Deep Generative Models for Image Synthesis: GANs, VAEs, and Diffusion Approaches

Zhuolin Li[1], Xiangyu Liu[2] and Yibo Lu[3, *]

[1]Nanjing Foreign Language School FanShan Campus, Nanjing, 210000, China

[2]Xiamen JiuXi Senior High School, Xiamen, 361000, China

[3]Nanjing University of Posts and Telecommunications, Nanjing, 210000, China

*Corresponding author: P23000309@njupt.edu.cn

Abstract. This paper provides a comprehensive overview of the evolution and current state of deep generative models in image synthesis, with a focus on Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models. It identifies persistent challenges, including difficulty in generating high-resolution details, training instabilities, and limited controllability. The analysis demonstrates that Diffusion Models have emerged as the dominant paradigm, effectively overcoming previous limitations through their robust training dynamics and superior detail preservation. By modelling data distribution via gradual noise addition and removal, diffusion models achieve state-of-the-art image quality while maintaining structural coherence. Their integration with text-conditioning mechanisms enables precise semantic control, facilitating diverse applications such as high-fidelity text-to-image synthesis, medical image enhancement, and accelerating creative design. The review further examines practical implementations, including SRGAN for super-resolution, CVAE for image restoration, and diffusion-based text-to-image systems, highlighting their real-world impact across healthcare, media, and scientific imaging. The work concludes that diffusion models, particularly with latent-space optimisation, represent the most promising direction for future research, balancing unprecedented quality with increasing computational efficiency to enable broader deployment in critical domains.

Keywords: Image Synthesis; Image Super-Resolution; Image Enhancement and Restoration.

1. Introduction

In recent years, deep learning and neural networks have pushed image generation to new heights. At first, models could only create simple textures or low-quality images. Now they can make high-resolution images with rich details and realistic effects [1,2]. The use of these models has spread from art and games to virtual reality and medical image reconstruction [3-6].

However, there are still many problems. High-resolution images are still hard to restore with fine texture. Results often have blurred edges, broken structures, or artefacts [3-7]. Some models, like GANs, are unstable. They may collapse in training and lose diversity. In complex cases, results are hard to control and explain. Computation is also expensive. Diffusion models require multiple steps for inference. This results in a prolonged delay and limits real-time use [8].

To address these problems, numerous generative models have been proposed. GANs use a game between the generator and the discriminator. They can make images with rich detail and natural structure. But training is very sensitive to hyperparameters [1]. VAEs use probability modelling. They give reasonable control in the latent space and can do semantic interpolation. But results are often blurry [9,10]. Diffusion models employ a step-by-step noise reduction and denoising process. They give better detail and texture than GANs or VAEs. They are also more stable and can adapt to many inputs [8,11-13]. With faster sampling, their slow speed is also improving.

These models are essential. They improve both efficiency and quality of digital content creation. They bring new tools to art, VR, film, and medical imaging [14-18]. They also help with data augmentation and cross-modal tasks. They can give more training samples to other models. Compared to old methods, deep generative models have stronger capabilities, greater flexibility, and broader applications [19].

Looking at history, GANs were first introduced in 2014. They quickly gained popularity and developed numerous robust versions, including StyleGAN and CycleGAN [1]. VAEs improved in probability modelling and latent structure. They produced new versions, including CVAE, β-VAE, and VQ-VAE [9, 15-18]. Diffusion models became known around 2020. Works like DDPM, Stable Diffusion, and Imagen have demonstrated high-quality results in multimodal tasks [8, 11, 12].

This paper will study GANs, VAEs, and Diffusion models. It will explain their principles, strengths, limits, and uses. It will also compare their results and features. The focus will be on the advantages of diffusion models in quality and efficiency. Finally, it will give ideas on their future trend and research value.

