# Text to image generation

Pragati Patil
*234161019*
*Data Science MTech*
IIT Guwahati
p.patil@iitg.ac.in

Prakhar Shankar
*234161007*
*Data Science MTech*
IIT Guwahati
p.shankar@iitg.ac.in

G Gayathri
*234161016*
*Data Science MTech*
IIT Guwahati
g.gutla@iitg.ac.in

## I. Abstract

In our paper, we introduce the Attentional Generative Adversarial Network (AttnGAN), an innovative model designed for fine-grained text-to-image generation. AttnGAN introduces a sophisticated attention mechanism that refines images through multiple stages, allowing for the synthesis of highly detailed images guided by textual descriptions. By attending to relevant words in the natural language input, AttnGAN is capable of generating intricate details in various subregions of the image. This layered attentional GAN dynamically selects conditions at the word level, enabling the generation of distinct parts of the image with precision. Our model not only enables the synthesis of images with rich visual content but also ensures coherence and fidelity to the input description. By leveraging attentional mechanisms, AttnGAN achieves stateof-the-art performance in text-to-image generation tasks, offering significant advancements in the field of artificial intelligence and computer vision.

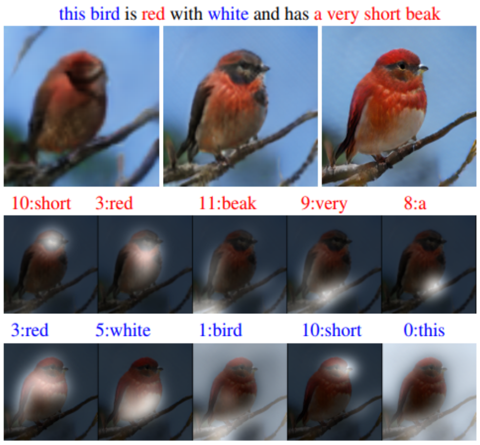


Figure 1. Example results of the proposed AttnGAN. The first row gives the low-to-high resolution images generated by $G_0$, $G1$ and $G_2$ of the AttnGAN; the second and third row shows the top-5 most attended words by $F_1^{attn}$ and $F_2^{attn}$ of the AttnGAN, respectively. Here, images of $G_0$ and $G_1$ are bilinearly upsampled to have the same size as that of $G_2$ for better visualization.

## II. Introduction

Automatically generating images according to natural language descriptions is a fundamental problem in many applications, such as art generation and computer-aided design. It also drives research progress in multimodal learning and inference across vision and language, which is one of the most active research areas in recent years.

The model consists of two novel components. The first component is an attentional generative network, in which an attention mechanism is developed for the generator to draw different sub-regions of the image by focusing on words. More specifically, besides encoding the natural language description into a global sentence vector, each word in the sentence is also encoded into a word vector. The generative network utilizes the global sentence vector to generate a low-resolution image in the first stage.

In the following stages, it uses the image vector in each subregion to query word vectors by using an attention layer to form a word-context vector. It then combines the regional image vector and the corresponding word-context vector to form a multimodal context vector, based on which the model generates new image features in the surrounding sub-regions.

This approach allows for the generation of images with fine-grained details that closely align with the textual description. The other component in the AttnGAN is a Deep Attentional Multimodal Similarity Model (DAMSM). With an attention mechanism, the DAMSM is able to compute the similarity between the generated image and the sentence using both the global sentence-level information and the finegrained word-level information. This component enhances the coherence and fidelity of the generated images by considering not only the overall meaning of the description but also the specific details conveyed by individual words. Together, these components enable AttnGAN to produce images that are not only visually appealing but also semantically meaningful, contributing to advancements in the field of multimodal learning and inference.

## III. Recent Works

In recent years, text-to-image generation has seen significant advancements driven by the exploration of various architectures and techniques. One notable recent work is the Generative Pre-trained Transformer 3 (GPT-3), a large language model developed by OpenAI. While primarily focused on natural language processing, GPT-3 has demonstrated impressive capabilities in generating images from textual prompts, albeit with limited control over the generated images' content.

Another notable advancement is the StackGAN architecture, which introduced a two-stage generation process to produce higher resolution images from text descriptions. StackGAN first generates a low-resolution image and then refines it to produce a high-resolution output. This approach improves the visual quality of the generated images, but it still struggles with capturing fine-grained details.

