset and the synthetic images obtained from generator. Mathematically, GAN training involves solving a minimax optimization problem:

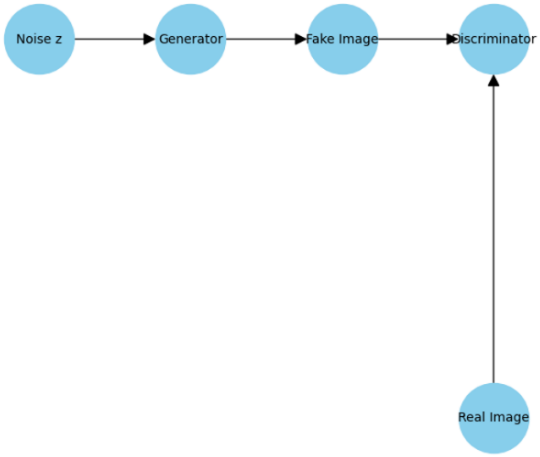
$$min_G min_D V(D,G) = E_{x\sim Pdata}[log\ log\ D(x)] = E_{z\sim p_z}[log\ (1 - D\big(G(z)\big))]$$

Where Pdata is the data distribution and z~p is the prior distribution of the noise vector z.

Several versions of GANs have been designed to improve stability and output quality, including:

- **DCGAN (Deep Convolutional GAN):** Uses convolutional architectures for both generator and discriminator, enhancing image quality for structured data.
- **WGAN (Wasserstein GAN):** Introduces the Wasserstein distance to improve training stability and reduce mode collapse.
- **StyleGAN:** Incorporates style-based latent manipulation to generate high-resolution, photorealistic images with controllable features.

Figure 1: GAN architecture showing Generator and Discriminator training loop



3.2 Variational Autoencoders (VAEs)

Variational Autoencoders (VAEs) encode input into a latent space and reconstruct it probabilistically. A VAE consists of three main components: an encoder, a decoder, and a latent variable z. The encoder maps an input image x to a probability distribution in the latent space, typically characterized by a mean μ and variance σ, whereas the decoder rebuilds the image from a sampled latent vector z. The training of a VAE is done to maximize the likelihood of reconstructed data while regularizing the latent space to follow a prior distribution, usually a standard Gaussian. This objective is formalized as the Evidence Lower Bound (ELBO):

$$L_{VAE} = E_{q(z|x)}[log\ log\ p(x|z)] - D_{KL}(q(x)||p(z))$$

Here, q(z|x) is the approximate posterior obtinaed by the encoder, p(x|z) is the likelihood of the data with latent variable given, and D_{KL} is the Kullback-Leibler divergence that confirms the latent distribution remains close to the prior p(z). The first term encourages accurate reconstruction, while the second term regularizes the latent space that allows VAEs to create varied and smooth outputs. VAEs are widely used in applications such as data augmentation, anomaly detection, and interpretable generative modeling as a consequence of training and meaningful latent representations.

Figure 2: VAE Encoder-Decoder Architecture

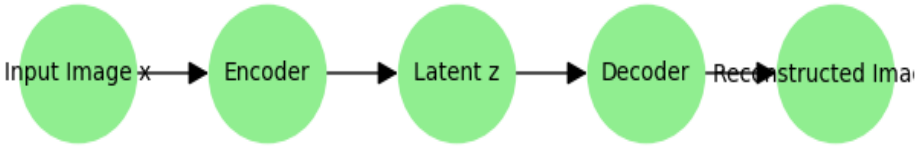


Figure 2 illustrates the Variational Autoencoder (VAE) workflow. The encoder processes the input image x and maps it to a latent space that is represented by a probability distribution. A latent vector z is sampled and passed through the decoder to rebuild the actual image. This setup lets the VAE learn a useful latent format and enables diverse, smooth image generation.

VAEs do a lot more than just crunch numbers. They learn the underlying patterns in data, which makes them pretty useful in different areas. One big thing people use them for is data augmentation. Basically, VAEs can dream up new samples that look a lot like the real stuff, so machine learning models get exposed to more variety and end up stronger. They're also handy for spotting anomalies. If you feed something weird into a VAE, the gap between what you put in and what comes out usually jumps, making it easier to catch outliers.