2. Related Principles

2.1. Problems in Image Processing Technology

In the study and application of image generation and processing, several key problems remain that impact quality and practicality. The first challenge is the scarcity of high-quality, realistic images. Although deep generative models have made progress in low- and medium-resolution tasks, they continue to struggle with restoring fine details and textures in high-resolution images. Problems such as blurry edges, distorted structures, and artefacts are common [1,2]. The second challenge is the instability in training, especially in models represented by Generative Adversarial Networks (GANs). Their optimisation follows a game process, which often leads to mode collapse, where the generated samples lack diversity and some modes are ignored [3-6]. A third problem is the lack of controllability and interpretability. Conditional generation models can use text, labels, or reference images to provide some control. However, under complex scenes or multimodal conditions, the output remains challenging to predict [9]. Additionally, the limitations of computing resources and inference efficiency must not be overlooked. Diffusion models require multiple iterations in inference. This makes the process costly and slow, which is unsuitable for applications that need real-time response [8].

2.2. Principles of Image Processing Technology

The theoretical basis of modern image generation comes from generative modelling. The main goal is to learn the probability distribution of training data and then generate new samples that follow this distribution. Different methods provide different approaches. Probabilistic generative modelling uses maximum likelihood estimation (MLE) or variational inference to approximate the real distribution. Adversarial learning, represented by GANs, improves generation quality by training a generator and a discriminator in a game process [3-6]. The encoder–decoder structure, represented by Variational Autoencoders (VAEs), maps input data to a latent distribution through an encoder and then reconstructs samples through a decoder, allowing for controllable generation and latent space interpolation [9,10]. The diffusion–reverse diffusion process, represented by diffusion models, adds Gaussian noise step by step to the data, then removes the noise step by step using a trained denoising network, and finally generates high-fidelity images [8, 11, 12]. These theories provide a solid foundation for GANs, VAEs, and diffusion models, and point to future directions in structure optimisation and multimodal fusion [13].

3. Related Technologies and Their Principles

3.1. Generative Adversarial Networks (GANs)

3.1.1 Working Principle and Advantages and Disadvantages

Generative Adversarial Networks (GANs) were proposed by Goodfellow et al. in 2014. The core consists of a generator and a discriminator. The two are optimised in a minimax game. The generator takes random noise and outputs fake images. The discriminator attempts to determine whether the

input is real or fake. During training, the generator improves its output to fool the discriminator, while the discriminator learns to detect fake samples [3-6]. GANs can generate high-resolution and detailed images. They are widely used in unconditional generation, conditional generation, and cross-domain transfer. Many variants, such as StyleGAN and CycleGAN, show strong performance in specific tasks [6]. However, GANs are very sensitive to hyperparameters and initial conditions. They are often unstable in training and may suffer mode collapse, which reduces the diversity of samples [5]. Their latent space is also hard to control, which limits smooth interpolation. Despite these challenges, GANs play a crucial role in computer vision and continue to drive the advancement of generation technology.

3.1.2 Case Studies

DCGAN first applied convolutional neural networks to both the generator and discriminator, and it also used batch normalisation to improve stability [3]. StyleGAN introduced a mapping network and layer-wise style control in the generator, which achieved high-quality and controllable human face generation [6]. CycleGAN used cycle consistency loss to realise unpaired cross-domain image translation, making breakthroughs in style transfer and image-to-image translation [6].

3.2. Variational Autoencoders (VAEs)

3.2.1 Working Principle and Advantages and Disadvantages

Variational Autoencoders (VAEs) were proposed by Kingma and Welling in 2013 [9]. They are probabilistic generative models that combine deep neural networks with Bayesian inference. The encoder maps input data to parameters of a latent distribution. A latent variable is sampled using the reparameterization trick. The decoder then reconstructs the image from this latent variable. Training maximises the Evidence Lower Bound (ELBO) to approximate the real data distribution. Compared with GANs, VAEs are more stable in training and converge faster. Their latent space is more interpretable, which supports interpolation and feature manipulation [10]. They can also be extended into Conditional VAEs (CVAEs) for controllable generation [10,15]. However, VAEs often produce blurry results and struggle to capture complex textures as effectively as GANs [18]. To address this, new variants such as CVAE, β-VAE, and VQ-VAE improve latent space modelling and image details [15-17]. These models also show potential in multimodal generation, feature disentanglement, and long-range dependency modelling [18].

