Research on Text-to-Image Generation Method Based on GAN

Yinuo Liu[a]
College of Information Engineering, Northwest A&F University, Shaanxi, 712100, China

Keywords: GAN, Text-to-Image Generation Method, Cycle Consistency.

Abstract: The main task of text-to-image is to generate images that are true and clear according to the text content. As one of the representative methods, generative adversarial network (GAN) occupies an important position in the implementation of text images. According to the different requirements of text image generation, the GAN network generated based on text images is divided into three major functions: improving content authenticity, enhancing semantic correlation, and promoting content diversity. In response to the above needs, this article analyzes the authenticity of the content from the perspective of improving the quality, fine particle size enhancement, contextual enhancement, and dynamic adjustment of the content of the stack structure, analyzed the authenticity of the content from the stack structure. The perspective of extraction, semantic layout, and cycle consistency analyzes enhanced semantic correlation function and analyzes the diversity of content diversity from the perspective of training mechanisms and text processing. The thesis focuses on predecessors' representative methods' basic process and design ideas. The predecessor method is used to compare and analyze the predecessor methods through the existing data set. Forecasting and prospects will help researchers to further promote this field.

1 INTRODUCTION

In recent years, the rapid development of deep learning has promoted the innovation and application of natural language processing, computer vision, and image generation technology. Text image generation occupies an important position in this field. Using natural language text as input to generate corresponding images shows a powerful multimodal interactive ability. In 2024, Ajay Kumar et al. AI text is used to transform the image tools Dall E2 and Midjourney to generate unique facial signs related to medical conditions, such as thyroid dysfunction and Hoven syndrome. These tools generate high-quality images based on professional medical texts, solving confidentiality problems in the use of traditional patients. In the future, this technology is expected to find new application scenarios in more fields.

Text image generation can be understood as text-to-image (T2I) style migration, which aims to generate realistic images that meet the requirements of a given text. The original TextCNN core idea is to apply the convolutional neural network (CNN) to text classification to extract text characteristics (Kim, 2014). To handle more complicated tasks, deep convolutional neural network DCNN adds more convolutional layers and full connection layers based on CNN to learn more abstract and higher-level features (Atwood and Towsley, 2015). However CNN-based text images have the problem of generating efficiency and low accuracy. After that, VAE is known for stable training and continuous potential space generation capabilities. It is suitable for image reconstruction and noise, but there are still problems with low -generating image resolution. With the complexity of text description and the demand for image generation, generative adversarial network (GAN) generates continuously developing. GAN not only has advantages in generating high-quality and high-resolution images but also its generator and identifier game learning. To sort out more clear development ideas, the functions in the process of generating the development of text images have been continuously improved, and GAN is divided into three stages.

The first is to improve the authenticity of the content. The primary task of image generation focuses on generating the authenticity of the image, as is text image generation. Cross-domain feature fusion to generate a confrontation network (CF-

[a] https://orcid.org/my-orcid?orcid=0009-0009-8499-1991

Liu and Y.
Research on Text-to-Image Generation Method Based on GAN.
DOI: 10.5220/0013510500004619
In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, pages 124-130
ISBN: 978-989-758-754-2

GAN) as a type of network framework improves the residual module to facilitate the full extraction features (Zhang, Han, and Zhang, et al., 2022). Feature fusion enhanced response modules (FFEM) and multi-branch residual modules (MBRM) are refined to the generated images in a deep fusion.

The second is to enhance semantic correlation. To attain the superior grade of image creation, it not only needs to improve the authenticity of generating images but also emphasizes the consistency of generating images and text semantics in T2I. Considering the limitations of generating detailed description objects, PMGAN emphasizes the consistency of generating images and text semantics (Yu, Yang, and Xing, 2024). The text's characteristics are extracted by this model using the CLIP text encoder, and the model maps the text and images to a common semantic space to ensure that the generating image can fully reflect the rich information in the text input.

After solving the authenticity and semantic correlation of the image, you also need to pay attention to the diversity of generation. Compared with the early GAN, RiFeGAN has rich text descriptions based on attention subtitle matching models, selecting and refining candidate subtitles (Cheng, Wu, and Tian, et al., 2020). This model uses a self-attention embedding mixture to extract features and uses multiple subtitles to generate combined network synthetic images, thereby effectively improving the diversity of generating content.