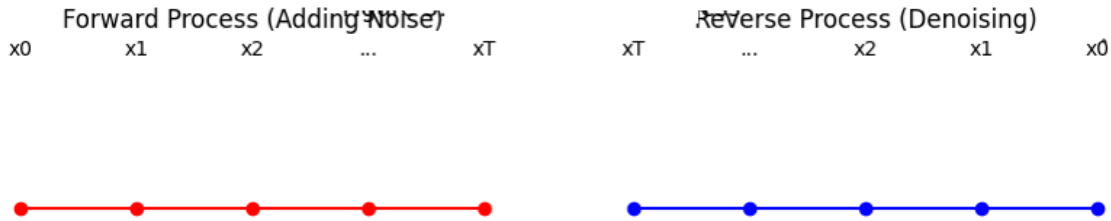
VAEs really shine when it comes to stable training and their easy-to-understand latent space. They don't suffer from the training headaches or mode collapse that often trip up GANs, so you can usually train them on all kinds of datasets without much drama. That latent space? It's compact and continuous, which makes it simple to play around with—tweak a few numbers, and you can generate new images with specific changes.

Still, VAEs aren't perfect. Their biggest issue is the output images—they're usually a bit blurry, especially next to the crisp results you get from GANs. That's because VAEs focus on capturing the overall data distribution instead of chasing perfect, photorealistic images. So, if you need super high-quality visuals, they might not be your first pick. Even so, their reliability and the way you can interpret what's going on under the hood keep them popular in both research and real-world projects.

3.3 Diffusion Models

Diffusion Models create images by iteratively denoising a random noise input until a high-quality image is produced. Unlike GANs, which learn a direct mapping from noise to image, diffusion models gradually refine the image through a series of denoising steps, making them highly stable and can generate high-fidelity outputs. This iterative approach allows the model to get fine-grained details and produce realistic textures, resulting in images that are often of great visual quality than the ones generated by other generative models.

Figure 3: Diffusion Model Process



Diffusion models create images by iteratively denoising a noisy input, beginning from random noise and progressively reconstructing the image. The forward process fills noise on multiple stages, while the reverse process removes noise step by step to produce a high-fidelity output. This process is depicted in Figure 3, where the forward and reverse stages are clearly separated, demonstrating how the model transforms noise into a coherent image.

Diffusion models are popping up everywhere these days. You'll see them behind text-to-image tools like DALL·E 2 and Stable Diffusion, turning your words into detailed, photorealistic pictures. Scientists use them too—think of creating synthetic microscope or space images to boost research or test out new ideas.

In the creative world, artists and content creators rely on these models for high-res image generation, digital art, media projects, and building out virtual environments.

But they're not perfect. The biggest headache? They eat up a lot of computing power. Every image means running the network back and forth several times, so both training and generating images can get expensive and slow, especially if you want results right away. Still, people keep pushing for faster, smarter versions. Research on speeding things up is moving fast, and honestly, diffusion models keep looking like one of the most exciting paths forward for making top-tier images.

