Advancements In Text-To-Image Synthesis: A Comprehensive Review Of Techniques, Challenges, And Future Directions

SOUMYA ASHWATH[1], DR. JYOTHI SHETTY[2]
Department of CSE, Nitte (Deemed to be University) NMAM Institute of Technology (NMAMIT), Nitte, INDIA

Abstract: Recent advancements in text-to-image synthesis are explored through innovative approaches designed to address key challenges in generating realistic images from textual descriptions. These approaches include IIR-Net, CRD-CGAN, GALIP, Transformer-based methods, StyleGAN-T, and OPGAN. Each model introduces distinct techniques such as Image Information Removal (IIR), attention mechanisms, CLIP integration, and object-centric architectures, aiming to enhance fidelity, diversity, semantic consistency, and object modelling accuracy. Evaluation across diverse datasets demonstrates their superior performance over existing methods, highlighting improvements in editability, photorealism, and control over the synthesis process. Furthermore, future research directions are discussed, emphasizing the need for refining text alignment, advancing object modelling techniques, and exploring personalized GAN approaches to further advance text-to-image synthesis.

Key-words: Image synthesis, Machine learning, Image generation, GAN, Generative models, CG-GAN, Object modelling

Received: March 7, 2024. Revised: August 13, 2024. Accepted: September 8, 2024. Published: October 10, 2024.

1. Introduction

In many real-world contexts, like art generating, CAD, and image editing, image synthesis—the process of making synthetic images from various sources including text, drawings, sounds, or existing images—plays a crucial role. In recent years, this has sparked a great deal of curiosity among researchers. Convolutional Neural Networks (CNNs), Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), image retrieval, and diffusion models are some of the main types of image synthesis approaches. Despite covering all of these topics, this study focuses on GAN-based image synthesis because of its popularity and how quickly research in this area is progressing. Additionally, the potential of diffusion model-based image synthesis is discussed as a promising avenue for future research.

CNNs, widely used in visual tasks, excel in reducing image dimensionality while preserving vital information through convolution layers. By transforming input data into a distribution across latent space, VAEs—which consist of an encoder and a decoder—strive to reduce reconstruction errors. In GANs, a generator and a discriminator play a minmax game. In this game, the generator creates realistic images in an effort to trick the discriminator, who then determines whether the images are genuine or not. Sketch-to-image synthesis, which began with sketch-based image retrieval systems, today uses deep convolutional neural networks (CNNs) and generative adversarial networks (GANs) to create complex color images from basic sketches. Nevertheless, the process of synthesising complicated scenarios from drawings continues to be difficult.

Text-to-image synthesis aims to visually represent human-written sentences while preserving their semantic meaning. Initially relying on supervised approaches analysing word-to-image correlations, the field has shifted towards unsupervised deep learning approaches, notably GANs, for generating images from text. While synthesizing realistic images of single objects has seen progress, generating scenes with multiple objects remains a challenge.

Image-to-image synthesis maps input images from one domain to output images in another. This process entails maintaining the content of the input images while maybe altering certain features. Image-to-image synthesis models have extensively used GANs due of their efficacy with unknown data.

Speech-to-image synthesis creates images with comparable meaning to voice input. Recent research focuses on developing models capable of converting sounds into images, addressing complexities in machine perception within this domain. Overall, these diverse approaches collectively contribute to advancing the field of image synthesis, paving the way for various applications and future research endeavours.

Table 1. Shows the categories of image synthesis with the methodologies used.

Text to Image	Sketch to Image	Image to Image	Speech to Image
Traditional and CNN	Image retrieval based	GAN based	GAN
VAE	Deep CNN		
GAN	GAN		
Masked			

Table 2: Summary of methodologies reviewed

Title	**Features**	**Dataset**	**Evaluation metrics**
IIR-Net	-Integration of Image Information Removal (IIR) module -Two-stage model: conditional diffusion and IIR module -Enhanced editability and fidelity balance	CUB, Outdoor scenes, COCO	LPIPS, CLIP scores, qualitative assessments
CRD-CGAN	Focus on category-consistency and relativistic diversity constraints Integration of attention loss, diversity loss, relativistic conditional loss, and category-consistent loss components	Caltech-UCSD Birds-200-2011, Oxford 102 flower, MS COCO 2014	Photorealism, diversity, sensitivity to word attention
GALIP	Utilization of pretrained CLIP models in discriminator and generator Improved synthesis efficiency and quality Faster synthesis speeds	Challenging datasets	Training efficiency, synthesis speed, image quality
Text-to-LayoutGAN	Synthesis of layout from text and layout from layout to images may be modelled simultaneously. Emphasis on precise textual-visual alignment	Custom datasets	Layout Quality Score (bounding box distribution errors, spatial relationships)

