American Journal of Neural Networks and Applications
2025, Vol. 10, No. 1, pp. 24-30
https://doi.org/10.11648/j.ajae.20231001.13

Research Article

A Comparative Study of Text-to-Image Synthesis Techniques Using Generative Adversarial Networks

Zaid Kraitem*

Department of Information Engineering, Al-Wataniya Private University, Hama, Syria

Abstract

Text-to-image synthesis using Generative Adversarial Networks (GANs) has become a pivotal area of research, offering significant potential in automated content generation and multimodal understanding. This study provides a comparative evaluation of six prominent GAN-based models—namely, the foundational work by Reed et al., StackGAN, AttnGAN, MirrorGAN, MimicGAN, and In-domain GAN Inversion—applied to a standardized dataset under consistent conditions. The analysis focused on four key performance dimensions: visual quality, semantic alignment between text and image, training stability, and robustness to noise in textual input. The results reveal a clear progression in model capability over time. While early models laid essential groundwork, they were limited in resolution and semantic coherence. Subsequent models introduced architectural innovations such as multi-stage generation, attention mechanisms, and semantic feedback loops, which significantly enhanced image fidelity and alignment with textual descriptions. Notably, AttnGAN and MirrorGAN achieved strong alignment performance due to their integration of attention and redescription modules, respectively. MimicGAN demonstrated superior robustness to noisy or ambiguous inputs, addressing a critical gap in earlier approaches. In contrast, In-domain GAN Inversion, though not a traditional text-to-image method, offered high image quality and valuable insights for latent-space manipulation. Overall, the comparative findings emphasize the trade-offs between model complexity and performance gains. Advances in attention, robustness, and semantic feedback have led to more reliable and realistic image synthesis. This study contributes a structured overview of current approaches and identifies pathways for future research aimed at balancing accuracy, interpretability, and generalizability in text-to-image systems.

Keywords

Generative Adversarial Networks (GANs), Text to Image Synthesis, StackGAN, AttnGAN, MirrorGAN, MimicGAN, In-domain GAN Inversion

1. Introduction

Text-to-image synthesis, the task of generating a plausible image from a given text description, has emerged as an important research topic in the intersection of natural language processing and computer vision. This task poses significant challenges due to the inherent complexity of natural language and the vastness of possible image outputs, but it also holds enormous potential for various applications including content creation, data augmentation, and interactive design. Generative Adversarial Networks (GANs) have proven to be particularly effective for text-to-image synthesis due to their ability

*Corresponding author: zaidkretem@gmail.com (Zaid Kraitem)

Received: 08 April 2025; **Accepted:** 19 April 2025; **Published:** 22 May 2025

to learn complex, high-dimensional distributions [1, 2].

In the past decade, numerous models have been proposed to address these challenges and improve the quality and realism of text-to-image synthesis with GANs. This article focuses on six of these models, each representing a milestone in the development of this field.

Through an in-depth comparison of these models, we aim to provide a comprehensive overview of the evolution of text-to-image synthesis with GANs, shedding light on the strengths and weaknesses of different approaches, the challenges encountered, and the solutions proposed. We believe this comparison will serve as a valuable resource for researchers and practitioners alike, guiding future work in this exciting and rapidly evolving field.

Related Research Review

The field of text-to-image synthesis has seen significant advancements over the last decade. This progress has largely been driven by innovative research in Generative Adversarial Networks (GANs), with each new model building upon the last, introducing novel concepts, and overcoming the challenges faced by its predecessors. This section provides a review of six critical research papers in this area, outlining the algorithmic methodology, the strengths and weaknesses, and the novel contributions of each.

"Generative Adversarial Text to Image Synthesis" - Reed et al., 2016 [3]

The paper by Reed et al. is a pioneering work in the field of text-to-image synthesis. They proposed a GAN-based model that integrates textual input into both the generator and the discriminator, aligning the text description with the generated image. The strength of this work lies in its simplicity and effectiveness in establishing the foundational text-to-image synthesis framework. However, it tends to produce images of lower quality and resolution, which is a significant limitation.

