iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75737

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

Analysis of GAN-Driven Methods for Text-to-Image Conversion

Bhagwantkour Sahebsingh Jimeedar
Shri Guru Gobind Singh Ji Institute of Technology, Nanded, 431601, India

Abstract: The primary objective of text-to-image generation is to produce realistic and visually clear images that accurately correspond to the given textual descriptions. Among numerous approaches, Generative Adversarial Networks (GANs) have emerged as a key method in achieving effective text-based image synthesis. Depending on the generation goals, GAN-based text-to-image models can be categorized into three main functional areas: enhancing the realism of generated content, improving semantic alignment between text and image, and increasing the diversity of generated outputs. To address these areas, this study examines improvements in content authenticity through quality optimization, fine-grained detail enhancement, contextual refinement, and adaptive structural adjustments. From the standpoint of structural design, semantic extraction, spatial arrangement, and cycle consistency, the research also explores methods for strengthening semantic correlation. Furthermore, it investigates strategies to enhance content diversity through refined training mechanisms and effective text preprocessing approaches. This work provides an in-depth review of key methodologies proposed by prior researchers, emphasizing their design frameworks and processing pipelines. Comparative analyses are performed using established benchmark datasets to evaluate model performance and identify improvement opportunities. Finally, this research outlines future directions and potential developments to encourage continued progress in the field of text-to-image generation.
Keywords: GAN, Text-to-Image Generation Method, Cycle Consistency.

I. INTRODUCTION

In recent years, the rapid progress of deep learning has greatly driven advancements in fields such as natural language processing, computer vision, and image synthesis technologies. Within these areas, text-to-image generation has become a key research direction. This technique demonstrates strong multimodal interaction capabilities by converting natural language input into corresponding visual outputs. AI-based generators like DALL·E 2 and Midjourney are prime examples, capable of creating distinctive facial representations corresponding to specific medical conditions such as thyroid disorders and Hoven syndrome. These models produce high-quality synthetic medical images derived from textual descriptions, helping address patient privacy issues encountered with traditional medical imaging. Looking forward, text-to-image generation holds promising potential for expansion into various other application domains. Text-to-image generation, often regarded as a form of text-to-image (T2I) style transfer, seeks to create realistic images that satisfy the context and constraints of the input text. The foundational concept of TextCNN involves applying Convolutional Neural Networks (CNNs) to text classification tasks to capture essential textual features (Kim, 2014). For more complex applications, Deep Convolutional Neural Networks (DCNNs) extend the basic CNN architecture by incorporating additional convolutional and fully connected layers, enabling the extraction of more abstract and high-level feature representations (Atwood and Towsley, 2015). However, CNN-based text-to-image models tend to face limitations such as low generation efficiency and reduced output accuracy. Subsequently, Variational Autoencoders (VAEs) emerged, offering stable training performance and continuous latent space generation capabilities, making them particularly suitable for image reconstruction and denoising tasks. Nevertheless, their primary limitation lies in producing low-resolution images. As text descriptions became more complex and expectations for image clarity increased, Generative Adversarial Networks (GANs) began playing a crucial role in improving generation quality. GANs are particularly effective in generating high-resolution, photorealistic images through the adversarial training between a generator and discriminator. As research evolved, GAN-based text-to-image frameworks developed progressively across three major functional stages.

The first stage focuses on improving the realism of generated imagery. In this stage, methods such as Cross-domain Feature Fusion GAN (CF-GAN) were introduced to strengthen residual modules, improving the efficiency of feature extraction. Models using Feature Fusion Enhanced Response Modules (FFEM) and Multi-Branch Residual Modules (MBRM) further enhance deep-level feature integration, resulting in more lifelike outputs. The second stage aims to strengthen semantic alignment between text and