3.2.2 Case Studies

CVAE introduces conditional information, such as class labels or text, into both the encoder and decoder, thereby enabling controllable generation [10,15]. Building on this, β-VAE modifies the ELBO by introducing a weight β to enhance disentanglement in the latent space, thereby facilitating the learning of separable semantic features [16,17]. Furthermore, VQ-VAE discretises the latent space through vector quantisation, enhancing the fidelity of image details. When combined with autoregressive models, it further improves the modelling of long-range dependencies [18].

4. Application

With the continuous development of the deep generation model. Its application scene has been extended to actual tasks, and it has made significant achievements in image enhancement, Cross-modal generation, and Image super-resolution.

4.1. The Application of Generative Adversarial Networks (GAN) in Image Super-Resolution

Image super-resolution refers to the process of reconstructing low-resolution images into high-resolution counterparts. It is a Hot research direction in computer vision. The core contribution of SRGAN (Super-Resolution Generative Adversarial Network) in the field of image super-resolution lies in overcoming the visual quality bottleneck of traditional methods through adversarial learning mechanisms [1]. Traditional techniques only optimise pixel-level errors, resulting in overly smooth generated images and a lack of realistic texture details. Although the objective indicators are good, a

certain difference remains when observed by the human eye compared to the actual image [2]. SRGAN innovatively introduced the generative adversarial framework. The generator takes low-resolution images as input and reconstructs high-resolution images through a deep residual network structure. The discriminator distinguishes between the generated images and the real ones, thereby prompting the generator to learn more realistic high-frequency texture information [1]. This design enables the model to no longer merely pursue pixel matching but to focus on the image quality that the human visual system truly perceives.

The key breakthrough lies in the design of the perception loss function. This loss function integrates content loss based on deep features and adversarial loss. By utilising the pre-trained VGG network to extract high-level semantic features from images, it ensures that the generated results are consistent with real pictures in terms of semantic structure. The content loss employs a pre-trained VGG network to extract high-level semantic features from images, ensuring that the generated results maintain semantic structural consistency with the ground truth images. The adversarial loss, guided by the discriminator's feedback, prompts the generator to produce realistic texture details. This combination not only enhances the resolution of the reconstruction results but also makes them subjectively closer to real photos, solving the problem of traditional methods being clear but distorted. For instance, when dealing with facial images, SRGAN can generate natural skin textures and eyelash details, rather than the blurry and smooth areas produced by traditional methods [1].

In practical applications, SRGAN has demonstrated significant value. In the field of cultural heritage protection, it can transform blurry historical photos into clear images, thereby recreating precious historical moments [3]. In medical diagnosis, enhancing the details of CT or MRI images helps doctors detect minute lesions [4]. The security surveillance system utilises it to improve the recognition rate of faces and license plates in low-quality videos [5]. Streaming media services offer users a higher-quality video experience through this technology, even in situations with limited bandwidth [6]. In addition, in specialised fields such as satellite remote sensing and microscopic imaging, SRGAN can also effectively enhance the analysability of images [7].

4.2. The Application of Variational Autoencoder (VAE) in Image Enhancement and Restoration

Image enhancement and restoration, as an essential task in the field of computer vision, aims fundamentally to optimise the quality and integrity of images through algorithmic means to meet the visual analysis requirements of specific application scenarios. The Conditional Variational Autoencoder (CVAE) is a significant model in this direction, building upon the VAE. CVAE constructs a generative model for image enhancement and restoration tasks by embedding external conditional variables into the variational inference framework [9]. The core idea lies in constraining the latent space with conditional information, enabling the generation process to achieve targeted inference based on specific contexts [10]. In image restoration tasks, CVAE takes the known image regions as conditional inputs and achieves probabilistic reconstruction of the missing areas by modelling the parametric representation of the conditional posterior distribution, effectively avoiding the structural breaks and semantic incoherence problems caused by traditional interpolation methods [14]. To further enhance the quality of generation, researchers have proposed a CVAE-GAN hybrid architecture. By introducing an adversarial training mechanism, the detail generation capability is optimised, making the repaired area more consistent with the real data distribution in terms of texture continuity, edge sharpness, and global consistency [15]. In specific application scenarios, CVAE demonstrates high adaptability. For the spectral selective attenuation characteristics of underwater images, the model achieves colour correction and contrast enhancement by learning the mapping relationship between environmental conditions and image quality [16]. In low-light image processing, CVAE effectively suppresses noise and restores detail levels by jointly modelling the distribution characteristics of lighting conditions and clear images [17]. The advantage of CVAE lies in its organic integration of prior knowledge and data-driven features. The introduction of conditional variables not only enhances the model's ability to represent the target distribution but also ensures the diversity and

