# Trends in Text to Image Synthesis (T2I) using Generative Adversarial Networks

Dr.Venkatesan R[1], Priyanka S[2]
{vr.cse@psgtech.ac.in[1], priyankasukumar26@gmail.com[2]}

Professor[1], PG Scholar[2], Dept. of Computer Science and Engineering,
PSG College of Technology, Coimbatore, India

**Abstract:** The two most evident modalities of humans are language and vision. Any system that aids interaction between human beings and Artificial Intelligence (AI) is rooted upon these two. Text-to-Image synthesis (T2I) powered by Natural Language Processing (NLP) and deep Generative Adversarial Networks (GANs) replicates this phenomenon. The logical relationship between semantics and vision guides T2I, that attempts to translate highly detailed natural language textual descriptions to pixel-level details. The human concept of attention is leveraged and conceptualized by deep attentional multi-layered GANs. Mimicking the human thinking processes of visualizing the scenes in mind while speaking and listening can be extensively used in various AI applications that craves brain-like comprehending potency. The advancement of a multitude of GANs that focused on semantic consistency, high-resolution photo-realistic images and diversity in synthesis has been investigated in this article.

**Keywords:** Generative Adversarial Networks, deep attentional, multi-layered, natural language textual descriptions, photo-realistic, semantic consistency, Text-to-Image synthesis

## 1. Introduction

With the advent of Generative Adversarial Networks (GANs) (Good fellow et al., 2014), text-to-image synthesis (T2I) has extensively progressed by successfully applying it to natural language processing and computer vision. T2I bridges the semantic gap by generating photorealistic images that incorporates visual realism. Generating images that are indistinguishable from real images in accordance with the natural language textual descriptions which have fine grained semantics is achieved. It requires the system to possess human-like intelligence in order to understand the rich lexical grounding and synthesize images like how the human brain visualizes it. T2I has diverse use cases ranging from photo-editing, pre-design visualization in construction or products in computer-aided design, realistic teaching environment, brainstorming new concepts and crime scene investigation. The ubiquitous usability of T2I instigated this review article.

## 2. Literature Survey

A deep architecture GAN formulation translating human-written single sentence descriptions to image pixels is pioneered by the matching-aware GAN-INT-CLS (Reed et al.,

ICCAP 2021, December 07-08, Chennai, India

DOI 10.4108/eai.7-12-2021.2315122

2016) for Deep Convolutional Generative Adversarial Network (DC-GAN). The discriminative and generalizable yields of deep convolutional and recurrent networks for texts in the zero-shot learning sense text representations lead to this work of learning to capture the most important visual details as a text feature representation and using them to synthesize a compelling realistic image. Multimodality is a very natural application for GANs, the deep symmetric structured joint embedding for textual descriptions of visuals for robustness is used. This is the first ever original differentiable architecture that translates from the character level to pixel level in an end-to-end fashion. Style transfer and content matching as per the descriptions have been untwined for the bird poses and atmosphere transfer. Images conditioned on text are strikingly multimodal with plentiful credible configurations of pixels that rightly depict the description. Higher resolution images for diverse textual representations are a potential extension. The inclusion of object location and pose with captions in T2I is addressed by Generative Adversarial What-Where Network (GAWWN). GAWWNs (Reed et al., 2016) synthesizes 128 X 128 resolution images based on informal text descriptions and object location provided what content to be drawn in which location. GAWWNs are bounding-box and key point-conditional models. The bird locations are controlled by the former while the individual part locations are addressed by the latter. Generation of high-resolution images are backed by decomposition of the entire task into easier sub problems which does not constraint the test time also. An unsupervised or weakly supervised saturated architectural advancements for improved text-to-human image synthesis performance can be explored.

It is observed that starting with low-resolution images, GANs then progressively increase the resolution by adding layers in an incremental fashion that allows the training to first discover large-scale structure of the image distribution and later paying attention to finer scale details rather than learning all scales simultaneously (Karras et al., 2017). In Progressive GANs (ProgGANs) batch normalization eliminated covariate shifts that handled the GANs prone to signal magnitude escalation due to unhealthy competition between them. The actual issue in constraining signal magnitudes and competition is addressed here with two non-learnable parameters. Healthy progressive training of both generator and discriminator by equalized learning rate - Ensures dynamic range and same learning speed for all weights by deviating from careful weight initialization at the beginning by scaling them explicitly during run time; pixelwise feature vector normalization in generator: prevents spiraling out of magnitudes by normalizing the feature vector in each pixel to unit length in the generator after each convolutional layer. The training time could have been reduced and generation of smaller images is considerably stable because there is little class information and limited modes.