More recently, BigGAN and its variants have shown remarkable progress in generating high-quality and diverse images across various domains. BigGAN leverages large-scale training and sophisticated architectures to produce visually realistic images from class labels. While not specifically designed for text-to-image generation, BigGAN's capabilities have inspired further research in this area.

Furthermore, StyleGAN and its variants have introduced new techniques for controlling the style and appearance of generated images. By disentangling the latent space, StyleGAN allows for more precise manipulation of image attributes such as pose, expression, and background.

In the domain of multimodal learning, CLIP (Contrastive Language-Image Pre training) has emerged as a powerful model capable of understanding both text and images in a unified framework. By learning to associate images with their textual descriptions through contrastive learning, CLIP achieves impressive performance on various vision and language tasks, including zero-shot image classification and image generation from textual prompts.

These recent works demonstrate the ongoing progress and diverse approaches in text-to-image generation, ranging from large-scale language models to specialized architectures designed for generating high-quality and diverse images guided by textual descriptions.

## IV. Challenges and open research directions

Text-to-image generation presents several challenges and opportunities for future research. In this section, we discuss some of the key challenges and open research directions in the field.

### *A. Fine-Grained Detail Generation*

Despite recent advancements, generating fine-grained details that align closely with textual descriptions remains a challenge. Models often struggle to capture subtle nuances and specific attributes mentioned in the text.

### *B. Controlled Generation*

Controlling the attributes and characteristics of generated images remains an important area of research. Techniques for controlling style, pose, appearance, and other attributes are needed to enable more precise image generation according to user preferences.

### *C. Multi-Modal Fusion*

Integrating information from both text and images in a meaningful way poses challenges in multimodal fusion. Developing effective techniques for combining textual and visual information to generate coherent and semantically meaningful images is an ongoing area of research.

### *D. Evaluation Metrics*

Developing robust evaluation metrics for assessing the quality and fidelity of generated images is crucial. Existing metrics such as Inception Score (IS) and Fréchet Inception Distance (FID) have limitations and may not fully capture the quality of generated images.

### *E. Data Efficiency and Generalization*

Improving the data efficiency and generalization capabilities of text-to-image models is essential for real-world applications. Models should be able to generalize to unseen textual descriptions and produce high-quality images with limited training data.

### *F. Cross-Domain Generation*

Extending text-to-image generation to diverse domains and scenarios presents exciting research opportunities. Crossdomain generation involves generating images from textual descriptions that span multiple domains, such as fashion, architecture, and nature scenes.

### *G. Ethical and Bias Considerations*

Addressing ethical concerns and biases in generated images is essential. Models should be designed to mitigate biases present in training data and ensure that generated images are inclusive and representative of diverse populations.

### *H. Interactive Generation*

Enabling interactive text-to-image generation where users can provide feedback and guide the generation process in realtime is an area of growing interest. Developing models that can incorporate user feedback to refine generated images is a promising direction for future research.

### *I. Scalability and Efficiency*

Scaling up text-to-image generation models while maintaining efficiency is a challenge. Techniques for training and deploying large-scale models efficiently are needed to handle increasingly complex textual descriptions and produce high-quality images in real-time.

### *J. Explainability*

Enhancing the explainability of text-to-image generation models is crucial for understanding how they generate images from textual descriptions. Interpretable models can help users understand the underlying mechanisms and make informed decisions about the generated outputs.

Addressing these challenges and exploring these research directions will further advance the field of text-to-image generation and unlock its full potential in various applications.

## V. Application of Text-to-image generation

Text-to-image generation has a wide range of applications across various domains. In this section, we explore some of the key application areas where text-to-image generation techniques are being applied.

### *A. Art Generation*

Text-to-image generation enables artists and designers to quickly translate their ideas and concepts into visual representations. Artists can use textual descriptions to generate initial sketches or concept art, facilitating the creative process.

### *B. E-commerce and Product Visualization*

In e-commerce, text-to-image generation can be used to generate realistic product images from textual descriptions. This allows customers to visualize products before making a purchase, improving the online shopping experience.

### *C. Content Creation*

Content creators, such as bloggers, social media influencers, and marketers, can use text-to-image generation to create engaging visuals for their content. By generating images from textual descriptions, they can enhance the visual appeal of their posts and attract more audience engagement.

### *D. Virtual Try-On*

Text-to-image generation can be applied in virtual try-on systems, where users can describe clothing items or accessories, and the system generates virtual representations of how they would look when worn. This is particularly useful in the fashion industry for online shopping and styling services.

