Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/79/20241286

A survey of generative models used in text-to-image

Jingjing Xu[1,4,*,†], Jiahao Du[2,5,†], Junyi Wang[3,6]

[1]School of Business, Zhejiang Wanli University, Ningbo, Zhejiang, China
[2]School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China
[3]Shenzhen College of International Education, Shenzhen, Guangdong, China

[4]tsujaejing03914@163.com
[5]Jiahao.du21@xjtlu.edu.cn
[6]s21213.wang@stu.scie.com.cn
*corresponding author
[†]These authors contributed equally to this work and should be considered co-first authors.

Abstract. The emergence and rapid development of neural networks have been pivotal in advancing text-to-image generative models, with particular emphasis on generative adversarial networks (GANs), variational autoencoders (VAEs), and augmented reality (AR). These models have greatly enriched the field, offering diverse avenues for image generation. Critical support has been provided by databases such as MS COCO, Flickr30K, Visual Genome, and Conceptual Captions, along with essential evaluation metrics, including Inception Score (IS), Fréchet Inception Distance (FID), precision, and recall. In this comprehensive review, we delve into the mechanisms and significance of each model and technique, ensuring a holistic examination of their contributions. Both GANs and VAEs stand out as significant models within image generative frameworks, each excelling in distinct aspects. Therefore, it is imperative to discuss both models in this review, as they offer complementary strengths. Additionally, we include noteworthy models such as augmented reality to provide a well-rounded assessment of the current advancements in the field. In terms of datasets, MS COCO offers a diverse and extensive collection of images, serving as a cornerstone for model training. Other datasets like Flickr 30k, Visual Genome, and Conceptual Captions contribute valuable labeled examples, further enriching the learning process for these models. The incorporation of widely recognized metrics and methodologies in the field allows for effective evaluation and comparison of their relative significance. In conclusion, the field's recent achievements owe much to the integration of its various components. VAEs and GANs, with their unique strengths, complement each other, while metrics and datasets play complementary roles in advancing the capabilities of generative models in the context of text-to-image synthesis. This survey underscores the collaborative synergy between models, metrics, and datasets, propelling the field toward new horizons.

Keywords: Generative Models, Text-to-Image, Variational Autoencoders.

DOI: 10.54254/2755-2721/79/20241286

1. Introduction

Text-to-image models are machine learning models that generate visual content based on textual descriptions. These models typically include a language and a generative model. In 2014, Ian Goodfellow and his colleagues put forward the concept of GAN, which simultaneously trains two neural networks: a generator and a discriminator. This design significantly promoted the research in generative models, including the text-to-image model, and around 2015, the first text-to-image model was introduced [1]. The images generated by it were low-resolution, and the given prompts were barely discernible from the pictures. However, it can create images that do not exist in the training data. Currently, commercialized text-to-image models are approachable to the vast majority, including those without a computer science prerequisite.

Variational Auto-encoders (VAEs) and Augmented Reality (AR) are prominent in text-to-image generative models, with promising applications across many sectors. VAEs are particularly noted for their ability to learn complex data distributions and generate new, distinctive outputs, marking them a significant contributor to the advancement of text-to-image models. AR, on the other hand, overlays digital information onto real-world environments, improves user interaction and perception, and aids time series analysis to predict future trends. These capabilities are crucial for enhancing visualization and aiding in more informed decision-making processes.

2. Technical Principles

2.1. Image Sources for Text-to-Image Models

2.1.1. Microsoft Common Objects in Context (MS COCO). A benchmark in the field of computer vision and object identification, Microsoft Common Objects in Context (MS COCO) is a sizable image recognition, segmentation, and captioning dataset [2]. With over 330,000 diverse and complex images, it provides a comprehensive collection for training and evaluating various computer vision tasks [3]. One of the most complete datasets for object detection and segmentation tasks, it contains more than 2.5 million object instances classified with 80 different object types. It includes annotations such as bounding boxes, segmentation masks, and key point annotations for certain object categories, enabling the development and evaluation of models for tasks like object detection, instance segmentation, and pose estimation. In addition to the object annotations, MS COCO also provides textual descriptions or captions for a subset of images. These captions capture the semantic meaning and contextual information of the images, making the dataset suitable for tasks like image captioning and text-to-image synthesis. To facilitate fair comparison and evaluation, the MS COCO dataset defines several evaluation metrics, including mean Average Precision (mAP) for object detection and instance segmentation, as well as BLEU and METEOR scores for image captioning. These metrics allow researchers to quantitatively assess the performance of their models and compare them with state-of-the-art approaches. Figure 1 shows the structure of technical principles.