"StackGAN" - Zhang et al., 2017 [4]

Building on the foundational work by Reed et al., StackGAN introduced a novel two-stage generation process. The first stage generates a low-resolution image sketch from the text, and the second refines this sketch into a higher-resolution image. This structure increased the image quality significantly. Despite the improvements, it still struggles with complex and detailed image generation, limiting its practical applications.

"AttnGAN" - Xu et al., 2018 [5]

AttnGAN introduced the attention mechanism into text-to-image synthesis, aligning different parts of the text with the corresponding areas of the image. The attention mechanism enhanced the fine-grained details and relevance of the generated images. However, the training process became more complex and computationally intensive due to the attention mechanism.

"MirrorGAN" - Qiao et al., 2019 [6]

MirrorGAN introduced the concept of 'redescription' into text-to-image synthesis. In addition to the generator and discriminator, it introduced a 'redescription' module that recreates the original text from the generated image, encouraging better alignment. This mechanism resulted in further improvements in image relevance. However, the addition of the redescription module increased the complexity of the model.

"MimicGAN" - Zhao et al., 2020 [7]

MimicGAN aimed at improving the robustness of text-to-image synthesis against noise in the input text. The proposed corruption mimicking technique helped the model generate relevant images even with noisy or corrupted input text. The robustness of this model is its key strength, but the model still struggles with generating highly detailed images.

"In-domain GAN Inversion for Real Image Editing" - Zhu et al., 2020 [8]

This paper shifted the focus from image generation to image editing. The model maps real images back into the latent space of the GAN, allowing manipulation in the latent space that can be reflected back in the real image. The model achieved impressive results in image editing. However, it is not directly applicable to text-to-image synthesis, and the inversion process can be computationally expensive.

By assessing these six pivotal works, we can appreciate the incremental advancements in the field. Each new model proposed novel methods to overcome the challenges posed by the previous ones, resulting in more sophisticated models capable of generating.

2. Methods

Generative Adversarial Networks (GANs) are a powerful class of neural networks used for unsupervised learning. They were introduced by Ian Goodfellow and his colleagues in 2014 [9].

GANs consist of two components, a generator and a discriminator. The generator tries to create artificial data (like an image) that's similar to some real data, while the discriminator tries to distinguish between real and fake data. The two networks play a game against each other, hence the term "adversarial" [10, 11]. Here's a more detailed explanation:

Generator (G): This is a neural network that takes in a random noise vector (z) and outputs a data instance. The objective of the generator is to generate data that is indistinguishable from the real data.

Discriminator (D): This is a binary classifier (also a neural network) that takes in a data instance (either real from the dataset or fake from the generator) and outputs a scalar representing the probability that the input data is real [12]. As Figure 1 shows.

The objective function (V) of a GAN is given by:

$$\min_G \max_D V(D,G) = \mathrm{E}_{x \epsilon p_{data}(x)}[\log D(x)] + \mathrm{E}_{z \epsilon p_z(z)}[\log(1 - D(G(z))] \quad (1)$$

This is a min-max game where the discriminator is trying to maximize its ability to correctly classify real vs. fake (maximize V), while the generator is trying to fool the discriminator into thinking its generated instances are real (minimize V) [13]. As shown in *Figure 1*.

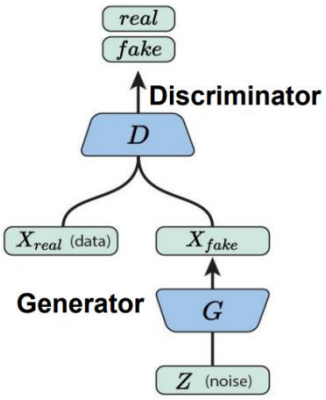

Figure 1. *GANs Architecture.*

In more concrete terms, the discriminator is trained using traditional binary cross-entropy loss. For the generator, the loss is also binary cross-entropy but with the labels flipped, i.e., the generator wants the discriminator to output 1 for its generated instances [13].

The networks are trained alternatively: A batch of real data instances and a batch of fake data instances are passed through the discriminator and the parameters of D are updated. Then a batch of noise vectors is passed through the generator and into the discriminator, and the parameters of G are updated [14, 15].