### *E. Storytelling and Narrative Generation*

Authors and storytellers can use text-to-image generation to illustrate their narratives with visual content. By generating images that correspond to key scenes or characters described in the text, they can create more immersive storytelling experiences.

### *F. Education and Training*

Text-to-image generation can be used in educational settings to create visual aids, diagrams, and illustrations for teaching materials. It can also be used in virtual reality and simulation environments for training purposes.

### *G. Medical Imaging*

In medical imaging, text-to-image generation can assist healthcare professionals in visualizing medical conditions and treatments. By generating images from medical reports or descriptions, doctors can better understand and communicate complex medical information.

### *H. Architectural Design*

Architects and urban planners can use text-to-image generation to generate architectural renderings and visualizations from textual descriptions of buildings and landscapes. This facilitates the design process and helps stakeholders visualize proposed projects.

### *I. Forensics and Law Enforcement*

Text-to-image generation techniques can aid in forensic investigations and law enforcement by generating composite sketches of suspects based on eyewitness descriptions or textual evidence.

### *J. Entertainment and Gaming*

In the entertainment industry, text-to-image generation can be used to create characters, scenes, and backgrounds for video games, animations, and virtual worlds. It enhances the immersive experience for players and viewers.

These applications demonstrate the versatility and potential impact of text-to-image generation across various fields, paving the way for innovative solutions and new opportunities.

## VI. Evaluation Metrics and benchmarks

Assessing the performance of text-to-image generation models requires reliable evaluation metrics and benchmarks. In this section, we examine some commonly used metrics and benchmarks in the field, specifically focusing on the CUB dataset.

TABLE I
INCEPTION SCORES

| GAN-INT-CLS | GAWWNStackGAN | StackGAN-v2 | PPGN | AttnGAN |
|---|---|---|---|---|
| 2.88 ± .04 | 3.62 ± .07 | 3.70 ± .04 | 3.82 ± .06 | 4.36 ± .03 |

Inception Score (IS):

Inception Score measures the quality and diversity of generated images by evaluating the distribution of generated images and their predicted class labels using an Inception-v3 classifier. Higher IS values indicate better quality and diversity in generated images.

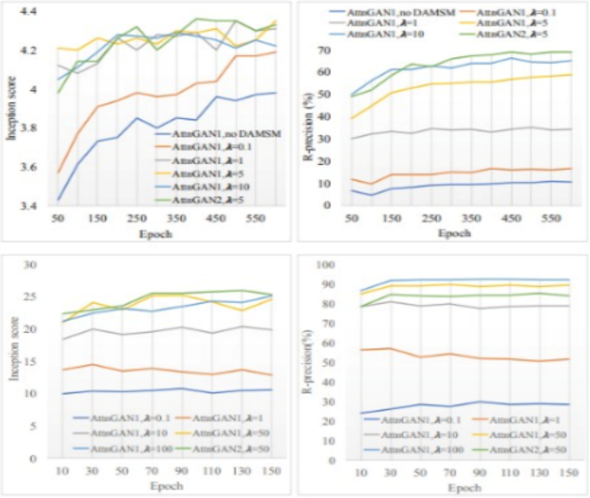


**Figure 3. Inception scores and R-precision rates by our AttnGAN and its variants at different epochs on CUB (top) and COCO (bottom) test sets. For the text-to-image synthesis task, $R = 1$.**

Fig. 1. DAMS Model

For the CUB dataset, the Inception Scores for various models are as shown in table 1.

R-Precision:

R-Precision measures the precision of relevant features among the generated images. It is computed as the percentage of relevant features correctly identified among all relevant features.

For the AttnGAN models on the CUB dataset, the R-Precision rates are as shown in table 2.

.

Precision and Recall:

Precision and recall metrics evaluate the accuracy of generated images in capturing specific attributes or features mentioned in the textual descriptions. Precision measures the proportion of relevant features among the generated images, while recall measures the proportion of relevant features that are correctly identified.