DOI: 10.54254/2755-2721/79/20241286

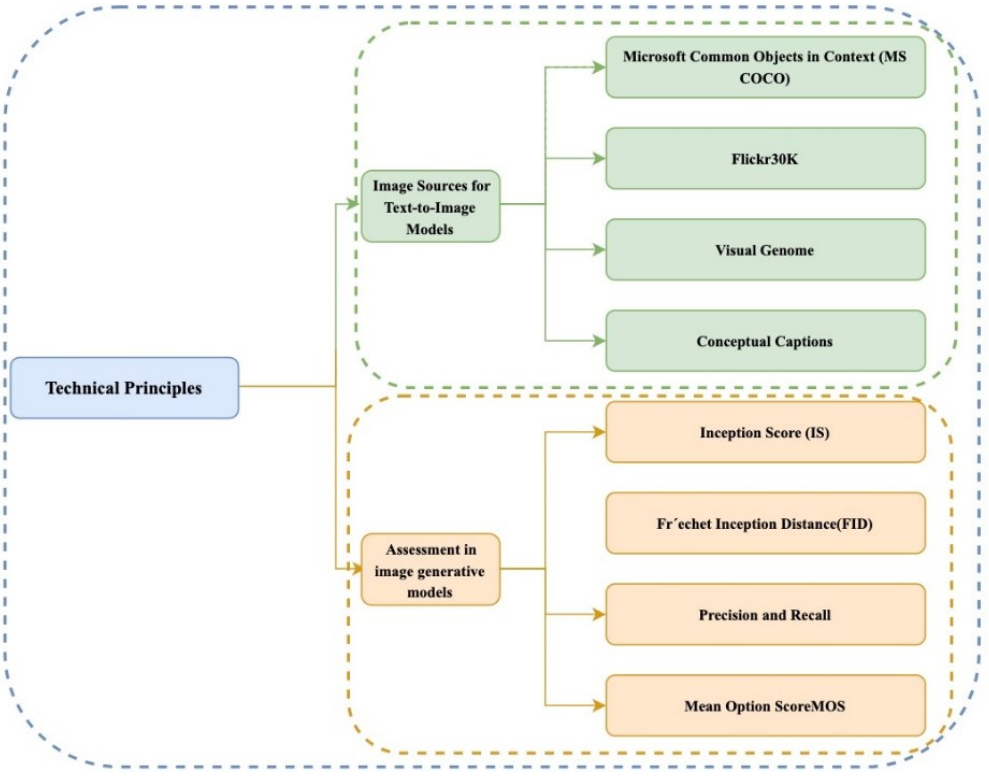

Figure 1. The Structure of Technical Principles

2.1.2. Flickr30K. The Flickr30K dataset is widely recognized and utilized within the field of computer vision for a variety of multimodal research endeavours [4], including image captioning, visual question answering, and multimodal representation learning. Comprising a total of 31,783 images sourced from the Flickr website, each accompanied by five English captions, this dataset serves as an invaluable resource for the examination of the intricate relationship between images and textual descriptions. The extensive utilization of the Flickr30K dataset in numerous studies has significantly contributed to the advancement of multimodal research. It has effectively established itself as a benchmark for evaluating the performance of diverse models and algorithms in tasks pertaining to image comprehension and natural language processing.

A notable attribute of the Flickr30K dataset lies in the provision of multiple captions for each image. This unique characteristic allows researchers to delve into the realm of diverse textual descriptions associated with the same visual content, thereby facilitating an exploration of the nuances inherent in language and image comprehension. Furthermore, the dataset's substantial size and wide-ranging assortment of images render it highly suitable for the training and evaluation of models employing real-world data.