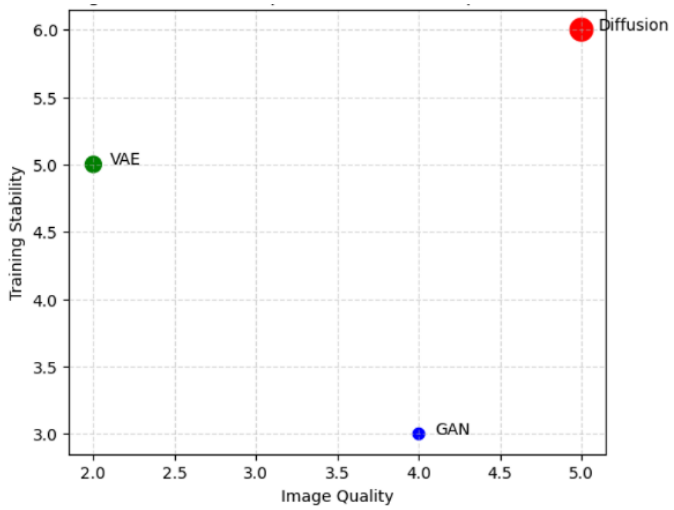
3.4 Hybrid and Emerging Models

As the field of generative modeling evolves, researchers are exploring hybrid architectures and emerging approaches that combine the strengths of existing models or introduce novel paradigms. VAE-GAN hybrids really stand out. They take the best from both worlds—VAEs give you stable training and meaningful latent spaces, while GANs deliver those sharp, high-quality images everyone wants. In these setups, you've got a VAE's encoder and decoder working together with a discriminator that checks if the images look real. The result? Outputs that make sense structurally and actually look good.

Then there's the rise of score-based generative models. These build on diffusion models, but instead of directly modeling the data, they use the score function. They keep refining noisy images step by step, a bit like diffusion models do, but usually with quicker convergence and better samples. People see these as a strong choice for high-fidelity image creation.

Transformers have also shaken things up in this space. You get models like ViT-GAN hybrids—Vision Transformers handle wide-range dependencies and the big-picture context in images, while GAN parts make sure the results look convincing. These transformer-based models are great at picking up on complex patterns and relationships, which opens up all kinds of possibilities: generating high-res scenes, conditional image synthesis, and even working across multiple types of data.

Figure 4: Comparative bubble chart — image quality (x-axis), training stability (y-axis), bubble size = computational cost; showing GAN, VAE, Diffusion positions.



Overall, hybrid and emerging generative models represent a **dynamic and rapidly advancing frontier**, combining the complementary strengths of classical VAEs, GANs, and diffusion models with novel neural architectures like transformers, and opening new possibilities for high-quality, controllable, and efficient image generation.

4. Results and Discussion

4.1 Comparison of Generative Models

Table 1: Comparison of Generative Models

Model	Architecture	Strengths	Limitations	Applications
GAN	Generator + Discriminator	High-quality images, versatile, realistic textures	Training instability, mode collapse, sensitive to hyperparameters	Image generation, super-resolution, style transfer, deepfakes
VAE	Encoder → Latent Space → Decoder	Stable training, interpretable latent space, probabilistic outputs	Blurry images, lower fidelity than GANs	Data augmentation, anomaly detection, medical imaging, compression
Diffusion Models	Iterative denoising process	High-fidelity images, stable training, handles complex distributions	Computationally expensive, slow sampling	Text-to-image (DALL·E, Stable Diffusion), scientific imaging, high-res synthesis

Table 1 summarizes the key characteristics, strengths, limitations, and applications of GANs, VAEs, and Diffusion Models. From the comparison, several observations emerge:

- Image Quality: If you want sharp, realistic images, GANs and Diffusion Models usually come out on top. VAEs? They're easier to train and their latent spaces actually make sense, but the images they spit out look a bit blurry. That's just the way they work—their reconstructions lean on randomness, so you never get that crisp finish.
- Training Stability: GANs can be a headache. They like to get stuck making the same few images over and over if you don't tune them just right. VAEs don't have that problem. They're way steadier, so a lot of people use them for things like anomaly detection. Diffusion Models take a lot of computing power, but they're reliable and pump out a wide variety of high-quality images.
- Computational Complexity: GANs need a fair amount of resources, but Diffusion Models are real resource hogs—those repeated denoising steps add up fast. If you're short on computing muscle, VAEs are your best bet. They're lightweight and get the job done.
- Application Suitability: People use GANs for all sorts of creative stuff: art, sharpening up images, even deepfakes. VAEs are great for things like making new training data, medical imaging, or anywhere you want to understand what's going on inside that latent space. Right now, Diffusion Models are leading the pack in high-resolution and text-to-image generation—think DALL·E 2 or Stable Diffusion.