	per object Introduction of Layout Quality Score metric		
StyleGAN-T	Enhanced capacity, stable training, and improved text alignment Utilization of truncation for improved text alignment Suggested avenues for future research	Large-scale datasets	Sample quality, speed, text coherence
OPGAN	Introduction of Semantic Object Accuracy (SOA) metric for evaluating object modelling Consistent outperformance over baseline architectures	Custom datasets	Semantic Object Accuracy, qualitative
Single-Stage Text-to-Image	Training in a single stage using a single discriminator and generator Utilization of deep residual networks and sentence interpolation strategy	Custom datasets	Realism, diversity, training efficiency
DM-GAN	Leveraging dynamic memory module for initial image enhancement Integration of memory writing and response gates Superior performance compared to existing approaches	Real-world datasets	Qualitative and quantitative measures
KT-GAN	Multi-stage approach with object-driven attention layers Utilization of Fast R-CNN based object-wise discriminators Substantial performance enhancement over state-of-the- art	COCO dataset	Inception score, FID score, qualitative assessments

Text-to-image synthesis aims to visually represent human-written sentences while preserving their semantic meaning. Initially relying on supervised approaches analysing word-to-image correlations, the field has shifted towards unsupervised deep learning approaches, notably GANs, for generating images from text. While synthesizing realistic images of single objects has seen progress, generating scenes with multiple objects remains a challenge.

The alignment of text with image content and the maintenance of coherence within created images are also problems that text-to-image synthesis must take into consideration. OPGAN addresses these issues by explicitly modelling individual objects within images, resulting in high accuracy in producing realistic images from complex textual descriptions. Furthermore, Obj-GANs leverage object-driven attention layers to enhance image synthesis quality, achieving significant performance improvements over prior models. When taken as a whole, these fresh perspectives advance text-to-image synthesis to a cutting-edge level and pave the way for exciting new possibilities in the area.

2. Related Work

Zhang, Zhongping, et al., In this work [1] IIR-Net, an innovative text-to-image editing model designed to address shortcomings of existing methods by integrating an Image Information Removal (IIR) module. Through selective removal of colour and texture details from the original image, IIR-Net ensures better preservation of text-irrelevant content and mitigates issues related to overfitting and information concealment. It comprises two key stages: a conditional diffusion model that leverages the original image as supplementary control, and the IIR module to tackle concerns regarding identical mapping. Results from experiments conducted on CUB, Outdoor Scenes, and COCO datasets showcase superior performance in balancing editability and fidelity compared to previous approaches, with notable enhancements in LPIPS and CLIP scores observed particularly in COCO evaluations. Furthermore, qualitative assessments underscore the model's adeptness in modifying desired attributes while upholding the integrity of the original content.

The researchers (Hu, Long, et al., 2004) Using category-consistency and relativistic diversity constraints as priorities, this research [2] introduces CRD-CGAN, a new conditional generative adversarial network (GAN) developed for image generation from textual descriptions. CRD-CGAN integrates diversity loss, relativistic conditional loss, attention loss, and category-consistent loss to improve word attention sensitivity, realism estimation, and visual coherence in generated images. Thorough experiments on the Caltech-UCSD Birds-200-2011, Oxford 102 flower, and MS COCO 2014 datasets reveal that CRD-CGAN surpasses state-of-the-art techniques in photorealism and image diversity. Notably, CRD-CGAN adeptly captures nuances in textual descriptions, maintains relative authenticity in generated images, and ensures consistency with the main visual features of the corresponding categories, affirming its efficacy across various datasets encompassing complex scenes and multiple categories. The GALIP was introduced by Tao, Ming, et al., which was a novel [3] method for efficiently generating high-fidelity complex images from textual descriptions while maintaining control over the synthesis process. GALIP utilizes pretrained CLIP models in both the discriminator and generator components. The CLIP-based discriminator accurately assesses image quality, while the CLIP- synthesis. This integration results in enhanced training efficiency, which requires a substantially lower amount of data and parameters in comparison to other methodologies that are already in use, while yet producing results that are equivalent. The synthesis rates achieved by GALIP are much quicker, and it inherits the smooth latent space properties that are distinctive of GANs. The results of experimental assessments reveal that GALIP performs very well on difficult datasets, demonstrating its capacity to produce complex images of a better quality. In addition, the incorporation of the understanding model (CLIP-ViT) into the generative framework highlights the possibility of synergies between the understanding model and the generative model, which suggests that there may be opportunities for the creation of generic large-scale models from a future perspective.