2.1.3. Visual Genome. The Visual Genome dataset has been widely recognized and utilized in the field of computer vision for various research endeavours, including object recognition, image captioning, and visual relationship understanding [5]. This collection is made up of over 100,000 web-sourced photos, each meticulously annotated with detailed scene graphs.

The annotations in the Visual Genome dataset capture objects, attributes, and relationships within each image, providing a structured representation of the visual content and its semantic relationships. These scene graphs offer valuable insights into the intricate details of the images, enabling researchers to explore and comprehend visual relationships at a granular level.

The availability of the Visual Genome dataset and its comprehensive annotations has significantly contributed to the advancement of computer vision research. Researchers heavily rely on these

DOI: 10.54254/2755-2721/79/20241286

annotations to develop and evaluate models for tasks such as object recognition, attribute prediction, visual relationship detection, and scene understanding. The dataset serves as a benchmark for assessing the performance of different algorithms and approaches in these tasks, facilitating the comparison and advancement of computer vision models.

In conclusion, the Visual Genome dataset is a highly valuable resource in the field of computer vision, providing extensive annotations that enable researchers to explore and advance various tasks related to object recognition, image captioning, and visual relationship understanding.

2.1.4. Conceptual Captions. The Conceptual Captions dataset represents a vast compilation of image-caption pairs meticulously curated to aid developments in computer vision and natural language processing. This comprehensive collection encompasses a wide spectrum of images, each accompanied by descriptive captions, thereby serving as a highly valuable resource for various tasks such as image captioning, multimodal representation learning, and cross-modal retrieval [6].

The Conceptual Captions dataset, with an astonishing collection of roughly 3.3 million image-caption pairings, is one of the most extensive publicly available datasets for the purpose of comprehending images and generating languages. These images were painstakingly acquired from the internet and represent a wide spectrum of aesthetic themes and scenarios. The captions themselves are meticulously crafted through a combination of automated techniques and human validation, ensuring the provision of high-quality textual descriptions.

The Conceptual Captions dataset has been extensively utilized in the development and evaluation of image captioning models. Its substantial size and diverse array of images and captions provide researchers with a unique opportunity to explore the challenges associated with generating accurate and diverse textual descriptions for visual content. Moreover, this dataset has been effectively employed in cross-modal retrieval tasks, facilitating the retrieval of relevant images based on textual queries, or vice versa.

2.2. Assessment in Image Generative Models

2.2.1. Inception Score(IS). The Inception Score represents a widely recognized metric employed in the evaluation of the quality and diversity of generated images [7]. Its computation entails the determination of the conditional entropy derived from the predicted probability distribution of the generated images, utilizing the Inception-v3 classifier. This classifier enhances the precision of the assessment, thereby ensuring a more accurate evaluation. A higher Inception Score denotes a heightened level of quality and diversity exhibited by the generated images. The calculation of the Inception Score involves the classification of the generated images and the subsequent computation of the conditional entropy associated with the predicted labels. This measure effectively captures the uncertainty or diversity inherent in the predicted labels, considering the generated images. By encompassing both the quality and variety components of the created pictures, the Inception Score offers a comprehensive and meticulous evaluation of their overall performance. Consequently, it serves as an invaluable tool in the assessment of image generation.

2.2.2. Fr´echet Inception Distance(FID). Benchmarking is an integral component in the assessment of algorithmic effectiveness and performance in the disciplines of computer vision and image synthesis. In the realm of generative models, benchmarking assumes a critical role in the evaluation of image quality and diversity. One widely employed metric for this purpose is the Fr´echet Inception Distance (FID) [8], a metric of similarity between the distributions of produced and real pictures. Developed by Heusel et al. in 2017, the FID metric harnesses the capabilities of deep neural networks and statistical analysis to provide a quantitative evaluation of image quality. Its underlying principle is that a well-trained generative model should not only produce visually appealing images but also accurately capture the statistical properties inherent in real image distributions. The FID metric leverages feature representations derived from a pre-trained Inception-v3 network. FID evaluates the dissimilarity between