rationality of the generated results through the probabilistic framework. The advantage of CVAE lies in its organic integration of prior knowledge and data-driven features. The introduction of conditional variables not only enhances the model's ability to represent the target distribution but also ensures the diversity and rationality of the generated results through the probabilistic framework [14]. Compared to traditional autoencoders, CVAE establishes a more stable conditional relationship in the feature space, enabling the generation of enhanced images that significantly improve visual quality while maintaining content consistency [15]. In practical applications, this technology has been successfully deployed in multiple fields: in remote sensing image processing, it is used to repair areas obscured by clouds to improve the accuracy of surface analysis; in the medical imaging field, it is employed to eliminate motion artefacts in CT scans to assist in locating lesions [18]. These applications have verified the engineering value of CVAE in complex image degradation problems and demonstrated the technical advantages of the approach that combines generative models with domain knowledge in depth.

4.3. The application of Diffusion Models in text-to-image generation

The diffusion model represents a significant breakthrough in the field of text-to-image generation in the realm of generative artificial intelligence. Its core concept stems from the progressive modelling of data distribution [8]. This method achieves the controllable generation of high-quality images by constructing a learnable mapping from a simple noise distribution to a complex data distribution. In the text-to-image task, the diffusion model transforms the image generation process into a sequential denoising Markov chain, where the initial state is pure Gaussian noise, and the final state is the target image. This process consists of two key stages: the forward diffusion process gradually adds noise to transform the original image into a standard normal distribution; the reverse generation process learns the conditional probability distribution for restoring the image from the noise, by using neural networks to predict the noise components at each step and gradually removing them [8].

The introduction of text conditions is the key to achieving semantically controllable generation. By integrating the semantic features extracted by the text encoder as conditional signals into the reverse generation process, the model can establish the semantic correspondence between the text description and the visual content. This conditional mechanism is typically implemented using the cross-attention mechanism, allowing each spatial position in the generation process to focus on the relevant semantic information of the text, thereby ensuring a high degree of semantic consistency between the generated image and the text description [11]. The proposal of the potential diffusion model has further enhanced the practical value of this technology. By performing the diffusion process in the compressed potential space rather than the original pixel space, it significantly reduces computational complexity while maintaining generation quality [12]. This method not only keeps the high level of fidelity of the generated images but also achieves a good balance between computational efficiency and generation quality, laying the foundation for large-scale applications.

The successful application of diffusion models in text-to-image generation demonstrates their unique advantages in multimodal learning, enabling them to effectively capture the complex correlations between language descriptions and visual representations [13]. Its progressive generation feature allows the model to produce detailed and structurally sound images, avoiding the common mode collapse problem that occurs in traditional generation models. This technology has been widely applied in the field of creative design, enabling designers to transform abstract concepts into visual sketches quickly; the advertising industry utilizes it to generate diverse marketing materials efficiently; the education sector employs it to visualize complex concepts to assist in teaching; in game development, it accelerates the conceptual design of characters and scenes; in film and television production, it is used for concept art and pre-visualization; and it has reverse applications for visually impaired individuals to provide image descriptions, etc [19]. These applications fully demonstrate the powerful ability of the diffusion model in connecting language and visual modalities, providing an essential paradigm for the development of multimodal artificial intelligence.

5. Conclusion

The paper examined three primary models for image generation: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models. The paper explained how they work, their strengths and weaknesses, and where they can be applied. GANs can produce sharp and realistic images and are helpful in tasks such as image super-resolution; however, they are challenging to train and can collapse. VAEs are easier to train and have a clear latent space, which makes them suitable for controllable generation and image repair; however, the images they produce often appear blurry. Diffusion models are newer and stronger. By adding and removing noise step by step, they create images with fine details and natural textures. They also work well with text-to-image tasks. However, they require a significant amount of time and computing power.