A innovative way to addressing the difficulty of preserving semantic consistency in text-to-image synthesis is presented in this study [4] by J. Liang and colleagues. This method models text-to-layout creation and layout-to-image synthesis simultaneously. This method uses Transformer models to reframe text-to-layout creation as sequence-to-sequence. Unlike the usual technique, which struggles with text-derived object spatial distribution patterns. It figures out where things are in relation to one another in the layout by taking use of sequential dependencies. When it comes to layout-to-image synthesis, the model places an emphasis on exact textual-visual semantic alignment per item. The Layout Quality Score is a novel measure that takes spatial connections and mistakes in the distribution of bounding boxes into consideration while conducting evaluations. The recommended technique beats the present state-of-the-art methods when it comes to predicting layouts and producing visuals from text, according to the findings of exhaustive tests that were carried out on three distinct datasets respectively.

(Axel Sauer et al.) In order to address the difficulties of large-scale text-to-image synthesis, the authors of the aforementioned article [5] present StyleGAN-T, which has the following benefits: better text alignment with controlled variance, consistent training across various datasets, and increased capacity. StyleGAN-T creates better samples and processes them faster than early GANs and even distilled diffusion models. StyleGAN-T struggles with connecting characteristics and objects and creating consistent text inside images, similar to DALL·E 2 using CLIP. Though it may increase runtime, a broader language model may help. Truncation is identified as a means to improve text alignment, yet it differs from diffusion model guidance, indicating the need for alternative methods. Furthermore, the work identifies prospective directions for future research, such as the refinement of super-resolution stages and the exploration of customized GAN techniques that are comparable to those seen in diffusion models.

T. Hinz et al., The paper [6] introduces OPGAN, a novel GAN architecture aimed at addressing challenges in generating images from intricate textual descriptions by explicitly modelling individual objects within images. A novel measure called Semantic Object Accuracy (SOA) is suggested for quantitatively assessing these models. It checks whether the produced images include the items listed in the input caption. A user study validates that SOA aligns with human judgment better than other metrics like the Inception Score. OPGAN consistently outperforms baseline architectures in both quantitative and qualitative evaluations, highlighting its effectiveness in producing realistic images. Furthermore, SOA evaluation highlights current struggles in modeling rare or complex objects, underscoring the need for ongoing advancements in text-to-image synthesis techniques.

"Souza et al." (D. M.) [7] propose a revolutionary neural architecture that achieves state-of-the-art performance with single-stage training using a single generator and discriminator, departing from the standard method to text-to-image synthesis. This innovation represents a significant departure from the current technique. In contrast to earlier approaches that relied on multi-stage training to overcome difficulties in integrating data from various modalities and training GANs at high resolutions, this method makes use of deep residual networks and an innovative sentence interpolation strategy to effectively learn a smooth conditional space. By showcasing the effectiveness of this architectural shift, the paper pioneers a new direction for text-to-image research, emphasizing the potential for exploring innovative neural architectures in this domain.

The Dynamic Memory Generative

Adversarial Network (DM-GAN) is given in the research article [8] for text-to-image synthesis to address difficulties with existing approaches. A memory writing gate is used by DM-GAN in order to emphasize significant text information depending on the content of the image. A dynamic memory module improves early images, especially badly designed ones. To accurately merge memories and images, a response gate is employed. Evaluation on real-world datasets validates DM-GAN's superior performance compared to existing approaches across qualitative and quantitative metrics. Despite its advancements, DM-GAN acknowledges limitations in handling complex multi-subject layouts and suggests avenues for future research to refine initial image generation capabilities. S. H. Tan and his fellow researchers, [9]KT-GAN is a novel framework that is described in this study for the purpose of creating fine-grained text-to-image conversions. Key processes include the Semantic Distillation Mechanism (SDM) and Alternate Attention Transfer Mechanism (ATM). Through the real-time modification of word attention weights and image sub-region attention weights, AATM is able to continuously enhance the quality of essential word and image details. By directing the training of a text encoder with an image encoder that was trained for an Image-to-Image assignment, SDM is able to enhance both the encoding of text features and the quality of images. We can see that KT-GAN significantly outperforms baseline approaches in experimental validation on public datasets, with competitive outcomes across all of our assessment measures. The fact that it successfully bridges the gap between text and image demonstrates its usefulness.