TABLE II
R-PRECISION

| Method | R-Precision(%) |
|---|---|
| AttnGAN1, no DAMSM | 10.37 ± 5.88 |
| AttnGAN1, = 0.1 | 16.55 ± 4.83 |
| AttnGAN1, = 1 | 34.96 ± 4.02 |
| AttnGAN1, = 5 | 58.65 ± 5.41 |
| AttnGAN1, = 10 | 63.87 ± 4.85 |
| AttnGAN2, = 5 | 67.82 ± 4.43 |

Human Evaluation:

Human evaluation involves subjective assessment by human annotators to evaluate the visual quality, coherence, and relevance of generated images. Annotators may be asked to rate images based on predefined criteria or provide qualitative feedback.

Dataset-Specific Metrics:

Some datasets used for text-to-image generation tasks come with specific evaluation metrics tailored to the dataset's characteristics. For example, CLEVR (Compositional Language and Elementary Visual Reasoning) dataset includes metrics for assessing the model's ability to perform visual reasoning tasks.

Perceptual Similarity:

Perceptual similarity metrics assess the perceptual similarity between real and generated images based on human perception. These metrics consider factors such as color, texture, and structure similarity.

User Studies:

User studies involve collecting feedback from human users to assess the subjective quality and utility of generated images. Users may be asked to compare images generated by different models or provide feedback on the realism and relevance of generated images.

Benchmarks:

Benchmarks provide standardized datasets and evaluation protocols for benchmarking the performance of text-to-image generation models. Common benchmarks include the MS COCO (Microsoft Common Objects in Context) dataset and the CUB (Caltech-UCSD Birds) dataset.

Domain-Specific Evaluation:

Some domains, such as medical imaging or architectural design, may require domain-specific evaluation metrics tailored to the particular requirements and challenges of those domains.

Long-Term Evaluation:

Long-term evaluation assesses the robustness and generalization capabilities of text-to-image generation models over extended periods. It involves testing models on unseen data and evaluating their performance over time.

By employing appropriate evaluation metrics and benchmarks, researchers can quantitatively and qualitatively assess the performance of text-to-image generation models, leading to more reliable and reproducible research outcomes.

## VII. DATA

The data used for training and evaluation in this project is the Caltech-UCSD Birds-200-2011 (CUB) dataset. The CUB dataset contains images of birds with corresponding textual descriptions, making it suitable for text-to-image generation tasks.

### A. CUB dataset:

*1) Description:* The CUB dataset consists of 200 bird species with a total of 11,788 images. Each image is accompanied by a textual description of the bird species.

*2) Characteristics:* The images in the CUB dataset vary in resolution and quality, capturing various poses, backgrounds, and lighting conditions of birds in natural settings.

*3) Training and Evaluation:* The AttnGAN model is trained using a subset of the CUB dataset, with images and corresponding textual descriptions used as paired inputoutput examples. The model's performance is evaluated on a separate subset of the dataset to assess its ability to generate realistic images from textual descriptions.

### B. Other Potential Datasets

In addition to the CUB dataset, there are several other datasets that can be used for text-to-image generation tasks:

*1) COCO (Common Objects in Context):* The COCO dataset contains a large collection of images depicting everyday scenes with multiple objects and corresponding textual descriptions.

*2) CelebA:* The CelebA dataset consists of celebrity face images with associated attributes and textual descriptions, making it suitable for generating facial images from text.

*3) Oxford-102:* The Oxford-102 dataset contains images of flowers with textual descriptions, allowing for the generation of flower images based on textual input.

Data suitability:

Each dataset has its own characteristics and domain-specific features, making it suitable for different applications of text-to-image generation. The choice of dataset depends on the specific task requirements and the desired domain of application. In our project, we chose the CUB dataset due to its focus on bird images and corresponding textual descriptions. This alignment with our task objectives allows us to train the AttnGAN model to generate realistic bird images from textual descriptions effectively. The detailed descriptions provided in the CUB dataset enable the model to learn the fine-grained visual features of various bird species, enhancing its ability to produce accurate and visually coherent outputs.

## VIII. METHOD

The method section describes the architecture and components of the Attentional Generative Adversarial Network (AttnGAN), including the attentional generative network and the Deep Attentional Multimodal Similarity Model (DAMSM).

### A. Attentional Generative Network

The attentional generative network in AttnGAN consists of two main components: a text encoder and an image encoder.

### B. Text Encoder

The text encoder is implemented as a bidirectional Long ShortTerm Memory (LSTM) network, which extracts semantic vectors from the textual description. Each word in the sentence corresponds to two hidden states, one for each direction of the LSTM. The two hidden states are concatenated to represent the semantic meaning of the word.