StackGAN (Zhang et al., 2017) with two stages made the image synthesis into a more manageable subproblem. Conditioning augmentation technique activates multitudinous smoothening in the inherent conditioning. Stacked-GAN for high resolution image synthesis with a Conditioning Augmentation stabilized GAN training encourages diversity of generated samples. Stage-I GAN sketches the shapes and hues that are intuitive and primitive yielding low-resolution images with background layout drawn while Stage-II GAN corrects defects from the Stage-I image and infuses the details of objects from textual description as a high-resolution image. Sparsity due to the limited number of training text-image pairs hinders diversity of generated samples which is managed by the conditioned augmentation which could be improvised.

Extension of Deep Generator Network-based Activation Maximization (DGN-AM) for high-resolution images, activation maximization of one or multiple neurons with gradient ascent in a separate classifier network led to Plug and Play Generative Networks (PPGNs).

PPGNs (Nguyen et al., 2017) are a unified probabilistic image captioning interpretation model that generates 227 X 227 resolution images synthetically. A probabilistic framework that unifies activation maximization as energy-based models that are free to be designed and can be used in a plug and play manner with different priors forming new conditioned generative models. Lack of diversity in DGN-AM synthesized images that are biased towards the most highly activating class neuron has been overcome.

Earlier approaches demanded ground-truth layouts for generation. Whereas, inferring the semantic layout is further broadly applicable to diversified generation tasks. Envisioning semantic label maps is preferred over figuring out particular groups of structure for image generation. Hierarchical text-to-image synthesis (Hong et al., 2018) goes from coarse to fine by inferring semantic layout, which defines a scene's structure it offers fine-grained information such as the number of objects, object category, locality, size, and shape based on interest points in the scene. The employment of a series of generators allows for the gradual construction of a scene by fine-tuning the image's semantic structure: Box Generator - Creates a rough graphic layout from text embedding by training an auto-regressive decoder by lowering the negative log-likelihood of ground truth bounding boxes. Shape Generator - From the set of bounding boxes it predicts the shape of objects inside them which is built using RCNN and trained based on the GAN framework; Image Generator - Images are synthesized from text description and layout masks. Image synthesis for videos using powerful generators in plug and play fashion is an interesting extension. End-to-end training of image layout and subsequent generation might be emphasized.

Instability and sensitivity of GANs towards choice of hyper-parameters and difficulty in training while synthesizing high resolution images like 256 X 256 in StackGAN-v1 is improvised with advanced multi-stage GAN architecture. The StackGAN++(Zhang et al., 2018) is a successor of StackGAN-v1 for both conditional and unconditional generating problems. Stage-I and Stage-II GANs are almost the same as StackGAN-v1, but with an innovative conditioning augmentation approach to normalise the conditional GAN's training and enhance pixel density using the multi-distribution approximation with join conditional and unconditional distribution approximation and color-consistency regularization. Without conditional augmentation, Stage-I GAN collapses to nonsensical pictures owing to unstable training images, however this technique stabilises GAN training and strengthens diversity, fostering resilience to minor perturbations throughout much of the latent manifold. Image synthesis for diverse text embeddings and datasets rather than the state-of-the art birds and flowers is not experimented.

The lack of crucial fine-grained information at the word level in global syllable vectors hinders the fabrication sceneries and high-quality imagery. Thus, came to proposal the Attentional Generative Adversarial Network (AttnGAN) (Xu et al., 2018). Itlets on fine-grained T2I of distinct image subregions offering attention to pertinent natural language word descriptions. The attention-driven mechanism is accompanied by multi-stage refinement. The Deep Attentional Multimodal Similarity Model (DAMSM) learns two Neural Networks: The text encoder - Bi-directional Long Short-Term Memory (LSTM) extracts semantic vectors from test description; The image encoder - Images are converted into meaningful vectors using CNN. For training the AttnGAN generator, DAMSM computes the fine-grained image-text matching loss. that are conditioned on the most relevant words different sub-regions are drawn. The stratified attentional GAN spontaneouslysingles out the condition at the word level and produces different segments of the image on its own, which could be extended for complex scene generation using intricate descriptions.