Li, Wenbo, and colleagues To create complicated scene images from written descriptions, the authors of the article [10] provide Object-driven Attentive Generative Adversarial Networks, or Obj-GANs. Obj-GANs utilize a multi-stage approach featuring innovative object-driven attention layers to emphasize key objects based on pertinent words and pre-generated layouts, thereby improving the quality of image synthesis. Fast R-CNN-based object-wise discriminators provide precise object-level discrimination signals, improving text description and layout alignment. Obj-GAN beats state-of-the-art models on the COCO benchmark. This is shown by a 27% improvement in the Inception score and an 11% decrease in the FID score. The usefulness of object-driven attention in the generation of high-quality complex sceneries is highlighted in this work via the use of detailed comparisons with typical grid attention techniques.The above table represents a concise summary of each paper highlighting key features, datasets used and evaluation metrics.

Finding novel approaches to the complex problems that text-to-image synthesis entails, the study delves into this complex topic. These challenges encompass the effective representation of complex semantic relationships within scenes, the precise alignment of textual descriptions with resultant images, the accurate modelling of rare or intricate objects mentioned in input text, and the optimization of training and inference procedures for enhanced efficiency. Furthermore, the study recognizes that in order to thoroughly evaluate the quality of synthetic images, strong assessment measures are required.

While the proposed methods represent significant advancements, there remain gaps in fully comprehending and addressing these challenges. Future research directions may entail further refining existing models or pioneering new techniques to bridge these gaps, ultimately pushing the boundaries of text-to-image synthesis for diverse applications and domains.

Table 3. Summary of research gap.

Research Gap	Description

Handling Rare or Complex Objects	While some models excel at generating common objects, they may struggle with rare or complex ones. Research into better modelling and generation techniques for less frequent objects, potentially incorporating domain-specific knowledge or data augmentation strategies, could address this limitation.
Efficient Training and Inference	Many models require significant computational resources, hindering scalability and practical utility. Investigating more efficient training algorithms and model architectures to achieve comparable performance with reduced computational cost would enable wider adoption in real-world applications.
Evaluation Metrics	When it comes to semantic consistency and perceptual realism, in particular, existing assessment criteria could be falling short when it comes to capturing image quality. Improving the credibility and usefulness of text-to-image synthesis might be achieved by creating thorough metrics that are in line with human perceptual judgments. This would allow for more precise evaluations of the model's performance.
Handling Complex Semantic Relationships	Models need improvement in capturing complex semantic relationships between objects and scenes to produce more accurate and contextually rich images.
Enhancing Text-Image Alignment	Several models face challenges in aligning textual descriptions with generated images, especially in maintaining semantic consistency and coherence. Exploring novel approaches, such as advanced attention mechanisms or additional contextual information, could improve synthesis quality.

Datasets for Text-to-Image Synthesis

Various datasets are commonly utilized for text-to-image synthesis research. These datasets serve as essential resources for training and evaluating models in this domain:

- **MS COCO** (Microsoft Common Objects

in Context): MS COCO is a widely used dataset comprising over 330,000 images, each accompanied by at least 5 different captions. Its extensive collection covers diverse scenes and objects, making it suitable for tasks such as image captioning and text-to-image synthesis.

- **CUB-200-2011**: Featuring 200 different bird species, the Caltech-UCSD Birds-200-2011 collection has 11,788 images. Alongside images, it provides annotations including attribute labels, bounding boxes, and descriptions. Researchers often employ this dataset for fine-grained image recognition tasks and text-to-image synthesis involving birds.
- **Oxford-102 Flowers:** Annotated with category names and bounding boxes, the Oxford-102 Flowers collection contains 8,189 photos representing 102 different flower types. In particular, it finds widespread usage in flower-centric text-to-image synthesis and fine-grained identification.
- **Visual Genome**: The Visual Genome dataset consists of over 100,000 images densely annotated with object instances, relationships, and attributes. Its rich contextual information makes it valuable for generating images from textual descriptions, especially for complex scenes.
- **ADE20K**: Included in ADE20K are more than 20,000 photos focused on scenes that have been tagged with objects, their components, qualities, and connections. This dataset is instrumental in generating intricate scenes from textual descriptions.
- **COCO-Stuff**: An extension of MS COCO, COCO-Stuff includes pixel-wise annotations for 80 object categories and additional stuff categories like sky, grass, and road. It provides detailed semantic segmentation masks, aiding in the creation of realistic scenes from text.
- **FashionGen**: FashionGen contains images of fashion items along with textual descriptions, catering specifically to text-to-image synthesis tasks in the fashion domain.
- **Multi30K**: Multi30K is a multilingual dataset comprising 31,014 images paired with English and German descriptions. It facilitates research in cross-lingual text-to-image synthesis.