generated visuals. For superior generation quality, it is essential not only to enhance image realism but also to ensure that the produced images accurately reflect the semantic meaning of the input
text. The PMGAN model addresses this by utilizing a CLIP-based text encoder to extract textual features and map both text and image representations into a shared semantic space. This ensures that the synthesized images capture the full richness of the textual content. Once realism and semantic consistency are achieved, the next focus is on improving diversity in generation. Compared to early GAN approaches, RiFeGAN introduces an attention-based caption matching mechanism that enriches text descriptions and provides greater variation. Through the use of self-attention embedding mixtures and multiple caption candidates, RiFeGAN generates synthetic images with broader diversity and more nuanced details, thereby expanding the expressive range of generated content. This paper first discusses the significance of text-to-image synthesis, emphasizing both its theoretical research implications and practical applications. It then presents a novel classification framework based on the functional evolution of adversarial T2I models, analyzing improvements across different development stages. Performance metrics, evaluation measures, and application domains are also discussed. Finally, the study provides an outlook on future research directions aimed at further advancing the field of text-to-image generation.

II. T2I GENERATION MODEL CLASSIFICATION BASED ON FUNCTION

A. T2I Generation Based on Improving Content Authenticity

In the initial settings, A discriminator and a generator contribute to a GAN, and they are taught with mutually competing objectives. To track the identifier, the generator is taught to produce samples that are dispersed toward the real data, while the discriminator is tuned to distinguish between the fake and genuine samples that the generator produces. It shows huge potential in simulating complex data distribution, but GAN is difficult to train. When training GAN generates high-resolution (256*256) real images, a common failure phenomenon is Essence Table 1 shows a research method generated based on text image generation based on the authenticity of content.

Table 1: In summary of text image generation strategy that improves content authenticity

Model	Innovation	Advantage	Limitation
StackGANs	Decomposition of the production task of difficulty into a sub-problem with a gradual goal	train GAN stably to generate high-resolution images	Related multiple identification devices at different stages of the network.
HfGAN	End-to-end network with adaptive fusion multi-level features	Just one identifier can generate photo-level realistic images	The generated image is lost with the corresponding details of the word level
DualAttn-GAN	Introduce Double-attention module	Strengthen local details, text details, and image details corresponding to	When the initialization quality of the initial image is not high, the quality of the initial image will not be too good to refine the initial image again
CF-GAN	The comparatively straightforward and creative residual network topology is capable of fully extracting characteristics.	The target objects that solve the images are incomplete and the texture structure is not detailed	
DGattGAN	Introduced a cooperative sampling mechanism, decoupled object, and background generation	Details of fine particle size on the actual target object	
DMGAN	Introduce dynamic storage modules to refine the vague image content	Can accurately generate images from the text description	Different sentences with the same meaning might create different images.

1) *Improve the Quality of the Stack Structure*

In 2018, Zhang and others proposed a stacking GAN (StackGANs). By breaking the generating task into a child problem, it is stable to train GAN to generate high-resolution images. StackGAN-v1 has designed a two-stage architecture for T2I synthesis. Stage-I GAN generates low-resolution images, and Stage-II GAN generates high-resolution images based on this. Subsequently, the StackGAN-v2 introduced a multi-stage architecture, including multiple generators and identifiers. Compared to StackGAN-v1, it produces multi-scale pictures of the same language environment from various tree branches, demonstrating a steadier training impact. The preceding approach can be improved by breaking down the challenging task of producing high-quality images into a few more manageable subproblems. Nevertheless, there are still insufficient. For example, input is a global sentence vector that loses fine-grained word-level information, which means the generated image loses the corresponding details.

2) *Fine Particle Size Enhancement*

To address the issue of incomplete target objects in the image and refinement of the texture structure, 2022, Zhang et al. Proposed cross-domain characteristics fusion to generate confrontation network CF-GAN. This framework contains characteristic FFEM and MBRM. It is refined to generate images through deep fusion vector characteristics and image features. MBRM is an innovative residual network structure that can effectively extract features.

3) *Context Background Enhancement*

In 2021, Zhang and others proposed the dual- generator's attention GAN (DGattGAN) to pay attention to the objects and backgrounds in the input text to solve the high-quality problem of generating images. This model establishes two independent generators, decoupled objects, and background generation, and introduces cooperation sampling mechanisms to promote collaboration between the two. At the same time, asymmetric information feeding schemes are used to synthesize each generator based on the receiving semantic information. Through effective dual-generating machines and attention mechanisms,
DGATTGAN can generate fine-grained details on the target object.