The family of deeply supervised Convolutional Neural Networks (CNNs) is used to reconcile convergence between generator and discriminator while stably modelling the huge pixel space in high-resolution images and guaranteeing semantic consistency with the hierarchically nested adversarial (Zhang et al., 2018) objective. High-Definition Hierarchically Nested (HD-GANs) is an end-to-end method for modelling high-resolution image statics and generating photographic images with a generator that resembles a straightforward vanilla GAN, without any need for multi-stage training and various overlapping text conditioning or added class label supervision. Hierarchical-nested adversarial objectives are supplemented inside the nested networks, regularizing mid-level representations that capture sophisticated image statistics. Multi-purpose adversarial loss assures semantic consistency and image realism. Style transfer using sentence interpolation is modelled. Consolidation of several subtasks is braced by the mid-level representations. End-to-end high quality level image synthesis is achievable with a sole vanilla-like GAN. Playing adversarial games by distancing the generator using various leveled adversarial intentions is a unique aspect.

While the state-of-the-art T2I approach considers imparting the object of focus in the synthesized image, the Multi-Conditional Generative Adversarial Network (Park et al., 2018) preserves the background information as well. The background feature from the input image is extracted in the synthesis block that uses a convolution and batch normalization (BN) which composes the core component of MC-GAN.

False positives that occur due to geometric or semantic artifacts due to the inability of set-level constraints to learn the ‘instance-level correspondences’ which are the set-shared high-level content pertaining to identifiable objects is hammered away by the first attention mechanism integrated GANs (Ma et al., 2018) that decomposes the task of instance-level image translation with enhanced controllability exploiting instance-level jointly using a Deep Attention Encoder (DAE). The Deep Attention Generative Adversarial Network (Da-GAN) improves pose morphing and data augmentation with dramatic drop in the missing modes. DAE projects samples into the latent space upon integrating attention mechanisms. They are constrained with instance-level image transaction followed by set-level image translation which mitigates the mode collapse problem, domain adaptation and object transfiguration besides T2I.The mechanism of weak supervised attention causes incorrect attention results in some cases which seeks a more robust algorithm.

Global sentence vector based GANs miss out object level details in text embeddings while the distorted people shape in AttnGAN (Xu et al., 2018) with semantic meaning of picture layout not is not desired. Obj-GANs (Li et al., 2019) effectively captures and utilizes fine-grained word i.e., object level information for T2I. Two novel components namely, object-driven attentive generator and object-wise discriminator are used. In a multi-stage coarse-to-fine process given the text description along with a pre-generated semantic layout, Obj-GAN synthesizes high-resolution images. Fast-Region Based Convolutional Neural Network (RCNN) discriminator and image region conditioned on class label and word context vector within the bounding box is computed at every stage. ObjGAN has impressive photo-realistic images even for complex scenes with good generalization ability.

| GAN Architecture | Significance | Attribute of focus |
|---|---|---|
| GAN-INT-CLS [3] | RNN encoder with GAN encoder | Architectural novelty |
| GAWWN [4] | Bounding box and key point for parts | Visual realism |

| | | |
|---|---|---|
| ProgGAN [5] | Incremental addition of layers | High resolution |
| StackGAN [6] | Multi-stage generation process | Architectural novelty |
| PPGN [7] | Activation maximization priors | High resolution |
| Hierarchical T2I [8] | Scene based semantic layout | Semantic consistency |
| StackGAN++ [9] | Unconditional generative tasks stabilization | Diverse synthesis |
| AttnGAN [10] | Complex scenes leveraging sub-regions | Semantic consistency |
| HD-GAN [11] | Hierarchical nesting capturing complex image statistics | High resolution |
| Mc-GAN [12] | Input image background preservation | Visual realism |
| Da-GAN [13] | Instance-level image transaction with set-level image translation | Diverse synthesis |
| Obj-GAN [14] | Text embeddings over object-driven generator and object wise discriminator | Visual realism |
| Dm-GAN [15] | Fuses image and memory representation with response gates | Semantic consistency |
| MirrorGAN [16] | Aligns underlying semantics via re-description | Semantic consistency |
| Text- SeGAN[17] | Discriminator measures semantic relevance over class prediction | Diverse synthesis |

**Table 1:** Summary of related work highlighting their uniqueness and aspect of focus

The significant increase in the use of GANs for image and video creation is highly reliant on the quality of the initial images. The image refining procedures, in general, employ unaltered text representation. Each word, however, has a distinct amount of significance. A GAN model paired with a dynamic memory component (Zhu et al., 2019) was presented to create high-quality images even if the starting image was weakly generated. A memory writing gate in Dynamic Memory Generative Adversarial Networks (Dm-GAN) allows it to retrieve precise language based on the original image. Architecture was a response gate for adaptively fusing information from vision and memory. The DAMSM loss improves the conditionality of produced images based on text definition. The first pictures, which are made up of crude shapes and the incorrect hue, are improved. The layout of multi-subjects in the original image has a significant impact on the final outcome. Better organized powerful model for the initial stage is crucial.