Figures 1, 2, and 3 below demonstrate the various datasets used for the text-to-image synthesis method.

Fig 1. Textcontrol GAN model

Fig 2. Example of T21 Dataset

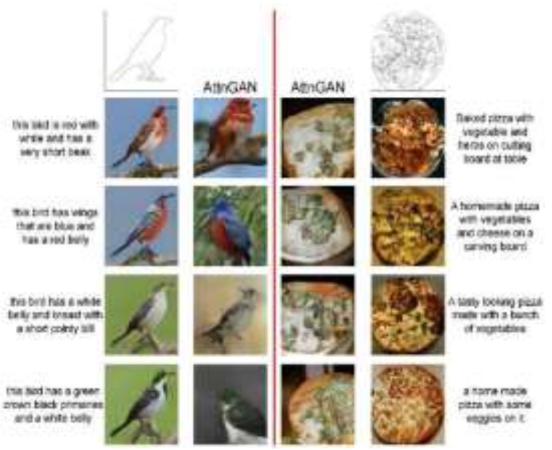

Fig 3. The CUB and MSCOCO dataset for AttnGAN

Researchers may examine different facets of text-to-image synthesis in different areas using these datasets, which provide varied visual material and written descriptions.

3. Applications

Thanks to its capacity to connect visual representations with written descriptions, text-to-image synthesis has grown into a flexible technology with uses in several fields. It adds vibrant visual context to product listings, which improves the purchasing experience, and it changes the game for online product display by taking written descriptions and turning them into realistic visuals. Similarly, within educational contexts, text-to-image synthesis serves as a valuable tool for creating illustrative materials from textual content, facilitating comprehension and bolstering knowledge retention among learners. In the field of design, it enables swift prototyping and visualization of ideas, empowering designers to iterate and refine concepts efficiently. In addition, text-to-image synthesis is a driving force in the entertainment and gaming industries, allowing for the creation of more engaging virtual worlds, characters, and stories. Additionally, in healthcare settings, this technology aids healthcare professionals by generating visual representations of medical conditions outlined in patient records, thereby assisting in accurate diagnosis and treatment planning. The overall potential of text-to-image synthesis is enormous across many disciplines, and it is already driving innovation and progress in many different areas.

Table 4: Applications of Text -to – Image Synthesis

Domain	Applications
E-commerce	When it comes to online shopping, text-to-image synthesis is essential for turning written product descriptions into photorealistic visuals. This process significantly enhances the shopping experience by providing consumers with visual context, thereby improving their understanding and decision- making process.
Education	Within the educational domain, text-to-image synthesis serves as a valuable tool for generating visual aids from textual content. These visual aids aid in comprehension, knowledge retention, and engagement among learners, making complex concepts more accessible and memorable.
Design	Text-to-image synthesis facilitates rapid prototyping and visualization of design concepts, enabling designers to swiftly iterate and refine their ideas. By providing visual representations of textual descriptions, this

	technology enhances the creative process, streamlining design workflows and fostering innovation.
Entertainment/Gaming	The entertainment and gaming business relies heavily on text-to-image synthesis, which enables the development of captivating virtual worlds, characters, and stories. By converting textual descriptions into lifelike images, this technology enhances user experiences, increasing engagement and immersion in virtual worlds.
Healthcare	In healthcare settings, text-to-image synthesis assists medical professionals by generating visual representations of medical conditions described in patient records. These visual representations aid in accurate diagnosis, treatment planning, and communication among healthcare providers, ultimately improving patient care and outcomes.

4. Future Directions

Future research in text-to-image synthesis should prioritize several key areas for advancement. Improving models to keep produced visuals and textual descriptions consistent semantically is the first order of business. This calls for the creation of more advanced systems that can accurately align visual components with text and capture subtle semantic links. Additionally, enhancing the efficiency of large-scale text-to- image synthesis is crucial, involving exploration of strategies to optimize training algorithms and model architectures to reduce computational resource requirements while maintaining synthesis quality and performance. Furthermore, improving the diversity and realism of synthesized images, particularly in modelling rare or complex objects mentioned in input text, necessitates the integration of domain-specific knowledge and advanced data augmentation techniques. There is still a pressing need to provide all-encompassing assessment metrics that can reliably measure the precision and accuracy of produced images, taking into account both quantitative and qualitative factors like perceptual realism and semantic consistency. Lastly, integrating multimodal understanding into the text-to- image synthesis process could be beneficial by leveraging insights from understanding models like CLIP to enhance the synthesis process and improve alignment between textual descriptions and generated images.