DOI: 10.54254/2755-2721/79/20241286

the two distributions by comparing the multivariate Gaussian distributions of these feature representations for both actual and produced pictures. A lower FID score signifies a closer match, indicating that the generated images closely resemble the statistical properties of real images. Benchmarking FID involves the evaluation of different generative models or algorithms utilizing a standardized dataset. This dataset typically comprises real images that serve as a reference distribution. The generated images produced by various models are then compared to the real image distribution using the FID metric. This benchmarking process provides valuable insights into the capabilities and limitations of different generative models, facilitating meaningful comparisons between various approaches.

2.2.3. Precision and Recall. Benchmarking precision and recall is a crucial aspect of evaluating the performance of models and algorithms in various information retrieval and classification tasks [9]. Precision and recall are widely used evaluation metrics that provide insights into the effectiveness and completeness of a system's output. Precision is the fraction of accurately detected positive cases among all positive instances projected. It is determined by dividing the number of true positives (TP) by the total number of true positives and false positives (FP):

Recall, in other words, counts the percentage of positive cases that were properly detected out of all the actual positive instances. The ratio of true positives to the sum of true positives and false negatives (FN) is used to calculate it: Benchmarking precision and recall involves comparing the performance of different models or algorithms on a standardized dataset or set of test cases. The dataset typically contains ground truth labels or annotations that serve as the reference for evaluating the model's predictions. To conduct a benchmark, a performance evaluation framework is established, where the models or algorithms are applied to the dataset, and their precision and recall values are computed. These values are then compared to identify the best-performing methods or to analyse the benefits and drawbacks of different strategies. Precision and recall are frequently combined in the context of information retrieval to evaluate the calibre of search results. A high precision indicates that the system retrieves a small number of irrelevant results, while a high recall suggests that the system retrieves a large proportion of the relevant results.

2.2.4. Mean Opinion Score(MOS). Mean Opinion Score (MOS) serves as a crucial benchmarking metric for evaluating the subjective quality of multimedia content, providing a standardized framework for performance assessment [10]. MOS benchmarking involves subjective tests where a diverse group of observers rates multimedia stimuli. The aggregated and averaged ratings result in the MOS, representing the overall perceived quality. This metric enables researchers to objectively compare the performance of multimedia processing algorithms and systems, aligning with human perception. The benchmarking process includes selecting a representative group of observers, designing a test protocol with stimuli and rating scales, and employing statistical analysis techniques to ensure result reliability. MOS benchmarking is widely applied in audio and speech processing for assessing codecs, enhancement algorithms, and noise reduction techniques. In video processing, it aids in evaluating codecs, quality enhancement algorithms, and streaming systems. Additionally, MOS finds application in assessing virtual reality (VR) and augmented reality (AR) systems, emphasizing the immersive experience's quality [10].

3. Model Introduction

We will cover the methodologies and strategies used in generative models for text-to-image generation in this part. These approaches have made major contributions to the area, improving the generated image quality and realism.

3.1. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) have brought about a revolution in text-based image generation. Consisting of a generator and a discriminator, GANs collaborate to produce images that are both realistic and diverse. The generator learns to create images by starting with random noise or input

DOI: 10.54254/2755-2721/79/20241286

prompts [11]. The generator refines the output graphics using sophisticated transformations to coincide them with the specified text prompts. The discriminator, on the other hand, plays the role of differentiating real images from those generated by the generator. By training the discriminator with a diverse dataset, it acquires the ability to discern the nuances that distinguish real and generated images. It is worth mentioning that there are various types of GANs.

These types are categorized based on their distinct characteristics and objectives, and each category is explored to unveil its underlying principles and techniques. The following categories are discussed.

3.1.1. Improved Training Methods. This category focuses on GAN variants that address the training challenges faced by conventional GANs. Methods such as Wasserstein GANs (WGANs), Least Squares GANs (LSGANs), and energy-based GANs (EBGANs) have been proposed to stabilize training, mitigate mode collapse, and enhance sample quality. These approaches redefine loss functions or introduce new adversarial training methods to ensure more reliable and efficient GAN training.