The cooperation sampling mechanism they introduced may be very useful, because any dual generator architecture in the GAN model can benefit from this mechanism.

4) *Dynamic Adjustment*

To improve the authenticity of the content of the image, in the process of refining the existing images, unchanged text representations were used. Zhu and others put up a dynamic memory generation confrontation network (DM-GAN) to generate high- quality images. When the initial image is not generated well, this method introduces a dynamic storage module to refine the vague image content. Their techniques enable precise picture generation from text description by designing a memory writing entry to pick relevant text information depending on the original image content. To adjust to the data and picture characteristics read from memory, it also makes use of the response door.

This method effectively solves the problem of the quality of refining the initial image again will not be too good if the original image's quality is low initialization.

B. *Text Image Generation Based on Enhanced Semantic Correlation*

To generate high-quality images, not only need to improve the authenticity of generating images but also emphasize the consistency of generating images and text semantics in-text images. If the semantic information extracted in the text in the text is not enough, images produced by several texts with the same meaning may differ in certain ways. Table 2 shows the research method generated by text images based on enhancing semantic correlations.

1) *Attentive Mechanism*

Liu et al. put forward a knowledge transfer generating confrontation network (KT-GAN) and introduced alternating attention transfer mechanisms (AATM) and semantic distillation mechanism (SDM). AATM gradually highlights important words and enriches the details of the image. SDM guides text encoder training to generate better text features to improve image quality.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

Table 2: Research on text image generation of semantic correlation

Model	Innovation	Advantage	Limitation
AttnGAN	Pay attention to related words in natural language description	Synthetic images in different sub-regions of different sub-regions	It helps to increase the content details of the given information in the input text, but it cannot solve the problem that the input text is more abstract
SEGAN	Introduce attention competition module (ACM)	Pay attention to the weight and improve stability and accuracy, more attention than ATTNGAN	
KT-GAN	Introducing alternating attention transfer mechanism (AATM)	Update the weight of the attention alternate, highlight the important word information	
ControlGAN	Based on ATTNGAN, follow the multi-stage architecture	Allows users to manipulate the object attribute without affecting the generation of other content	
PMGAN	Text and pictures can be mapped to a shared semantic space using the CLIP encoder.	The rich semantic data contained in the text input may be fully utilized by the model.	
SDGAN	Extract semantic public points to achieve the consistency of image generation	Reserve semantic diversity and details to achieve fine-grained image generation	The problem of local text processing is solved, but the overall macro adjustment needs to be improved
TCF-GAN	Introduce DAMSM loss	Improve text utilization, monitor similarity between text, and improve semantic consistency	
ALR-GAN	Adaptive layout refine (ALR) loss to balance hard features and easy-to-character matching	Used to refine the layout of synthetic images adaptive without any auxiliary information	The diversity of the content of the image is not high
RII-GAN	DFAD uses specific synchronous dual-mode information extraction structures to improve semantic consistency	The dependence on text description is reduced in the introduction of the layout structure in the generator	
MirrorGAN	Learn T2I generating by re-description	With the use of the dual T2I and I2T adjustments, the text's meaning restoration loss	

2) Semantic Enhancement

For the semantic problems in T2I generation and the limitations of the detailed description object, Yu et al. In 2024, the generating model-based generating confrontation network. This model's generator and identification device make use of several pre-training models. PMGAN extracts the first image features from the text using the CLIP text encoder and then extracts the image feature to provide input for unconditional and conditional identifiers using a pre-trained CLIP image encoder. The CLIP encoder associates pictures and text with the same semantic space, helping to generate high-quality images. In addition, PMGAN also uses DAMSM text encoders which are pre- trained to excerpt the semantic embedded of thick granularity and fine particle size as a condition input guidance image. To improve the effectiveness of the embedded, each upper sample fusion module of the model has an attention fusion module and a deep fusion module to make full use of the rich semantic information in the text.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

3) Semantic Commonality Extraction

Zhou and others proposed a single-level network TCF-GAN to generate a highly detailed image from the text. This method relieves the details of the details of the stacking network and a large amount of calculation. It only needs one generator. TCF-GAN introduces text convergence modules (TAM) and text connection fusion (TCF) blocks and improves text utilization through door control recursive units (GRU) and upper sample blocks. In addition, the loss of depth multi- modal similarity model (DAMSM) is adopted to enhance the semantic consistency between the generated image and text.