5. Conclusion

In conclusion, there are many different ways to approach image synthesis, each of which has its own benefits and drawbacks. Text-to-image, sketch-to-image, image-to-image, and speech-to-image synthesis are three examples of these types of synthesis. Notably, GAN-based techniques have emerged as particularly effective, especially in generating lifelike images from textual descriptions. Recent innovations like OPGAN, Obj-GANs, IIR-Net, and CRD-CGAN signify substantial progress in overcoming obstacles related to semantic consistency, object representation, image manipulation, and category adherence. Additionally, the integration of multimodal understanding, utilization of pretrained models like CLIP, and exploration of novel neural architectures offer promising pathways for future investigation. As a result of these combined efforts, image synthesis approaches are constantly improving, which in turn opens up a wide range of practical applications.

References

[1] Zhang Z, Zheng J, Fang Z, Plummer BA. Text-to-image editing by image information removal. InProceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2024 (pp. 5232-5241).

[2] Hu T, Long C, Xiao C. Crd-cgan: Category-consistent and relativistic constraints for diverse text-to-image generation. Frontiers of Computer Science. 2024 Feb;18(1):181304.

[3] Tao M, Bao BK, Tang H, Xu C. GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (pp. 14214-14223).

[4] J. Liang, W. Pei and F. Lu, "Layout-Bridging Text-to-Image Synthesis," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 12, pp. 7438-7451, Dec. 2023, doi: 10.1109/TCSVT.2023.3274228

[5] Sauer A, Karras T, Laine S, Geiger A, Aila T. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. arXiv preprint arXiv:2301.09515. 2023 Jan 23.

[6] T. Hinz, S. Heinrich and S. Wermter, "Semantic Object Accuracy for Generative Text-to- Image Synthesis," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 3, pp. 1552-1565, 1 March 2022, doi: 10.1109/TPAMI.2020.3021209.

[7] D. M. Souza, J. Wehrmann and D. D. Ruiz, "Efficient Neural Architecture for Text-to-Image Synthesis," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9207584.Zhu M, Pan P, Chen W, Yang Y. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 5802-5810).

[8] H. Tan, X. Liu, M. Liu, B. Yin and X. Li, "KT-GAN: Knowledge-Transfer Generative Adversarial Network for Text-to-Image Synthesis," in IEEE Transactions on Image Processing, vol. 30, pp. 1275-1290, 2021, doi: 10.1109/TIP.2020.3026728

[9] Li W, Zhang P, Zhang L, Huang Q, He X, Lyu S, Gao J. Object-driven text-to-image synthesis via adversarial training. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019 (pp. 12174-12182).

[10] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In ICML, 2016:1-6. S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In NIPS, 2016:1-7.

[11] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In ICML, 2016:1-6.

[12] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In NIPS, 2016:1-7.

[13] E. L. Denton, S. Chintala, A. Szlam, R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks", Proc. 28th Int. Conf. Neural Inf. Process. Syst., pp. 2015:1486-1494.

[14] X. Wang, A. Gupta, "Generative image modeling using style and structure adversarial networks", Proc. Eur. Conf. Comput. Vis., pp.2016:318-335.

[15] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, S. Belongie, "Stacked generative

adversarial networks", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp.2017:1866-1875.

[16] ZHANG Han,XU Tao,LI Hongsheng,et al . Stackgan: Text to photo realistic image synthesis with stacked generative adversarial networks[J]. IEEE International Conference on Computer Vision, 2017 ,2(3):5908 5916. DOI: 10.l109/ICCV.2017, 629.

[17] ZHANG H, XU T, LI H, et al. StackGAN++: realistic image :synthesis with stacked generative adversarial nel works[J]. IEEE Trans actions on Pattern Analysis and Machine Intelligence, 2019 ,41(8): 1947-1962.

[18] X.Chen, D.Chen. A text generation image model based on classification reconstruction stack generation against network[J]. Journal of Huaqiao University (Natural Science Edition), 2019, 40(04): 549-555.

[19] Lin T-Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context[A].Computer Vision - ECCV 2014[C]. Springer, Cham, 2014:740-755.