3.1.2. Conditional GANs. Unlike traditional GANs, conditional GANs incorporate additional information during the generation process. These GAN variants generate samples that are conditioned on specific attributes, labels, or input samples. Image-to-image translation, text-to-image transformation, and style transfer problems all benefit from conditional GANs. Prominent examples include Conditional GAN (cGAN), Auxiliary Classifier GAN (ACGAN), and StackGAN.

3.1.3. Domain Transfer GANs. GANs have proven to be capable of transferring knowledge from one area to another. Domain transfer GANs generate samples in a target domain while preserving the characteristics learned from the source domain. CycleGAN, DiscoGAN, and UNIT are noteworthy GAN models in this category, enabling realistic image-to-image translation without the need for paired training data.

3.1.4. Progressive GANs. Standard GANs may encounter challenges in generating high-resolution images due to limitations in training stability. Progressive GANs tackle this issue by gradually growing the generator and discriminator during training. This technique allows the generator to refine image details incrementally, resulting in high-quality, realistic images at higher resolutions. Progressive GANs have achieved impressive results in image synthesis tasks.

3.1.5. Text-to-Image GANs. GANs have been extended to produce visually realistic images from textual descriptions. Text-to-image GANs utilize both textual and image domains to learn their relationship, enabling the synthesis of images based on textual input. Models such as StackGAN++, AttnGAN, and Mirror-GAN tackle this formidable task. The discriminator must adjust as the generator improves in order to retain accuracy in recognizing the increasing quality of the produced pictures. GANs' success is based on their adversarial training process. As the generator and discriminator continuously compete, they strive to outperform one another. This competition drives both networks to refine their capabilities. The goal of the generator is to create images that closely resemble real ones, while the discriminator aims to become more discerning in distinguishing between generated images and real ones. This dynamic interplay leads to an equilibrium where the quality of the generated images steadily improves. GANs have shown that they can produce high-quality, diversified, and aesthetically attractive pictures that match up to the specified text prompts. GANs have opened up new possibilities in the realms of art, design, and entertainment by transforming textual descriptions into vivid landscapes and synthesizing intricate objects. They can also be employed for video synthesis, image inpainting, and style transfer. The versatility of GANs solidifies their position as a prominent method in the realm of artificial intelligence.

In conclusion, Generative Adversarial Networks have revolutionized image generation from text. By leveraging the power of competition, GANs create high-quality, diverse, and visually appealing images that closely align with the provided text prompts. As GANs continue to evolve and improve, they will shape the future of image generation and push the boundaries of creativity.

DOI: 10.54254/2755-2721/79/20241286

3.2. Variational Auto-encoders (VAEs)

Variational Auto-encoders (VAEs) have become a prominent tool in text-to-image generation, drawing attention for their ability to learn data distribution and generate diverse samples [12]. Unlike Generative Adversarial Networks (GANs), VAEs utilize an encoder-decoder architecture to transform textual prompts into lower-dimensional latent spaces, enabling image generation [12]. Their strength lies in comprehensively capturing the intricate distribution of training data, making them valuable for artistic creation, content generation, and data augmentation. The encoder maps textual inputs to a compressed latent space, and the decoder leverages these latent vectors to generate images aligned with the given text prompts. VAEs stand out for their capacity to produce diverse and novel images by sampling from the learned latent distribution, showcasing versatility. They offer a probabilistic framework, allowing exploration and manipulation of the latent space, enabling users to actively influence specific attributes in the generated images. For instance, adjusting certain dimensions can alter color schemes or shapes. This control empowers users in the image generation process, enhancing creativity and customization potential.

In conclusion, VAEs have emerged as a powerful tool in text-to-image generation, adept at learning and capturing the distribution of training data. Through their encoder-decoder architecture, they facilitate the transformation of textual prompts into lower-dimensional latent spaces, fostering the generation of visually plausible and diverse images. With a probabilistic framework and the ability to explore and manipulate the latent space, VAEs offer versatility and customization in image generation, unlocking possibilities in art, design, and content creation [12].