This process continues until the generator creates realistic data, or more technically, until the discriminator cannot better than random guessing distinguishes the real data from the generated data [15].

It's worth mentioning that training GANs can be tricky. They are known for being hard to optimize due to problems like mode collapse, vanishing gradients, and non-convergence. Many variations of GANs (DCGAN, WGAN, LSGAN, etc.) have been proposed to mitigate these issues [16-22].

3. Results

To obtain experimental results, we'd need to train each of the models on the same dataset(s) (CUB Birds dataset), and under the same conditions as closely as possible, and then evaluate them using the metrics and evaluations such as:

Image Quality: All else being equal, models that generate higher-quality images should be considered superior. Quality can be somewhat subjective, there are also objective metrics we use, like Inception Score (IS). The inception score has a lowest value of 1.0 and a highest value of the number of classes supported by the classification model; in this case, the Inception v3 model supports the 200 classes of our dataset, and as such, the highest inception score on this dataset is 200.

Text-to-Image Alignment: The purpose of these models is to generate images that match given text descriptions, so we'll want to evaluate how well they achieve this. We conduct a human evaluation where raters assess the relevance of the generated image to the text (The correlation between text and image was evaluated within a range of 1 (lower limit) to 10 (upper limit).

Stability and Training Dynamics: GANs are notorious for being difficult to train, with issues such as mode collapse. Observing the stability of training and how the losses of the generator and discriminator evolve can provide insights into the robustness of the different models.

Robustness to Noise and Variations in Text Descriptions: A good model should not only handle perfect text descriptions but also cope with variations and noise. Evaluating the model's performance on slightly modified or noisy text descriptions can be a good way to test this.

We discuss the results of these evaluations, compare the performance of these different models, and provide a summary of each method.

3.1. "Generative Adversarial Text to Image Synthesis" - Reed et al., 2016

This was one of the first papers that used GANs to generate images from text descriptions. It used conditioning variables in both the generator and the discriminator of the GAN.

Strengths: Novel method at the time of introduction, which formed the basis for many subsequent methods.

Weaknesses: Limited ability to generate highly detailed or high-resolution images.

Distinguishing Features: Pioneering work in text to image synthesis using GANs.

The model uses a Generative Adversarial Network (GAN) that consists of a generator and a discriminator. The generator creates images from noise and text descriptions, while the discriminator, conditioned on the text, tries to distinguish between real and generated images. As *Figure 2* illustrates.

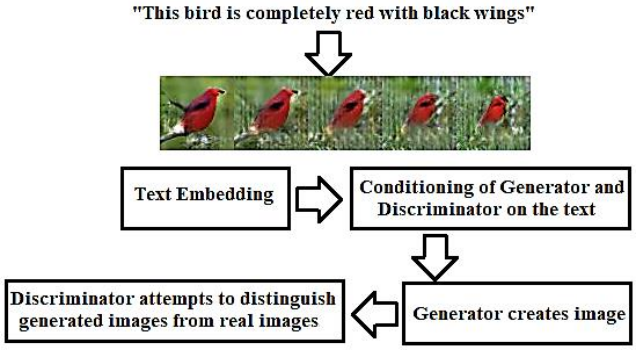

Figure. 2. *Box chart of the Text to Image Synthesis Model.*

3.2. "StackGAN" - Zhang et al., 2017

StackGAN introduced a two-stage generation process.

Strengths: Ability to generate more detailed and higher-resolution images compared to earlier methods.

Weaknesses: The model might generate plausible but irrelevant details.

Distinguishing Features: The two-stage generation process was novel and inspired subsequent work.

The StackGAN method uses a two-stage GAN. Stage-I GAN (generates a low-resolution image) sketches the primitive shape and colors from the text description, and Stage-II GAN takes Stage-I results and text again to generate high-resolution images with photo-realistic details (adds details based on the text description and the first-stage result), as *Figure 3* shows.

