# A Short Survey of AIGC in Computer Vision

|  |  |
| --- | --- |
| 杜思楠: | Concept Learning调研和写作 |
| 王渊睿: | Basic Generative Models调研和写作 |
| 孙前普: | Image Generation调研和写作 |
| 夏 瑞: | Introduction调研和写作、文章的组织排版 |

## 1 Introduction

Recently, Artificially Generated Content (AIGC) has garnered considerable attention from both the computer science community and society at large. Innovations in generative AI technology by major tech companies have led to the development of remarkable models like ChatGPT, DALL·E-2, and Codex. AIGC pertains to the application of generative AI algorithms to create specific content as directed by human instructions, in contrast to content produced through manual human effort. It can autonomously generate original content quickly by drawing upon the extensive knowledge databases acquired from vast datasets produced by humans. AIGC effectively diminishes the barriers to content creation and significantly enhances productivity across various domains, including advertising, education, and artistic endeavors. For instance, ChatGPT is an auto-regressive, large language model, trained on extensive corpora using reinforcement learning algorithms that incorporate human feedback. This training aligns the model's understanding with human intent, enabling it to engage in a wide array of interactions involving human text. In parallel, models like DALL·E-2 and Stable Diffusion possess the ability to generate high-quality images that can be considered masterpieces, based on descriptions provided by humans. This capability effectively dispels any skepticism regarding AI's potential to rival human-level artistic creativity. As depicted in Figure 1, generative AI is not confined to single-modality creation but can also facilitate cross-modality interactions and content generation. Nevertheless, due to the constraints imposed by the article's scope and the boundaries of my own expertise, our discussion on AIGC is limited exclusively to the domain of computer vision.

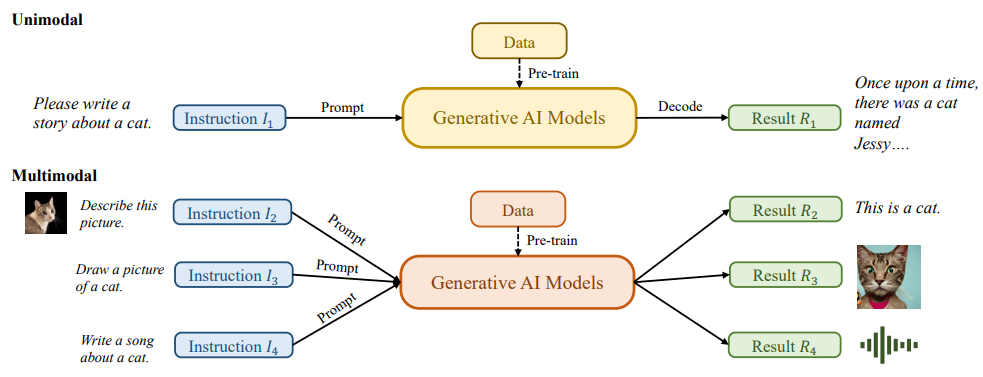


Fig. 1. Generative AI is typically categorized into two types: unimodal models and multimodal models. Unimodal models receive instructions from the same modality as the generated content, such as large language models (LLM) and image translation. In contrast, multimodal models can receive instructions from multiple modalities and generate results in various modalities, exemplified by text-to-image models (T2I) and visual question answering models (VQA).

## 2 Organization

The remaining sections of the survey are structured as follows. In Section 3, we provide an overview of the evolution of generative models in the field of computer vision, covering foundational concepts from GANs to the latest diffusion models. Section 4 delves into the diverse applications of generative AI across various domains, offering a comprehensive introduction to each area through an examination of related works.

## 3 Basic Generative Models

Image generation, a subset of generative modeling in deep learning, has witnessed significant advancements over the years, driven by models with increasing sophistication, performance, and capability. This review aims to provide an exhaustive overview of the current landscape of image generation research, focusing on major generative models like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Vector Quantized Variational Autoencoders (VQ-VAEs), Autoregressive Models, and Diffusion Models. We delve into influential papers for each model, summarizing the methodologies, challenges addressed, and contributions to the field. Finally, we encapsulate the current state and future prospects in image generation.

Image generation encompasses techniques that allow computers to create new images that appear similar to those in a given dataset. The field has seen extensive research and rapid advancements, particularly with deep learning models. The following sections discuss major categories of image generation models and their seminal research works.