3.3. Attention Mechanisms

Attention mechanisms are now integral in text-to-image generation, enhancing image quality and coherence [13]. They enable selective concentration on specific text sections, crucial for aligning textual prompts with image regions. A key advantage of attention mechanisms is their ability to improve the overall visual quality of generated images. Previously, text-to-image models struggled to capture intricate details accurately, resulting in less realistic and less effective conveyance of intended meaning. With attention mechanisms, models allocate resources more efficiently, resulting in higher fidelity images that closely resemble their textual descriptions. Attention mechanisms contribute to coherence and relevance by selectively attending to specific text sections, ensuring alignment between generated image regions and corresponding textual information. This alignment avoids inconsistencies and contradictions between the image and its description. Furthermore, attention mechanisms enable the model to consider the context of text prompts, leading to more coherent and meaningful image synthesis. Beyond individual image synthesis, attention mechanisms enhance understanding and interpretability of text-to-image models. Visualization of attention weights assigned to different text parts provides insights into how models generate images and which textual cues have the most influence on synthesis [13].

In conclusion, attention mechanisms revolutionize text-to-image generation, improving visual quality, coherence, and interpretability. With attention mechanisms, models can produce realistic, meaningfully aligned images closely resembling textual descriptions, enhancing understanding and usability. Ongoing evolution promises further advancements in generating high-quality, contextually aligned images from textual prompts [13].

3.4. Reinforcement Learning(RL)

Reinforcement Learning (RL) proves effective in text-to-image generation, optimizing the process by utilizing a reward-based approach for iterative improvement [14]. The model refines image generation capabilities through feedback on quality, creating lifelike outcomes aligned with textual descriptions. RL addresses challenges like the semantic gap, enhancing the model's understanding of semantics and visual intricacies. This iterative process produces images reflecting text nuances and contextual meaning. RL enables the model to explore diverse possibilities, learning which features to prioritize based on feedback. This adaptability optimizes the generation process, enhancing proficiency in creating images

DOI: 10.54254/2755-2721/79/20241286

aligned with the intended visual style. RL contributes to efficiency by continually refining parameters, reducing the need for extensive computational resources, particularly valuable in real-time generation scenarios. In conclusion, RL techniques enhance the efficiency and effectiveness of text-to-image generation, bridging the semantic gap and creating faithful depictions of textual descriptions [14].

3.5. Transfer Learning

According to a study by Weiss [15], transfer learning has emerged as a powerful technique in text-to-image generation, significantly improving model performance. The authors explain that the approach involves two crucial steps: pre-training on large-scale image datasets, such as ImageNet, and fine-tuning on text-to-image datasets. By pre-training on ImageNet, models gain a comprehensive understanding of visual concepts, object recognition, and image composition. This pre-training provides them with a solid foundation of image-related knowledge to generate visually appealing and contextually relevant images from text [15]. After pre-training on ImageNet, models undergo fine-tuning with text-to-image datasets, as highlighted by [15]. The authors emphasize that this step ensures that the models align their acquired image representations with the text-to-image generation task. Fine-tuning involves training on text-to-image datasets, which contain textual descriptions paired with corresponding images. During this training, models adapt their pre-trained visual knowledge to effectively generate images from text [15]. Furthermore, [15] notes that transfer learning enables models to overcome the limitations of training solely on text-to-image datasets. These datasets are often smaller and less diverse compared to large-scale image datasets like ImageNet. However, by leveraging knowledge from pre-training, models can compensate for the lack of training data and generate images that surpass the quality and complexity limitations of text-to-image datasets alone.

In conclusion, transfer learning, as discussed by [15], significantly enhances text-to-image generation models by pre-training on large-scale image datasets and fine-tuning on text-to-image datasets. This approach allows models to understand visual representations and effectively translate textual descriptions into high-quality images. The authors highlight the immense potential of transfer learning in pushing the boundaries of text-to-image generation and its promise for future advancements in this field.