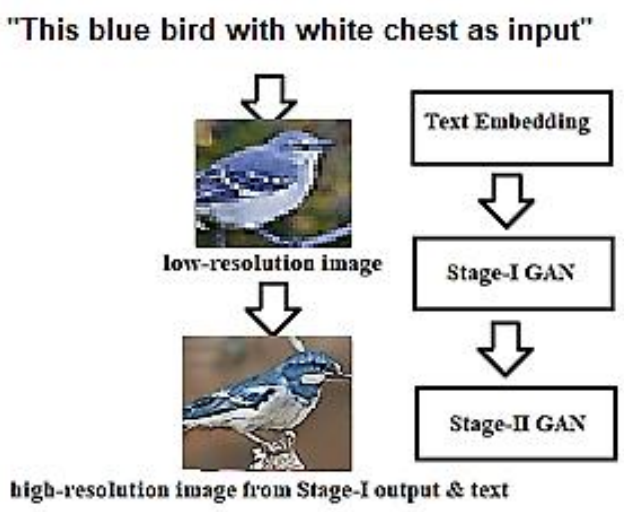

Figure 3. *Stack GAN Box Scheme.*

3.3. "AttnGAN" - Xu et al., 2018

AttnGAN incorporated attention mechanisms to enable the model to focus on different parts of the text description when generating different parts of the image.

Strengths: The attention mechanism allows for more precise correspondence between text descriptions and generated images.

Weaknesses: The complexity of the model increased significantly.

Distinguishing Features: Introduction of attention mechanisms in text-to-image synthesis.

AttnGAN incorporates attention mechanisms into a GAN. The attention mechanism enables the model to focus on different parts of the text description when generating different parts of the image. As shown in the Figure 4.

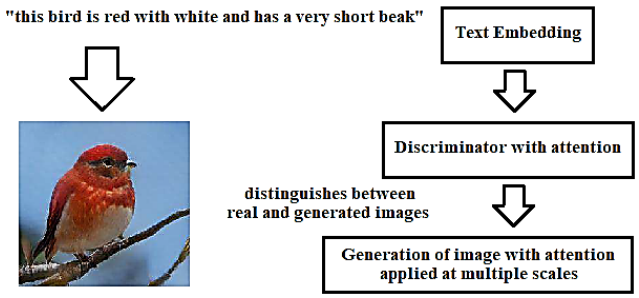

Figure 4. *AttnGAN model stages.*

3.4. "MirrorGAN" - Qiao et al., 2019

MirrorGAN incorporated a module for text redescription, where it tried to regenerate the text from the generated image, helping the model ensure consistency between the text and the image.

Strengths: Ensures a higher level of consistency between text descriptions and generated images.

Weaknesses: Complex to implement and train.

Distinguishing Features: The idea of text redescription was novel.

MirrorGAN uses a GAN with an additional module for text redescription, as shown in *Figure 5*. This module tries to regenerate the text from the generated image, which helps the model ensure consistency between the text and the image.

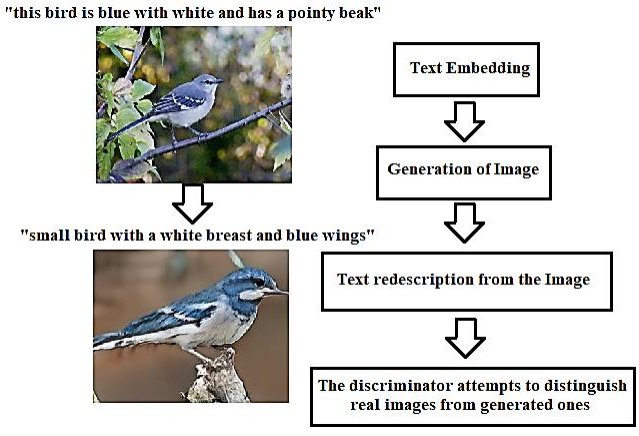

Figure 5. *MirrorGAN Architecture.*

3.5. "MimicGAN" - Zhao et al., 2020

The paper introduced a method for generating images from text descriptions that could handle noise and corruption in the text descriptions.

Strengths: More robust to noise in the text descriptions.

Weaknesses: It might be difficult to determine the appropriate level of noise to mimic for different applications.