GANs, a framework comprising two neural networks, the generator and the discriminator, competing against each other. This fundamental work laid the foundation for various GAN-based models. DCGAN use convolutional layers in both generator and discriminatorm, providing architectural norms that greatly stabilized the training of GANs. StyleGAN introduces a style-based generator enhancing the control over the generated images' features. This model significantly improved the quality of generated images. BigGAN demonstrated that scaling GANs by increasing the model size, batch size, and dataset diversity leads to an improvement in image quality. BigGAN set new records for the realism of generated images.

VAEs incorporate a variational approach within the encoder-decoder framework for generative modeling. They provided a scalable and deep generative model. β-VAE introduces a hyperparameter β to balance the latent channel capacity and the independence constraint, improving disentanglement in representations. VQ-VAE use vector quantization to produce discrete latent representations, enabling generative modeling of complex data distributions. Razavi et al. extended VQ-VAE to VQ-VAE-2, improving the hierarchical organization of the VQ-VAE for higher-fidelity images.

Autoregressive models are a family of generative models that decompose the joint distribution of an image into a product of conditional distributions. Unlike GANs and VAEs which map a latent vector to an image in one shot, autoregressive models generate images pixel-by-pixel or patch-by-patch, conditioning each step on previously generated portions of the image. The sequential nature of autoregressive models allows them to capture complex dependencies between pixels and model images with very high fidelity. PixelRNN uses an LSTM-based recurrent neural network to generate images pixel by pixel, capturing the full distribution of the data. PixelCNN, which utilizes a convolutional architecture to model the pixel distribution, improving computational efficiency compared to PixelRNN. Sparse Transformer use a sparse attention mechanism, offering improved performance and efficiency in autoregressive models for high-resolution image generation. Image Transformer adapted the Transformer model for image generation, capturing long-range dependencies with reduced computational complexity.

Diffusion models are a new class of deep generative models that have emerged as a powerful approach for unconditional image synthesis. The key idea is to model image generation as a process of stochastically adding noise to data, and then reversing this process. Specifically, diffusion models are trained to remove added noise from data by estimating conditional distributions over latent variables that represent gradual corruptions of the data. By composing a sequence of conditional denoising steps, high quality samples can be obtained from pure noise. DDPM uses a diffusion process to model the data distribution, demonstrating high-fidelity image generation. DDIM offers faster sampling and better quality images through modified training and sampling procedures. SDE(Song et al.) leveraged stochastic differential equations for score-based generative modeling, enhancing the flexibility and efficiency of diffusion models. Guided Diffusion is a conditioning framework for diffusion models, producing high-resolution, photorealistic images.

The field of image generation has achieved remarkable progress, with models generating ever more realistic images. GANs have emerged as prominent for their quality of generation, VAEs for their generative ability and understanding of latent spaces, Autoregressive models for capturing pixel-level dependencies, and Diffusion models for their principled approach to noise reduction and image creation.

Current research trends indicate a strong inclination towards improving image fidelity, resolution, and diversity of generative models while reducing computational costs and environmental impact. Furthermore, ensuring ethical use of these technologies remains a paramount consideration. Looking forward, the integration of multimodal learning, attention mechanisms, and understanding of 3D spaces appear as promising directions. The ultimate goal remains the same: to enhance the synergy between human creativity and AI's computational prowess in various domains, including art, design, healthcare, and entertainment.