4. Experimental Results

4.1. Data Diversity and Generalization

The datasets used in assessing text-to-image generating models are critical. Datasets such as MS COCO [2] have acted as a benchmark for numerous computer vision tasks due to their extensive collection of varied pictures and language descriptions. Similarly, the Flickr30K dataset, with several captions per image, allows for the investigation of various textual descriptions linked with the same visual material. Furthermore, with its extensive scene graphs containing objects, properties, and connections inside images, the Visual Genome dataset [16] has been essential in furthering studies in object identification, image captioning, and visual relationship comprehension. With its large collection of picture-caption pairings, making the Conceptual Captions dataset appropriate for image captioning, multimodal representation learning, and cross-modal retrieval tasks. The inclusion of such broad datasets not only proves the models' capacity to generalize across domains but also their versatility in dealing with various sorts of textual descriptions and pictures.

4.2. Model Performance

Within the quickly advancing field of text-to-image era, the strategies talked about have showcased surprising capabilities and headways. These strategies have altogether affected the quality, differing qualities, and interpretability of created pictures. However, numerous benchmarking metrics, such as the Inception Score (IS) and Fr´echet Inception Distance (FID), have been utilized to quantitatively evaluate the quality and diversity of images. The Inception Score, introduced by Salimans, assesses the quality of generated images based on the diversity and clarity of their class labels [7]. Conversely, the Fr´echet Inception Distance (FID), proposed by Heusel, computes the commonalities between feature

DOI: 10.54254/2755-2721/79/20241286

distributions of generated and actual pictures retrieved from a pre-trained Inception-v3 network [8]. Lower FID scores show that created pictures are closer to genuine picture conveyances in terms of visual highlights, demonstrating higher quality and authenticity. Additionally, it was highlighted how accuracy and recall measures have been used to assess how well models correlate produced pictures with textual descriptions [17]. Text-to-image creation has undergone a profound transformation thanks to Generative Adversarial Networks (GANs). This novel strategy makes use of a competitive framework with a generator and a discriminator that cooperate to improve each other's skills [18]. GANs have produced images that demonstrate an amazing level of realism and diversity as a result of this adversarial training process. These produced pictures are densely packed with fine-grained features and subtleties, making them extremely believable [18]. Beyond their potential to produce high-quality graphics, GANs are significant to be remarkably versatile and adaptable in a wide range of fields, from the creation of art to content. This flexibility highlights the wide range of applications and value that GANs provide in the quickly changing technological world of today. GANs, on the other hand, provide their own set of difficulties. Training them may be computationally demanding and necessitates precise hyperparameter tweaking. Mode collapse, in which the generator focuses on a small collection of pictures, is a typical problem. In contrast to GANs, Variational Auto-encoders (VAEs) take probabilistic modelling and exploration of a latent space approach to text-to-image synthesis. VAEs excel in capturing the underlying data distribution by mapping verbal prompts into a lower-dimensional latent space, allowing visuals to be generated that correspond to the statistical features of the training data. VAEs provide a distinct edge in terms of customization and control [19]. Users may actively adjust latent vectors to impact certain picture qualities such as colours, shapes, or styles. However, VAEs may experience difficulties in attaining the same degree of visual fidelity as GANs. While they efficiently capture data distribution, the produced pictures may lack the fine-grained features and realism exhibited by GANs. The incorporation of attention processes has had a substantial influence on text-to-image generation models, improving their performance in a variety of ways. Models can use attention mechanisms to selectively focus on certain portions of textual prompts, leading in pictures with higher visual quality and greater alignment with verbal descriptions [20]. This selective attention improves the authenticity, coherence, and relevance of created pictures, resulting in visuals that are not just aesthetically accurate but also semantically significant. To summarize, GANs, VAEs, and attention mechanisms have jointly pushed text-to-image production to new heights. They have overcome earlier restrictions, such as low visual quality and inconsistencies with verbal descriptions, while also opening up new avenues for creative content creation.