In spite of the compelling progress of GANs in generating visually realistic images, it is to some degree challenging to achieve semantic alignment of the produced image with the text input. MirrorGAN (Qiao et al., 2019) capitalises T2I’s notion of learning by redescription, which matches its underlying semantics with the provided text description. With three modules, a global-local collaborative attention model is seamlessly embedded in cascaded generators to preserve cross-domain semantic consistency and to smoothen the generating process. Semantic Text Embedding Module (STEM), Cascaded Image Generators' Global-Local Collaborative Attentive Module (GLAM), and Semantic Text REgeneration and Alignment Module (STREAM). Context Encoder (CE) supervises whether the generated images are semantically consistent and visually realistic using the text-semantics reconstruction loss. STREAM and other MirrorGAN modules are not jointly optimized with complete end-to-end training due to limited computational resources. Due to restricted computing resources, STREAM and other MirrorGAN modules are not jointly optimized with comprehensive end-to-end training. Subject enhancement is the fundamental approach for text embedding in STEM and picture captioning in STREAM. CycleGAN (Almahairi et al., 2018) is a complement to MirrorGAN that may be used to improve the model's capability for concurrently modelling cross-media component.

The mode-collapse problem of GANs not exhibiting anticipated diversity in synthesized images which is alleviated in Text-conditioned Semantic Classifier GAN (Text-SeGAN) (Cha et al., 2019). The selection of training samples is streamlined by inclusion of more negative examples which range from easy to hard over positive ones in the semantic space. Mini-batches of triplets with a real image of corresponding text, a real image with unmatched text and a fake image with corresponding text with the captions are employed for training that achieves remarkable diversity.

## 3. Deep Attentional Architecture

Most of the GANs adopt deep layered networks for enhancing the image quality and reality. Figure 1 represents the simplified architecture of AttnGAN (Xu et al., 2018). This particular structure introduced the attention technique which was enhanced predominantly by the basic architectures of ProgGANs(Karras et al., 2017), PPGNs (Nguyen et al., 2017)StackGAN (Zhang et al., 2017) andStackGAN++ (Zhang et al., 2018).

It is interesting to learn that the motivation behind this architecture is the transformation of the Encoder-Decoder analogy of T2I problem to Sequence-to-Sequence (Seq2Seq), which leads to sequential processing of images as well. The image generation is viewed as an m-stage process making use of the attention based Seq2Seq advancements. The text and image alignment after synthesis is investigated and their agreement is harkened.

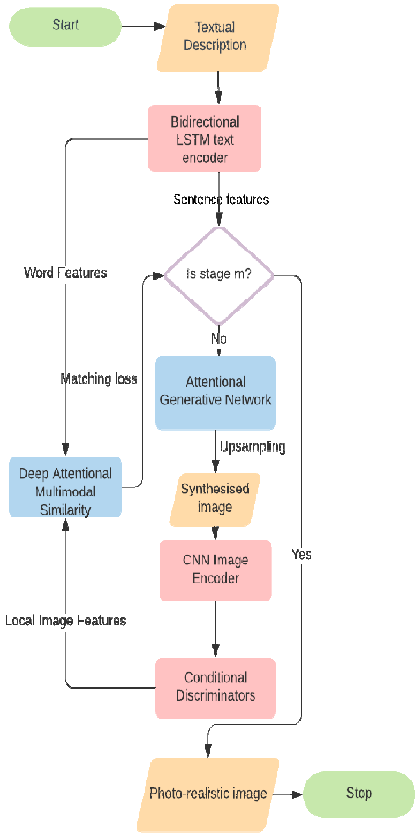

**Fig 1 :** Flow chart of deep attentional multi-layared gan

## 4. Benchmark t2i Datasets

### *4.1 COCO*

The Microsoft Common Objects in COntext (COCO)(Lin et al., 2014) is a comparatively complex dataset with shorter and fewer textual descriptions used for pretraining GANs. The number of raining samples is 80k, test samples is 40k and the number of captions per sample image is 5.

### *4.2 CUB*

The Caltech-UCSD Birds (CUB) dataset (Wahet al., 2010) is a 200 bird species annotated challenging dataset for most of the image classification systems. It is an object specific dataset used by almost all GAN based T2I models. The number of raining samples is 8855, test samples is 2933 and the number of captions per sample image is 10.

### *4.3 Oxford*

The Oxford (Nilsbacket al., 2008) is a 102-category flower dataset with matching text descriptions for almost 40 to 258 images per category.