This article first elaborates on the significance of text image generation which is the theoretical research value and application value of text image generation, and then proposes a new classification method based on the development stage of the adversarial T2I generation in terms of functional improvement, and discussed it at the same stage and different stages of the consideration, performance indicators, and application areas, then finally summarized and looked forward to its next research.

2 T2I GENERATION MODEL CLASSIFICATION BASED ON FUNCTION

2.1 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.

2.1.1 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 (Zhang, Xu, and Li, et al., 2018). 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 more steady 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.

2.1.2 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 (Zhang, Han, and Zhang, et al., 2022). 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.

2.1.3 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 (Zhang, Zhu, and Yang, et al., 2021). 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.

2.1.4 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 (Zhu, Pan, and Chen, et al., 2019). 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.

2.2 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.

2.2.1 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) (Tan, Liu, and Liu, et al., 2019). 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.

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	

2.2.2 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 (Yu, Yang, and Xing, 2024). 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.

2.2.3 Semantic Commonality Extraction

Zhou and others proposed a single-level network TCF-GAN to generate a highly detailed image from the text (Zhou, Wu, and Ye, et al., 2024). 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.

2.2.4 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 (Yuan, Zhu, and Yang, et al., 2024). 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.

2.2.5 Cyclic Consistency

Qiao and others proposed the Mirrorgan framework to improve T2I (T2I) generation (Qiao, Zhang, and Xu, et al., 2019). 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.

2.3 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.

2.3.1 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 (Cha, Gwon, and Kun., 2019). 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 promotes 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.

2.3.2 Text Processing

RIFEGAN proposes a new method of rich special synthesis for limited information in T2I synthesis (Cheng, Wu, and Tian, et al., 2020). 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 (Na, Do, and Yu, et al., 2022). 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.

3 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 (Szegeedy, Vanhoucke, and Ioffe., 2016).

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	/	/	/

4 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

Atwood J., Towsley D. 2015. Diffusion-Convolutional Neural Networks. arXiv E-Prints, arXiv:1511.02136.

Cha M., Gwon Y. L., Kung H. T. 2018. Adversarial Learning of Semantic Relevance in Text to Image Synthesis. arXiv E-Prints, arXiv:1812.05083.

Cheng J., Wu F., Tian Y., Wang L., Tao D. 2020. RiFeGAN: Rich Feature Generation for T2I Synthesis From Prior Knowledge. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10908–10917.

Kim Y. 2014. Convolutional Neural Networks for Sentence Classification. arXiv E-Prints, arXiv:1408.5882.

Na S., Do M., Yu K., Kim J. 2022. Realistic image generation from text by using BERT-based embedding. Electronics, 11(5), 764.

Qiao T., Zhang J., Xu D., Tao D. 2019. MirrorGAN: Learning T2I Generation by Redescription. arXiv E-Prints, arXiv:1903.05854.

Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. 2016. Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2818–2826.

Tan H., Liu X., Liu M., Yin B., Li X. 2021. KT-GAN: Knowledge-Transfer Generative Adversarial Network for T2I Synthesis. IEEE Transactions on Image Processing, 30, 1275–1290.

Yu Y., Yang Y., Xing J. 2024. PMGAN: pretrained model-based generative adversarial network for T2I generation. The Visual Computer.

Yuan H., Zhu H., Yang S., Wang Z., Wang N. 2024. RII-GAN: Multi-scaled aligning-based reversed image interaction network for T2I synthesis. Neural Processing Letters, 56(1).

Zhang H., Xu T., Li H., Zhang S., Wang X., Huang X., Metaxas D. N. 2019. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8), 1947–1962.

Zhang H., Zhu H., Yang S., Li W. 2021. DGattGAN: Cooperative Up-Sampling Based Dual Generator Attentional GAN on T2I Synthesis. IEEE Access, 9, 29584–29598.

Zhang Y., Han S., Zhang Z., Wang J., Bi H. 2022. CF-GAN: cross-domain feature fusion generative adversarial network for T2I synthesis. Vis. Comput., 39(4), 1283–1293.

Zhou H., Wu T., Ye S., Qin X., Sun K. 2024. Enhancing fine-detail image synthesis from text descriptions by text aggregation and connection fusion module. Signal Processing: Image Communication, 122, 117099.

Zhu M., Pan P., Chen W., Yang Y. 2019. DM-GAN: Dynamic Memory Generative Adversarial Networks for T2I Synthesis. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5795–5803.