5. Conclusion

To conclude, the field of text-to-image production has seen great progress, owing to the unique integration of numerous methodologies and techniques. VAEs have proved their ability to capture data distributions, allowing the production of different and distinct visuals from text descriptions [18]. Augmented Reality (AR) [21] has increased the usability of created pictures by augmenting real-world settings with virtual features, with applications in visualization, analysis, and decision-making. Benchmarking criteria like Inception Score, Fr´echet Inception Distance, Precision, and Recall have offered a quantifiable way to assess the quality, variety, and alignment of produced visuals with textual descriptions [7,22]. Also, the use of datasets such as MS COCO, Flickr30K, Visual Genome, and Conceptual Captions has enhanced the assessment process by offering extensive collections of photos and written descriptions [2,4]. These datasets have not only driven research but also acted as standards for evaluating the capabilities of various models across a wide range of computer vision and natural language processing applications. In conclusion, the advancements in text-to-image creation hold immense promise for transforming various industries and addressing real-world challenges. As these methodologies and techniques continue to evolve, we anticipate even more remarkable achievements in the generation of realistic, diverse, and contextually relevant images from textual descriptions. However, it is important to acknowledge the challenges of ethical considerations, data privacy, and the potential for misuse as these models become more sophisticated. While the potential significance of these models

DOI: 10.54254/2755-2721/79/20241286

in practical scenarios, such as content generation and decision-making, is vast, it comes with challenges related to data privacy, ethical concerns, bias and fairness, and the need for regulatory frameworks. Responsible research and development will be crucial in realizing the full potential of text-to-image models while mitigating potential risks and ensuring their ethical and secure use.

References

[1] Frolov, S., Hinz, T., Raue, F., Hees, J. & Dengel, A. Adversarial text-to-image synthesis: A review. *Neural Networks* **144**, 187–209 (2021).

[2] Lin, T.-Y. *et al.* Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755 (Springer, 2014).

[3] *Chen*, X. *et al.* Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[4] Young, P., Lai, A., Hodosh, M. & Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* **2**, 67–78 (2014).

[5] Krishna, R. *et al.* Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* **123**, 32–73 (2017).

[6] Tang, K., Zhang, H., Wu, B., Luo, W. & Liu, W. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6619–6628 (2019).

[7] Salimans, T. *et al.* Improved techniques for training gans. *Advances in neural information processing systems* **29** (2016).

[8] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS* (2017). URL https://api.semanticscholar.org/CorpusID:326772.

[9] Buckland, M. & Gey, F. The relationship between recall and precision. *Journal of the American society for information science* **45**, 12–19 (1994).

[10] Streijl, R. C., Winkler, S. & Hands, D. S. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems* **22**, 213–227 (2016).

[11] Goodfellow, I. *et al.* Generative adversarial networks. *Communications of the ACM* **63**, 139–144 (2020).

[12] Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[13] Wang, B. & Gong, N. Z. Stealing hyperparameters in machine learning. In *2018 IEEE symposium on security and privacy (SP)*, 36–52 (IEEE, 2018).

[14] Zhang, H. *et al.* Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 5907–5915 (2017).

[15] Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *Journal of Big data* **3**, 1–40 (2016).

[16] Xu, K. *et al.* Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* (2015). URL https://api.semanticscholar. org/CorpusID:1055111.

[17] Chen, X. *et al.* 3d object proposals using stereo imagery for accurate object class detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**, 1259–1272 (2016). URL https://api.semanticscholar.org/CorpusID:4552515.

[18] Goodfellow, I. J. *et al.* Generative adversarial networks. *Communications of the ACM* **63**, 139 – 144 (2014). URL https://api.semanticscholar.org/CorpusID:12209503.

[19] Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *CoRR* **abs/1312.6114** (2013). URL https://api.semanticscholar.org/CorpusID:216078090.

DOI: 10.54254/2755-2721/79/20241286

[20] Vaswani, A. *et al.* Attention is all you need. In *NIPS* (2017). URL https://api.semanticscholar.org/CorpusID:13756489.

[21] Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *CoRR* **abs/1409.0473** (2014). URL https://api.semanticscholar.org/ CorpusID:11212020.

[22] Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O. & Gelly, S. Assessing generative models via precision and recall. *ArXiv* **abs/1806.00035** (2018). URL https://api.semanticscholar.org/CorpusID:44104089.