The above method has reached the goal of enhancing semantic correlation from local text processing. Next, we will enhance the semantic correlation between text and generating images from the overall layout and adaptive adaptation of the overall layout.

4) Semantic Layout

Yuan and others proposed reverse image interaction to generate a confrontation network (RII-GAN), and achieve alignment of text and images by introducing the layout structure. Text encoders, adaptive emitting generators, reverse image interaction networks (RIIN), and dual- channel feature lying discriminators (DFAD) are all part of the network. To overcome the generating networks' lack of genuine image characteristics, RIIN adds real image distribution into the network. Each imitation block in the adaptive imitation generator enhances text information, while DFAD extracts the important features of images and text to improve semantic consistency.

5) Cyclic Consistency

Qiao and others proposed the Mirrorgan framework to improve T2I (T2I) generation. This framework contains three sections: STEM, GLAM, and STREAM. STEM creates vocabulary and sentences utilized in GLAM. GLAM uses a grade association structure to produce target pictures ranging from thick to thin and enhances the diversity and semantic consistency of the image through local word attention and global sentences. STREAM re-generates the same text from the genetic image as the given text description. For end-to-end training, visual realist confrontation loss and text-image pairing semantic consistency confrontation loss. At the same time, the loss of the loss of text based on cross-entropy is used to use the dual adjustment of T2I and I2T.

C. Text Image Generation Based on Content Diversity

After solving the authenticity and semantic correlation of generating images, to make the generated images more effectively solve user needs and provide more choice solutions, it is necessary to consider the diversity of content in the development process. Related research is shown in Table 3.

1) Training Mechanism

Similar to the text-segan proposed by AC-GAN and CGAN, Miriam Cha, and others proposed Text- SeGAN, the semantic correlation between text and images is estimated by the regression method, not prediction. This additional regression mission improves the diversity of the generator output, thereby alleviating the problem of pattern collapse.

Table 3: Research on text image generation of rich content diversity

Model	Innovation	Advantage	Limitation
AC-GAN	Auxiliary information such as category labels promote dual-shot mapping	It trains GAN stably to generate high-resolution images	For images with complex scenarios and multiple objects, text is the title that always describes the most obvious object or features in the image, and the details of the region and objects are often lacking
TAC-GAN	The category of predicting the image	Just one identifier can generate photo-level realistic images	
Text-SeGAN	Extra semantic correlation, training discriminator to estimate semantic correctness measurement	Strengthen local details, text details and image details corresponding to	

VQA-GAN	Q & A is selected as a locally related text, and the accuracy of VQA is used as a new evaluation indicator	Generate the description of the local image area or object, not the entire image	More quality indicators are required to achieve more complete assessment
RiFeGAN2	The family subtitle-matching model selects and refines the candidate subtitles from the ancestral knowledge	A novel approach to T2I synthesis	It is necessary to use more complicated methods in natural language understanding and image synthesis to further improve performance
GAN based on pre-training BERT	Fine-tune input text content using BERT	Acquire comprehensive textual data, commonly utilized in the domain of natural language processing	Optimize keyword extraction algorithm, use a small amount of data to generate higher resolution images from text generation.

2) Text Processing

RIFEGAN proposes a new method of rich special synthesis for limited information in T2I synthesis. This method uses attention-based subtitle matching models to select and refine the appropriate subtitles, embed the extraction features through self-attention, and use multi-subtitle attention to generate confrontation network synthetic images. Experiments have proven to significantly improve the quality of production. In addition, T2I generation models based on generating networks usually depend on the text encoder pre-trained by the image-text, which limits the acquisition of text-rich information. To this end, a method of using pre-training BERT as a text encoder is proposed. Through fine-tuning, the experimental results show that the method is better than the baseline model in quantitative and qualitative evaluation.