4.2 Datasets and Evaluation Metrics

Table 2: Popular Datasets and Evaluation Metrics

Dataset	Type	Resolution / Samples	Use Case

CIFAR-10	Object images	32×32, 60,000	Benchmark for image generation
CelebA	Face images	178×218, 200,000+	Facial image synthesis and attribute generation
ImageNet	Object images	Variable, 14M+	Large-scale image synthesis, pretraining
LSUN	Scene images	High-resolution	Scene understanding and image generation
MNIST	Handwritten digits	28×28, 70,000	Simple image synthesis, proof-of-concept models

Table 2 presents the commonly used datasets and metrics for evaluating generative models. Key insights include:

1. **Dataset Selection**

- Small-scale datasets like CIFAR-10 and MNIST are ideal for proof-of-concept models.
- Large-scale datasets such as ImageNet and LSUN are essential for training high-capacity GANs or Diffusion Models.
- CelebA is particularly useful for facial image synthesis and attribute manipulation tasks.

2. **Evaluation Metrics**

- **FID** effectively measures the similarity between real and generated image distributions and is widely used to benchmark GANs.
- **Inception Score (IS)** evaluates both quality and diversity but may fail in capturing subtle artifacts.
- Precision and Recall complement FID and IS by measuring coverage and diversity.

The area of generative modeling has witnessed significant advancements in recent years, with distinct trends emerging across different model families. Diffusion Models are rapidly gaining prominence, particularly in tasks that demand high-quality and high-resolution image synthesis. Their iterative denoising process allows them to capture fine-grained details and realistic textures that are often challenging for traditional GANs. As a result, diffusion-based architectures are increasingly preferred for - text-to-image generation, scientific imaging, and high-fidelity artistic synthesis.

Hybrid models are starting to really shine when it comes to balancing the strengths and weaknesses of different approaches. Take VAE-GAN combos, for instance. VAEs give you stable, reliable representations, while GANs are known for those crisp, lifelike images. Put them together, and you get the best of both worlds—less blurriness than a plain VAE, fewer headaches with unstable training than a straight-up GAN. This blend works especially well when you need images that look good but also keep their structure intact.

Still, there's no shortage of challenges. Diffusion models, for one, eat up a ton of computing power, which puts them out of reach for a lot of folks who don't have access to beefy machines. So, researchers are really focused on finding ways to make training and running these models less demanding. Then there's the question of how we actually judge what these models make. Right now, the usual metrics just don't capture everything—like how real, diverse, or controllable the generated images are. And yeah, there's the ethical side, too. Deepfakes and synthetic content are popping up everywhere, and without better safeguards or ways to spot fakes, the risks keep growing.

In the end, picking the right model isn't just about chasing the highest quality. You have to think about what you need for the job, what kind of hardware you have, and how to use these tools responsibly. When researchers weigh the options carefully, they're in a better spot to get great results while keeping things practical—and ethical.

5. Conclusion

Generative models, including GANs, VAEs, and Diffusion Models, have fundamentally transformed image synthesis by enabling high-quality, diverse, and realistic image creation. Each model presents unique strengths: GANs excel in realism, VAEs offer stable and interpretable latent representations, and diffusion models achieve state-of-the-art fidelity. These models have demonstrated significant impact across computer vision, medical imaging, creative industries, and gaming, while also raising challenges in computational efficiency, evaluation, and ethical usage. Future research focusing on efficiency, interpretability, multimodal capabilities, and ethical safeguards promises to further enhance the applicability and societal benefits of generative models. Overall, generative modeling continues to be a dynamic and rapidly advancing field, offering immense opportunities for innovation and practical applications.

5.1 Applications, Challenges, and Future Directions

Generative models have revolutionized multiple domains by making possible the creation of synthetic data. Generative models have really changed the game in computer vision. People use them for things like image inpainting, super-resolution, and style transfer—basically, fixing up photos, making them sharper, or giving them a whole new look. Over in medical imaging, these models create synthetic MRIs, CT scans, and X-rays. That's a big deal for AI training, since it helps build bigger, more private datasets for diagnosing diseases. In art and design, they open up all kinds of creative possibilities, from digital paintings and illustrations to design prototypes. Artists and designers can get more done, and try out new ideas faster.