## 5. Performance Evaluation Metrics

The GAN evaluation metrics must focus on two simple and classic properties awaited in GANs namely, fidelity and diversity. The former insists to generate high quality images while the latter focuses on how the capability of GANs in synthesizing wide range of images inherent in the given training dataset.

### *5.1 Fréchet Inception Distance (FID)*

FID (Heusel et al., 2018) acts as a standard quality metric in assessing the generated images of GANs. The evaluation compares the distribution of generated images with the distribution of training images using limited statistics such as the mean and covariance. FIDs require a large sample size of at least 10,000. Hence for very high-resolution images like 512x512 pixels it is computationally expensive. A lower FID corresponds to stronger relationship of synthetic image statistics to training image statistics which is preferred.

### *5.2 Inception Score (IS)*

The inception score (Salimans et al., 2016) is an objective metric that claims to correlate well with the human evaluation. IS judges only the distribution of generated images which heavily linked to the training images. IS fails to capture the inter-class diversity as well. A higher score indicates that the model generates better quality distinct images computed by taking an array of images and returning a single floating-point number.

## 6. Conclusion

In this work, a comprehensive study of the related work in GAN based T2I has been carried out. The architectural significance of deep attentional GANs has been depicted (Fig 1) and the standard dataset and evaluation metrics have been briefly discussed. The summary of the existing GAN architectures has been presented (TABLE I). The crux of T2I is appreciative of the aphorisms that emphasise how expressive is a single picture over thousands of words. Any computational system that persuades to empower artificial imagination can leverage these trends with desired augmentation making T2I using GANs a futuristic and amusing AI.

### References

[1] Nilsback ME, Zisserman A. Automated flower classification over a large number of classes. In 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing 2008 Dec 16 (pp. 722-729). IEEE.

[2] Wah C, Branson S, Welinder P, Perona P, Belongie S. The caltech-ucsd birds-200-2011 dataset.

[3] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In European conference on computer vision 2014 Sep 6 (pp. 740-755). Springer, Cham.

[4] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in neural information processing systems. 2014;27.

[5] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training gans. Advances in neural information processing systems. 2016;29:2234-42.

[6] Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H. Generative adversarial text to image synthesis. In International Conference on Machine Learning 2016 Jun 11 (pp.1060-1069). PMLR.

[7] Reed SE, Akata Z, Mohan S, Tenka S, Schiele B, Lee H. Learning what and where to draw. Advances in neural information processing systems. 2016;29:217-25.

[8] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems. 2017;30.

[9] Karras T, Alia T, Laine S, Lehtinen J. Progressive growing of gans for improved quality, stablility, and variation. ArXiv prepint arXiv:1710.10196. 2017 Oct 27.

[10] Zhang H, Xu T, Li H, Zhang S, Wang X, Huang Z, Metaxas DN. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision 2017 (pp.5907-5915).

[11] Nguyen A, Clune J, Bengio Y, Dosovitskiy A, Yosinski J. Plug & play generative networks: Conditional iterative generation of images in latent space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017 (pp. 4467-4477).

[12] Hong S, Yang D, Choi J, Lee H. Inferring semantic layout for hierarchical text-to-image synthesis. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 (pp. 7986-7994).

[13] Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN. Stackgan++:Realistic image synthesis with stacked generative adversarial adversarial networks. IEEE transactions on pattern analysis and machine intelligence. 2018 Jul 16;41(8):1947-62.

[14] Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, He X. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 1316-1324).

[15] 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 2018 (pp. 6199-6208).

[16] Park H, Yoo Y, Kwak N. Mc-gan: Multi-conditional generative adversarial network for image synthesis. arXiv preprint arXiv:1805.01123. 2018 May 3.

[17] Ma S, Fu J, Chen CW, Mei T. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 (pp. 5657-5666).

[18] Almahairi A, Rajeshwar S, Sordoni A, Bachman P, Courville A. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In International Conference on Machine Learning 2018 Jul 3 (pp. 195-204). PMLR.

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

[20] Zhu M, Pan P, Chen W, Yang Y. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019 (pp. 5802-5810).

[21] Qiao T, Zhang J, Xu D, Tao D. Mirrorgan: Learning text-to-image generation by redescription. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019 (pp. 1505-1514).

[22] Cha M, Gwon YL, Kung HT. Adversarial learning of semantic relevance in text to image synthesis. In Proceedings of the AAAI conference on artificial intelligence 2019 Jul 17 (Vol. 33, No. 01, pp. 3272-3279).