III. PROPOSED METHODOLOGY

This section explains the training process used for deep learning-based generative models. The approach employed Conditional Generative Adversarial Networks (GANs) in combination with Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) to generate coherent images from textual inputs. The dataset comprised flower images paired with corresponding text descriptions. To produce realistic images from text, both image and text data underwent preprocessing, including text cleaning and image resizing. Captions were extracted from the dataset, processed to build a vocabulary list, and then stored with their identifiers. The images were resized to uniform dimensions before being used as input to the proposed model.

RNNs were utilized to capture contextual relationships among words across time steps, facilitating effective text-to-image translation when paired with CNNs. The CNN component automatically identified visual features without manual annotation. Each text input was transformed into 256-dimensional word embeddings, which were then concatenated with a 512-dimensional noise vector. The model was trained using a batch size of 64 and a gated-feedback unit of 128, feeding both noise and text inputs into the generator. The architecture of the proposed system is detailed below.

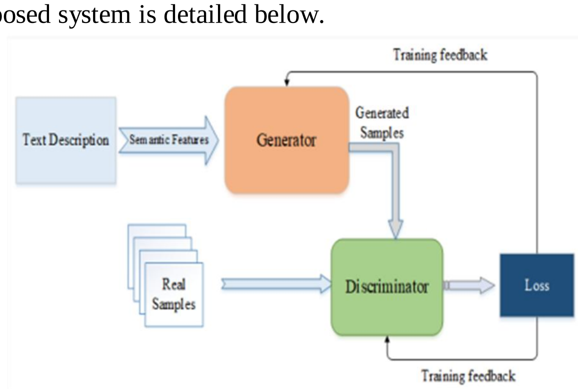

Figure 1. Architecture of the proposed method, which can generate images from text descriptions.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

The generator model leveraged the semantic content of text descriptions to map features into pixel representations, producing corresponding images. These generated images, along with real and mismatched text-image pairs from the dataset, were then input into the discriminator. The discriminator received various pairings—real images with correct captions, real images with incorrect captions, and synthetic images with real captions—to learn to distinguish between matching and non-matching examples. Real image-text pairs helped the model learn accurate alignment, while mismatched and fake pairs guided it toward differentiation.

During early training stages, the discriminator effectively classified real versus fake samples. As training advanced, computed loss values updated the model weights, refining both the generator and discriminator. Over time, the generator learned to create increasingly realistic images capable of deceiving the discriminator in distinguishing generated outputs from real ones.

IV. DATA SET AND RELATED INDICATORS

FID calculates the free distance between the distribution of synthetic images and the real image. The lower FID means that the distance between the generated image distribution and the real image distribution is closer.

R-precision is used to evaluate the visual- semantic similarity between the generating image and the corresponding text description, that is accuracy. By sorting the retrieval results between the extracted images and text features, the visual semantic similarity between the text description and the generated image is measured. If the real text of the image is described in the front R, it is related. The more closely the picture resembles the actual text description, the higher the R accuracy. IS scores are employed to assess the diversity and quality of images. First, the quality of the quality is evaluated by using the external image classifier (generally used on the ImageNet Inception-v3 network), and then use the information entropy distributed by different types of probability distribution The better, the better the diversity.

Table 4 is the evaluation indicator of the Chinese text image generation model mentioned above (based on COCO, CUB, and Oxford-102). It can be seen that PMAN (2024) combined with CLIP has very good data results in FID and IS.