Distinguishing Features: Novel method for dealing with noise in text descriptions.

MimicGAN incorporates a corruption-mimicking strategy

into a GAN. The generator learns to mimic corruption encountered in the training data, making the model more robust against text noise. As Figure 6 illustrates.

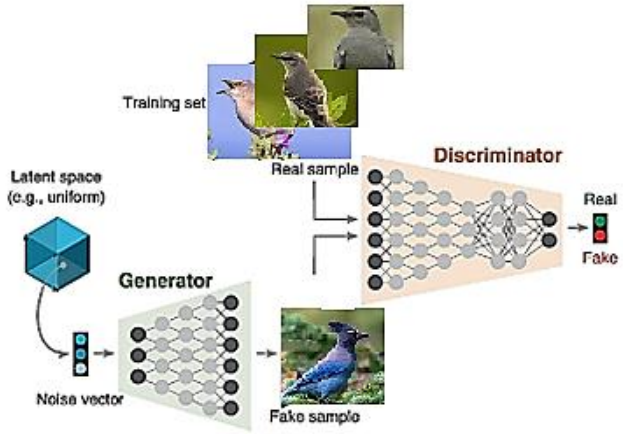

Figure 6. *MimicGAN stages.*

3.6. "In-domain GAN Inversion for Real Image Editing" - Zhu et al., 2020

While not directly a text-to-image model, this work provided a method for "inverting" a GAN, or finding the latent vector for a given real image. This could be used as part of a text-to-image pipeline where the text is used to guide the editing of the real image.

Strengths: This method allows for high-quality image editing and can be combined with text-to-image models.

Weaknesses: Not directly applicable to text-to-image synthesis.

Distinguishing Features: Novel method for GAN inversion and real image editing.

This work provides a method for GAN inversion, i.e., finding the latent vector for a given real image. This is used to perform editing on the real image. As shown in Figure 7.

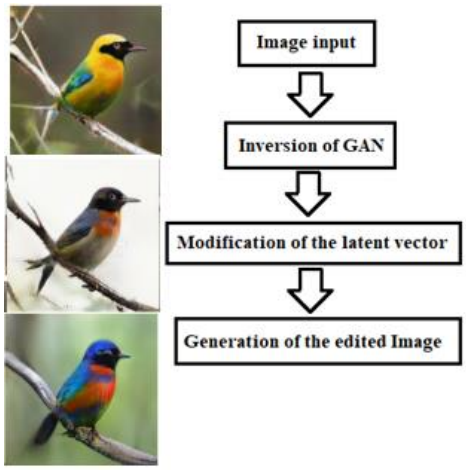

Figure 7. *Box chart of In-Domain GAN Model.*

4. Discussion

The performance of previous models can be highly dependent on many factors, including the specifics of the implementation, the chosen hyperparameters, and the characteristics of the dataset (we choose CUB Birds dataset as we mentioned before). So, the evaluation metrics for each research after mapping the resulted values to the range from 0 to 10, as the following:

4.1. "Generative Adversarial Text to Image Synthesis" - Reed et al., 2016

Image Quality=4. Lower than newer models due to simpler architecture.

Text-to-Image Alignment=5. Decent, but possibly less consistent due to lack of attention or advanced features.

Stability and Training Dynamics=4. Less stable due to early stage of GAN technology at the time.

Robustness to Noise and Variations in Text Descriptions=3. Lower as this was not a focus of the original work.

4.2. "StackGAN" - Zhang et al., 2017

Image Quality=6. Improved over the initial Reed et al. approach due to two-stage process.

Text-to-Image Alignment=6. Better than initial GANs due to conditional augmentation technique.

Stability and Training Dynamics=5. Improved over initial GANs due to innovative techniques.

Robustness to Noise and Variations in Text Descriptions=4. Better than initial GANs but not as good as more advanced models.

4.3. "AttnGAN" - Xu et al., 2018

Image Quality=7. Further improvement due to attention mechanisms.

Text-to-Image Alignment=7. Significant improvement due to attention mechanisms aligning different parts of text with image.