The entertainment and gaming industries have jumped in too. These models help build realistic characters, scenes, and even entire virtual worlds. That means less manual work—and less money spent—while still making everything look great. And in the bigger machine learning world, generative models boost performance by creating fresh data for training, which helps models learn better and handle more variety. Still, it's not all smooth sailing. Generative models run into real obstacles. In GANs, you get mode collapse, where the variety drops and everything starts to look the same. Diffusion models, on the other hand, are powerful but eat up a ton of computing power—they're slow and expensive to train. When it comes to judging quality, popular metrics like FID and Inception Score just don't tell the whole story. They often miss how realistic or diverse the images actually feel. There's also the problem of misuse—think deepfakes, privacy breaches, and bias sneaking in from the training data. If we want these systems to be reliable and accepted, we have to tackle these risks.

Researchers are pushing forward on a few fronts. They're working on making diffusion models faster and cheaper, without losing quality. Multimodal generative AI is on the rise—systems that can spit out images, videos, or audio from text prompts or other cues. That opens up a lot of new creative and practical uses. People also want generative models to be easier to explain, so we can actually understand why they produce what they do. And as these models get more powerful, ethical safeguards are becoming even more important—catching misuse, reducing bias, and keeping our data private. All of this is about making sure generative AI grows in a way that's responsible and trustworthy.

References

1. M. Klusch, J. Lässig, D. Müssig, et al., "Quantum Artificial Intelligence: A Brief Survey," KI - Künstliche Intelligenz, vol. 38, pp. 257-276, Nov. 2024. doi: 10.1007/s13218-024-00871-8.
2. E. Desdentado, C. Calero, M.Á. Moraga, et al., "Quantum computing software solutions, technologies,

evaluation and limitations: a systematic mapping study," Computing, vol. 107, art. 110, Apr. 2025. doi: 10.1007/s00607-025-01459-2.
3. G. Pilato and F. Vella, "A Survey on Quantum Computing for Recommendation Systems," Information, vol. 14, no. 1, art. 20, Dec. 2022. doi: 10.3390/info14010020.
4. M. Caleffi, A. Manzalini and A. S. Cacciapuoti, "Distributed quantum computing: A survey," Computer Networks, vol. 254, pp. 110672, Dec. 2024. doi: 10.1016/j.comnet.2024.110672.
5. H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen and S. Z. Li, "A survey on generative diffusion models," IEEE Trans. Knowl. Data Eng., 2024. doi: 10.1109/tkde.2024.3361474.
6. X. Zhang, X. Wei, W. Hu, J. Wu, Q. Li, Z. Zhang, Z. Lei, "A Survey on Personalized Content Synthesis with Diffusion Models," Mach. Intell. Res., vol. 22, pp. 817-848, Sep. 2025. doi: 10.1007/s11633-025-1563-3.
7. R. Jiang, G.-C. Zheng, T. Li, T. R. Yang, J.-D. Wang, X. Li, "A survey of multimodal controllable diffusion models," Journal of Computer Science and Technology, vol. 39, no. 3, pp. 509-541, Jul. 2024. doi: 10.1007/s11390-024-3814-0.
8. S. Fang, "A Survey of Data-Driven 2D Diffusion Models for Generating Images from Text," EAI Endorsed Transactions on AI & Robotics (AIRO), vol. 3, no. 1, Apr. 2024. doi: 10.4108/airo.5453.
9. D. Hemalatha, S. Dharshini, P. Ramesh, S. Priya, "AI Enabled MED Drone For Healthcare Application," Int. J. Innov. Res. Adv. Eng. (IJIRAE), vol. 11, no. 4, pp. 190-193, 2024. doi: 10.26562/ijirae.2024.v1104.01.
10. P. Mishra, P. Srivastav, P. Sharma, M. Atif, "AI-Based Medical Data Mining," Int. J. Innov. Res. Adv. Eng. (IJIRAE), vol. 11, no. 11, pp. 794-799, 2024. doi: 10.26562/ijirae.2024.v1111.01.