Table 4: Evaluation indicators of text image generation model

Model	**COCO**			**CUB**					
	FID	R-precision/%	IS	FID	R-precision/%	IS	FID	R-precision/%	IS
SDGAN	/	75.78	35.69	/	68.76	4.67	/	/	/
Text SeGAN	/	/	/	/	/	3.65	/	/	/
PMGAN	7.89	/	34.93	10.23	/	6.36	/	/	/
ALR-GAN	29.04	69.2	24.7	15.14	77.54	4.96	/	/	/
RII-GAN	19.01	/	/	12.94	/	5.41	/	/	/
AttnGAN	35.49	85.47	25.89	23.98	67.82	4.36	/	/	/
SEGAN	32.28	/	27.86	/	/	4.67	/	/	/
ControlGAN	/	82.43	24.06	/	69.33	4.58	/	/	/
cycleGAN	/	/	/	/	/	/	/	/	/
MirrorGAN	/	74.52	26.47	/	57.67	4.56	/	57.67	/
VQA-GAN	41.7	59.25	21.92	/	/	/	/	/	/
RiFeGAN	/	/	31.7	/	22.5	5.77	/	/	4.76
BERT + StackGAN	/	/	/	37.79	/	4.44	/	/	/
Metaphor Understanding	/	/	/	/	/		/	/	/
DGattGAN	/	/	/	/	/	4.45	/	62.45	3.48
DMGAN	32.64	/	30.49	16.09	/	4.75	/	/	/
DualAttn-GAN	/	/	/	14.06	/	4.59	40.31	/	4.06
StackGAN	74.05	/	8.45	51.89	/	3.7	55.28	/	3.2
StackGAN++	81.59	/	8.3	15.3	/	4.04	/	3.26	/
HfGAN	/	22.7	27.53	/	25.3	4.48	/	30.3	3.57
KT-GAN	30.73	24.5	31.67	17.32	32.9	4.85	/	/	/

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

V. CONCLUSIONS

With the rapid development of natural language processing and computer vision, this article reviews the T2I method founded on adversarial generative networks. According to the different requirements of text -generating images, the GAN network generated based on text images is divided into three major functions: improving content authenticity, enhancing semantic correlation, and promoting content diversity. It can be seen through the data in the chart. Image generation technical performance is continuously improved effectively.

While the quality, consistency, and semantics of the picture have all significantly improved with the present technique, there are still many difficulty points and the need for application expansion. In terms of content authenticity, in many application scenarios, such as interactive game image construction and medical image analysis, it is necessary to generate fine and real image generation. In terms of semantic correlation, text image generation technology can improve the efficiency of scene retrieval, increase the ability of artificial intelligence to understand the ability to understand artificial intelligence through text interaction, and have strong theoretical research value. For example, using text to generate videos has important research value. It is one of the future research directions, but more text and video evaluation methods need to be explored.

In terms of content diversity, diversified production outputs in the fields of art and design help inspire the creators' inspiration and promote the formation of creativity. In the field of human- computer interaction, text images can be added to human-computer interaction. For example, entering simple texts to generate a rich semantic image, has increased the ability to understand artificial intelligence, giving artificial intelligence semantics "imagination" And "creativity" an effective means to study the deep learning of machines. It is hoped that the content of this article will help researchers understand the cutting-edge technologies in the field and provide a reference for further research.

REFERENCES

[1] Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.

[2] Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.

[3] Liu, R.; Ge, Y.; Choi, C.L.; Wang, X.; Li, H. Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16377–16386.

[4] Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5907–5915.

[5] Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE Trans. Pattern Anal. Mach. Intell. **2018**, 41, 1947–1962. [CrossRef] [PubMed]

[6] Zhang, Z.; Xie, Y.; Yang, L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6199–6208.

[7] Cai, Y.; Wang, X.; Yu, Z.; Li, F.; Xu, P.; Li, Y.; Li, L. Dualattn-GAN: Text to image synthesis with dual attentional generative adversarial network. IEEE Access **2019**, 7, 183706–183716. [CrossRef]

[8] Kosslyn, S.M.; Ganis, G.; Thompson, W.L. Neural foundations of imagery. Nat. Rev. Neurosci. **2001**, 2, 635–642. [CrossRef] [PubMed]

[9] Zhang, Y.; Han, S.; Zhang, Z.; Wang, J.; Bi, H. CF-GAN: Cross-domain feature fusion generative adversarial network for text-to-image synthesis. Vis. Comput. **2022**, 1–11. [CrossRef]

[10] Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6.

[11] Kim, P. Convolutional neural network. In MATLAB Deep Learning; Springer: Berlin/Heidelberg, Germany, 2017; pp. 121–147. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. arXiv **2014**, arXiv:1406.2661. [CrossRef]