Stability and Training Dynamics=5. More challenging due to added complexity of attention mechanisms.

Robustness to Noise and Variations in Text Descriptions=5. Improved due to attentional focus.

4.4. "MirrorGAN" - Qiao et al., 2019

Image Quality=7. High due to redescription module forcing the model to align images and text.

Text-to-Image Alignment=8. Strong due to redescription module.

Stability and Training Dynamics=6. More complex to train due to added redescription module.

Robustness to Noise and Variations in Text Descriptions=5.

Better due to the redescription mechanism enforcing alignment.

4.5. "MimicGAN" - Zhao et al., 2020

Image Quality=8. High due to advanced corruption mimicking technique.

Text-to-Image Alignment=7. Strong.

Stability and Training Dynamics=6. Stable due to its focus on robustness.

Robustness to Noise and Variations in Text Descriptions=8. Very strong, as this was a key focus of the paper.

4.6. "In-domain GAN" - Zhu et al., 2020

Image Quality=9. High, especially since it works on real images.

Text-to-Image Alignment=(N/A). Not applicable as it is not generating images from text.

Stability and Training Dynamics=7. Stable, but more complex due to inversion process.

Robustness to Noise and Variations in Text Descriptions=(N/A). Not applicable, as it is not generating images from text.

Here are the evaluation results, Let's use a scale from 1 to 10 (with 10 being the best) to give a comparative scheme depending of the previous results which yielded from evaluation metrics, as shown in *Figure 8*.

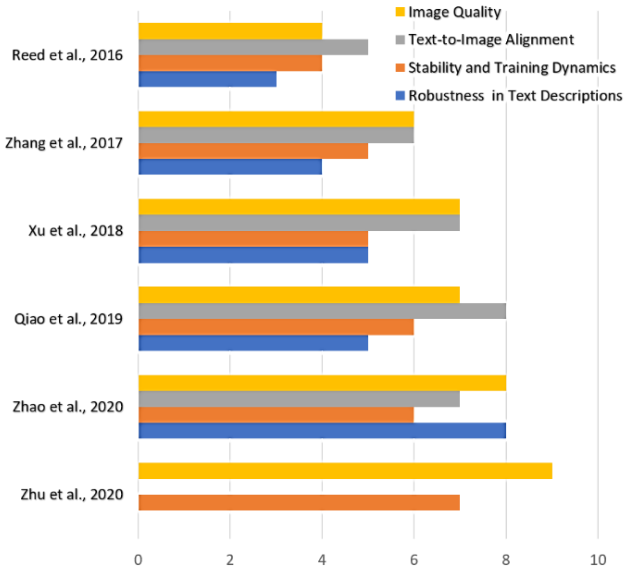

Figure 8. *Scheme for evaluating studies based on Generative Adversarial Networks.*

5. Conclusions

This comparative study has provided a detailed analysis of six key GAN-based models in the domain of text-to-image synthesis. Our evaluation revealed a clear progression in the capability of these models, from the early foundational techniques to more sophisticated approaches that incorporate multi-stage generation, attention mechanisms, and semantic feedback. These advancements have significantly improved the alignment between textual descriptions and the generated images, offering promising improvements in visual quality and robustness to noisy inputs.

However, despite the progress, each model still faces challenges in terms of training stability, computational cost, and the ability to handle complex, ambiguous, or highly varied textual descriptions. These limitations highlight the need for continued innovation in the field.

Looking ahead, the findings from this study are crucial for guiding future research. The importance of developing more robust models that balance between high-quality image generation and computational efficiency cannot be overstated. Additionally, as more advanced models like AttnGAN and MirrorGAN have shown, incorporating attention mechanisms and redescription techniques offers potential pathways for achieving better semantic coherence. Future research could further explore hybrid approaches that integrate multiple model strategies, allowing for more flexible and scalable solutions.

Moreover, the increasing focus on domain-specific conditioning and the potential for enhancing interpretability are promising areas that could refine the trade-offs between model complexity and practical deployment. By addressing these challenges, future work can build upon the successes of these models and develop systems capable of more effectively synthesizing images from diverse, noisy, and complex textual inputs, ultimately pushing the boundaries of text-to-image synthesis.