The paper also discussed some real-world applications. GANs are used in super-resolution, where they can rebuild missing details in low-quality pictures. VAEs are used for enhancing and repairing images, such as restoring old photos or filling in missing parts in medical images. Diffusion models are now very popular for text-to-image generation. With only a text description, they can create new pictures for design, games, or even science research. These examples demonstrate the power and utility of generative models in various fields.

Although progress is evident, numerous problems remain. It is still challenging to maintain every detail sharp in extremely high-resolution images. GANs are unstable during training, and VAEs often produce blurry results. Diffusion models are stable but very slow, requiring strong hardware. Additionally, controllability remains weak, and the models sometimes fail to yield results that align with user expectations.

In the future, research should try to make these models faster and lighter. It is also essential to make generation more controllable and easier to guide. Models should be able to combine various types of input, including text, images, and video, to facilitate practical analysis and decision-making. At the same time, people need to pay more attention to ethics, such as avoiding bias, fake content, or harmful use. If these problems are solved, image generation will be even more helpful in art, industry, and daily life.

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Ledig C, Theis L, Huszár F, et al. Photo-realistic single-image super-resolution using a generative adversarial network. IEEE Trans Pattern Anal Mach Intell, 2017, 40(2): 4681–4690.

[2] Dong C, Loy C C, He K, et al. Learning a deep convolutional network for image super-resolution. IEEE Trans Image Process, 2016, 25(2): 614–627.

[3] Yu Y, Ma X, Zheng Y, et al. Deep learning for historical document image super-resolution. Int J Doc Anal Recognit, 2020, 23(4): 395–411.

[4] Baur C, Albarqouni S, Navab N. Super-resolution for medical imaging: A survey. IEEE Trans Med Imaging, 2021, 40(9): 2449–2464.

[5] Liu J, Huang Y, Zhang W, et al. Real-time face super-resolution via generative adversarial networks for surveillance applications. IEEE Trans Circuits Syst Video Technol, 2019, 29(12): 3587–3600.

[6] Wang X, Yu K, Wu S, et al. ESRGAN: Enhanced super-resolution generative adversarial networks. Neurocomputing, 2020, 389: 107–119.

[7] Huang S, Chen Y, Lu W, et al. Remote sensing image super-resolution using generative adversarial networks. ISPRS J Photogramm Remote Sens, 2020, 162: 155–166.

[8] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. J Mach Learn Res, 2021, 22(1): 1–45.

[9] Kingma D P, Welling M. Auto-encoding variational bayes. Found Trends Mach Learn, 2019, 12(4): 307–392.

[10] Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding. Nat Mach Intell, 2023, 5: 545–556.

[11] Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. Neural Comput Appl, 2018, 29(5): 1195–1204.

[12] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. IEEE Trans Pattern Anal Mach Intell, 2023, 45(1): 10684–10695.

[13] Yu L, Zhao W, Zhang Z, et al. Multimodal learning with diffusion models: A comprehensive survey. Inf Fusion, 2024, 98: 101933.

[14] Iizuka S, Simo-Serra E, Ishikawa H. Globally and locally consistent image completion. ACM Trans Graph, 2017, 36(4): 1–14.

[15] Bao P, Liao J, Song Q, et al. CVAE-GAN: Fine-grained image generation through asymmetric training. Pattern Recognit, 2018, 79: 389–401.

[16] Liu Y, Wang C, Zhang H, et al. Underwater image enhancement via conditional variational autoencoders with multi-scale feature fusion. IEEE Trans Image Process, 2022, 31: 2985–2997.

[17] Wang R, Huang Z, Li Y, et al. Low-light image enhancement via conditional variational autoencoders with illumination-aware constraints. IEEE Trans Image Process, 2020, 29: 6872–6885.

[18] Wang R, Huang Z, Li Y, et al. Medical image restoration via conditional variational autoencoders with attention mechanisms. IEEE Trans Med Imaging, 2021, 40(8): 2176–2187.

[19] Chen X, Liu Y, Feng R, et al. Real-world applications of text-to-image generation: Case studies and lessons learned. ACM Trans Comput Hum Interact, 2023, 30(2): 1–24.