## 4 Applications

### 4.1 Image Generation

GANs consist of two neural networks: the Generator and the Discriminator. GANs function by engaging these networks in an adversarial training process with the primary objective of generating realistic images. The Generator takes random noise or other input data and strives to produce images closely resembling genuine ones. In contrast, the Discriminator evaluates images, distinguishing between those generated by the Generator and authentic ones. CGANs are a specialized GAN variant known for integrating conditional information into both the generator and discriminator networks. This additional information, often in the form of labels or class data, plays a pivotal role in guiding the image generation process. DCGANs enhance GAN stability and image quality by introducing deep convolutional neural networks into the GAN architecture. These networks are optimized for image synthesis tasks, consistently delivering superior results in generating realistic and intricate images. CycleGAN represents a robust GAN variant designed for unpaired image-to-image translation, notably enabling the transformation of images between domains without requiring a directly paired dataset for training. This innovative approach relies on cycle-consistency loss, ensuring the seamless reversibility of translated images. StyleGAN stands as a cutting-edge GAN architecture celebrated for its extraordinary mastery over the visual style of generated images. It empowers the fine-tuned manipulation of diverse elements within an image's style, encompassing attributes such as color, texture, and level of detail. In the context of image generation tasks, the Generator commonly accepts a random vector as input and progressively generates images that exhibit a high degree of realism. This unique capability positions GANs as an ideal choice for the creative generation of images. Following the training process, the Generator can process a variety of inputs, such as noise or textual descriptions, enabling it to produce images that adhere to specific criteria. This versatility proves to be particularly valuable in applications such as face generation, artistic content creation, and style transfer. The training of GANs frequently necessitates meticulous fine-tuning and extended training durations due to their susceptibility to issues of training instability and mode collapse. On occasion, the generator may become trapped in a repetitive cycle, resulting in the production of similar images, thereby leading to a deficit in image diversity.

VAEs are renowned for their capability to learn a compact and continuous representation of complex data, often used for tasks like image and text generation, data compression, and feature learning. VAEs employ probabilistic encoders and decoders to map data into a latent space and generate new data from this space, making them a powerful tool in generative modeling and data analysis. VQVAE, a novel family of models that merges VAEs with vector quantization to acquire a discrete latent representation. It has been demonstrated that VQVAEs exhibit the capacity to model extended, long-term dependencies by leveraging their compact discrete latent space. VQVAE-2 enhances the original model by introducing hierarchical structures, enabling more efficient modeling of complex data. DALL·E and CogView both adopt a two-stage training methodology to accomplish text-guided image generation. In the initial stage, VQVAE is trained to compress images, thereby mitigating computational overhead. In the subsequent stage, a Transformer decoder is leveraged to unify images and text into tokens for auto-regressive training. VQGAN utilizes a convolutional VQVAE to acquire knowledge of a codebook comprising context-rich visual components. This composition of components is further delineated through an autoregressive transformer architecture. The integration of a discrete codebook serves as the intermediary that connects these architectural components. Furthermore, the inclusion of a patch-based discriminator facilitates significant compression without sacrificing perceptual quality. This method harmoniously amalgamates the efficiency of convolutional methodologies with transformer-based high-resolution image synthesis.

Diffusion models draw inspiration from the principles of non-equilibrium thermodynamics. They establish a Markov chain of diffusion steps that progressively introduce random noise to data. Subsequently, they acquire the ability to reverse this diffusion process, enabling the generation of desired data samples from the noise. Notably, unlike VAEs, diffusion models are trained using a fixed procedure, and the latent variable maintains high dimensionality, equivalent to that of the original data. ADM proposes a gradient-guided method using a classifier to improve diffusion model sampling, and with the classifier gradient scale, it can be tuned to balance diversity and fidelity. ADM beats GANs and achieves state-of-the-art performance on unconditional and conditional image synthesis. CFGDM proves that it is still possible to run conditional diffusion steps by incorporating the scores from a conditional and an unconditional diffusion model without an independent classifier. Inspired by the capacity of guided diffusion models to produce highly realistic images and the proficiency of text-to-image models in handling open-ended prompts, GLIDE have applied guided diffusion to the realm of text-conditional image synthesis, which employs a text encoder to incorporate natural language descriptions as conditional inputs and train a diffusion model with 3.5 billion parameters to synthesize image. DALL·E-2's operation can be simplified into a three-step process. First, a given text prompt undergoes encoding using a specialized text encoder, resulting in a representation in a latent space. Next, a component called the "prior" transforms this text encoding into an image encoding, effectively capturing the underlying semantic information from the text. Finally, employing an image decoder, DALL·E-2 stochastically generates an image that visually embodies the conveyed semantic information, thus providing a visual manifestation of the initial text prompt. CogView-2 propose to utilize hierarchical transformers and local parallel autoregressive generation, which involves pretraining a 6B-parameter transformer with a flexible self-supervised task, the cross-modal general language model, and fine-tuning it for efficient super-resolution. Latent diffusion applies the diffusion process within a lower dimensional latent space to reduce memory and computational complexity. Incorporating cross-attention layers within the model's architecture enhances diffusion models, making them more robust and versatile in handling diverse conditioning inputs, including text and bounding boxes. This enhancement enables efficient high-resolution synthesis using a convolutional approach. Imagen's discovery highlights the effectiveness of large frozen language models, specifically T5 models trained solely on text data, as text encoders for text-to-image generation. Additionally, scaling the size of the frozen text encoder significantly improves sample quality more than scaling the size of the image diffusion model.