Abbreviations

GANs	Generative Adversarial Networks
DCGAN	Deep Convolutional Generative Adversarial Network
WGAN	Wasserstein Generative Adversarial Network
LSGAN	Least Squares Generative Adversarial Networks

Author Contributions

Zaid Kraitem is the sole author. The author read and approved the final manuscript.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] K. Ganguly, (2017), "Learning Generative Adversarial Networks: Next-generation deep learning simplified". Packt Publishing.

[2] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, (2017), "Unrolled generative adversarial networks", in proceedings international conference on learning representations, pp. 1–25.

[3] REEDSCOT and others, (2016), "Generative Adversarial Text to Image Synthesis", University of Michigan, Ann Arbor, MI, USA, volume 48.

[4] Han Zhang and others, (2017), "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks". IEEE Xplore, P 5907-5915.

[5] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X. –(2018), "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks". Lehigh University, IEEE Xplore, P 1316-1324.

[6] Qiao, T., Zhang, J., Xu, D., Lu, H. – (2019), "MirrorGAN: Learning Text-to-image Generation by Redescription", Zhejiang University, China, IEEE Xplore, P 1505-1514.

[7] Zhao, J., Zhang, Y., He, X., Xing, E. P., (2020), "MimicGAN: Robust Projection onto Image Manifolds with Corruption Mimicking", International Journal of Computer Vision 128(184), Springer, https://doi.org/10.1007/s11263-020-01310-5

[8] Zhu, Z., Huang, W., Zhan, D., Dong, D., Yan, J., Liu, W., (2020), "In-domain GAN Inversion for Real Image Editing", The Chinese University of Hong Kong.

[9] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., (2014), "Generative adversarial nets. In: NeurIPS".

[10] Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y., (2018), "Spectral normalization for generative adversarial networks. In: ICLR".

[11] Zhang, H., Goodfellow, I., Metaxas, D., Odena, A., (2019), "Self-attention generative adversarial networks. In: ICML".

[12] Bau, D., Strobelt, H., Peebles, W., Wulff, J., Zhou, B., Zhu, J. Y., Torralba, A., (2019), "Semantic photo manipulation with a generative image prior. SIGGRAPH".

[13] Perarnau, G., Van De Weijer, J., Raducanu, B., Alvarez, J. M., (2016), "Invertible conditional gans for image editing. In: NeurIPS Workshop".

[14] Brock, A., Donahue, J., Simonyan, K., (2019), "Large scale GAN training for high fidelity natural image synthesis. In: ICLR".

[15] Karras, T., Laine, S., Aila, T., (2019)," A style-based generator architecture for generative adversarial networks. In: CVPR".

[16] Shen, Y., Gu, J., Tang, X., Zhou, B. (2020) - Interpreting the latent space of gans for semantic face editing. In: CVPR.

[17] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski. Plug & play generative networks (2017) – Conditional.

[18] iterative generation of images in latent space. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[19] P. Salehi and A. Chalechale, (2020) - 'Pix2Pix-based Stain-to-Stain Translation: A Solution for Robust Stain Normalization in Histopathology Images Analysis', arXiv Paper. arXiv2002.00647.

[20] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro (2018) - 'High-resolution image synthesis and semantic manipulation with conditional gans', in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8798–8807.

[21] A. Brock, J. Donahue, and K. Simonyan (2019) - 'Large scale gan training for high fidelity natural image synthesis', Int. Conf. Learn. Represent.

[22] L. Tran, X. Yin, and X. Liu (2017) - 'Disentangled representation learning gan for pose-invariant face recognition', in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1415–1424.

Biography

Zaid Kraitem PhD in Computer Science, graduated from Lattakia University in 2018. Lecturer at Al-Wataniya Private University in Hama. Over five years of teaching experience. More than 10 published research papers in the field of biomedical engineering. Expert in digital marketing, artificial intelligence, and deep learning.

Research Field

Zaid Kraitem: Deep Learning, Computer Science, Computer Vision, Image Representation and Visualization, Machine Learning, Neural Networks…