### 4.2 Concept Learning

Recently developed large-scale text-to-image models have demonstrated remarkable capabilities, enabling the high-quality and diverse synthesis of images based on natural language text prompts. These models offer a significant advantage in the form of a robust semantic prior, acquired from an extensive dataset of image-caption pairs. This prior, for example, learns to associate the word "dog" with various instances of dogs, accommodating different poses and contextual variations within images. While these models excel in image synthesis, they currently lack the capacity to replicate the specific appearance of subjects from a provided reference set or generate novel interpretations of these subjects within distinct contexts.

Textual Inversion is a technique introduced in this context, aimed at representing visual concepts, such as objects or styles. It achieves this by creating fresh tokens within the embedding space of a fixed text-to-image model. This process leads to the generation of compact, personalized token embeddings. However, it's important to note that the method's expressive capacity is constrained by the immutability of the frozen diffusion model. DreamArtist utilizes a learning strategy known as positive-negative prompt-tuning to effectively balance the trade-off between preserving specified characteristics from the reference and ensuring precise control over the generation process. DreamBooth presents an innovative method for personalizing text-to-image diffusion models. By providing only a few subject images as input, a pre-trained text-to-image model is fine-tuned to establish a distinct association between the model and the subject. Once this integration is achieved within the model's output domain, it facilitates the generation of new, exceptionally realistic images of the subject, situated in diverse contextual scenarios. This capability is harnessed by leveraging the model's embedded semantic knowledge and applying a novel self-generated class-specific prior preservation loss. LoRA employs a strategy in which it locks the pretrained model weights and introduces rank decomposition matrices that can be fine-tuned in each layer of the Transformer architecture. This approach leads to a substantial reduction in the number of parameters that can be modified for downstream tasks. Custom Diffusion discovers that it is possible to achieve a high level of representational power for new concepts by optimizing only a limited set of parameters within the text-to-image conditioning mechanism. This approach not only enables rapid tuning, taking approximately six minutes. Moreover, they found that it is feasible to conduct joint training for multiple concepts or amalgamate multiple fine-tuned models into a unified model through a closed-form constrained optimization process. The approach employed by SVDiff involves fine-tuning the singular values within the weight matrices. This fine-tuning process results in a more compact and efficient parameter space, which helps reduce the risk of overfitting and language drift. Furthermore, SVDiff generates training data using the cut-mix technique and incorporates regularization penalties to limit the occurrence of multiple subject attention maps. ELITE introduces global and local mapping training techniques, in which a global mapping network is first trained to encode a concept image into multiple textual word embeddings, with one primary word for well-editable concept and other auxiliary words to exclude irrelevant disturbances and a local mapping network is further trained, which projects the foreground object into textual feature space to provide local details. Nevertheless, it's important to acknowledge that, owing to specific limitations in the model's architecture, the text alignment effect is only moderately achieved. The Cones framework introduces an innovative method featuring the utilization of concept neurons, which are specifically updated in the K and V layers of cross-attention for individual subjects. Moreover, in Cones 2, a technique is developed to generate composite images that incorporate multiple subjects. This is achieved by learning the residual of token embedding and exercising control over the attention map. The InstructPix2Pix system is capable of performing a variety of editing tasks based on human instructions, including object replacement, style changes, environment modifications, and artistic medium variations. This functionality is achieved by incorporating the latent features of reference images into the model's noise injection process. UMM-Diffusion presents an innovative concept known as Unified Multi-Modal Latent Diffusion. This approach processes input sequences that combine both text and images, with an emphasis on specific subjects, in order to generate customized images featuring these subjects. However, it's crucial to acknowledge certain limitations associated with this method. For example, it lacks support for multiple subjects, and its training data is sourced from LAION-400M, resulting in suboptimal performance when generating images related to less common or rare themes.