### C. Image Encoder

The image encoder is a Convolutional Neural Network (CNN) that maps images to semantic vectors. The CNN architecture used in AttnGAN is built upon the Inception-v3 model, which is pretrained on a large dataset for image classification tasks. The intermediate layers of the CNN learn local features of different sub-regions of the image, while the later layers learn global features of the image.

### D. Deep Attentional Multimodal Similarity Model (DAMSM)

The DAMSM consists of two neural networks that map subregions of the image and words of the sentence to a common semantic space. It measures the image-text similarity at the word level to compute a fine-grained loss for image generation.

### E. Word-level Image-Text Similarity

The DAMSM learns to compute the similarity between the generated image and the sentence using both the global sentence-level information and the fine-grained word-level information. It measures the semantic similarity between words in the sentence and corresponding subregions of the image.

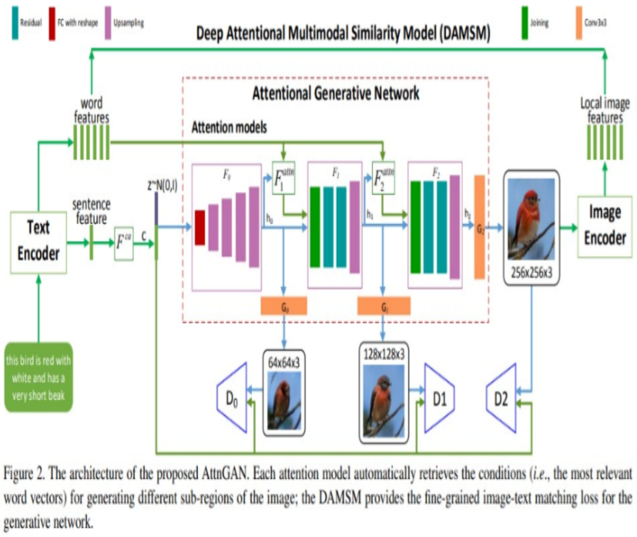


Figure 2. The architecture of the proposed AttnGAN. Each attention model automatically retrieves the conditions (*i.e.*, the most relevant word vectors) for generating different sub-regions of the image; the DAMSM provides the fine-grained image-text matching loss for the generative network.

Fig. 2. DAMS Model

### *F. Training and Optimization*

The text encoder, image encoder, and DAMSM are trained jointly using adversarial training.

The generator is trained to maximize the likelihood of generating realistic images that match the input textual descriptions, while the discriminator is trained to distinguish between real and generated images. The DAMSM is optimized to minimize the discrepancy between the similarities of image-text pairs in the semantic space.

### *G. Multi-Stage Refinement*

AttnGAN employs a multi-stage refinement process to generate high-resolution images from textual descriptions. In each stage, the generator refines the generated image based on the attentional information obtained from the DAMSM.

### *H. Architecture Overview*

The overall architecture of AttnGAN consists of multiple stages of refinement, with the attentional generative network generating images guided by the textual descriptions and the DAMSM providing feedback on the image-text similarity at the word level.

## IX. Experiments

Extensive experimentation is carried out to evaluate the proposed AttnGAN thoroughly. The important components of the AttnGAN, including the attentional generative network and the Deep Attentional Multimodal Similarity Model (DAMSM), are studied extensively to understand their contributions to the model's performance. Datasets for our method are evaluated on the Caltech-UCSD Birds-200-2011 (CUB) dataset, which contains a diverse collection of bird images with corresponding textual descriptions.

### *A. Visualization of Intermediate Results with Attention*

To gain insights into the learning process of the AttnGAN, we visualize its intermediate results with attention. The first stage of the AttnGAN (G0) primarily captures the primitive shape and colors of objects, resulting in low-resolution images. However, since only the global sentence vectors are utilized in this stage, the generated images often lack finer details described by exact words, such as the beak and eyes of a bird. Subsequent stages (G1 and G2) refine the generated images based on word vectors. The model learns to rectify defects in results of the previous stage and adds more details to produce higherresolution images. Some sub-regions or pixels of G1 or G2 images can be inferred directly from images generated by the previous stage. For these regions, the attention is equally distributed among all words in the text description. Conversely, for sub-regions with semantic meaning expressed in the text description, the attention is focused on their most relevant words. This demonstrates that the AttnGAN effectively learns to understand the detailed semantic meaning conveyed in the text description of an image.

### *B. Generalization Ability*

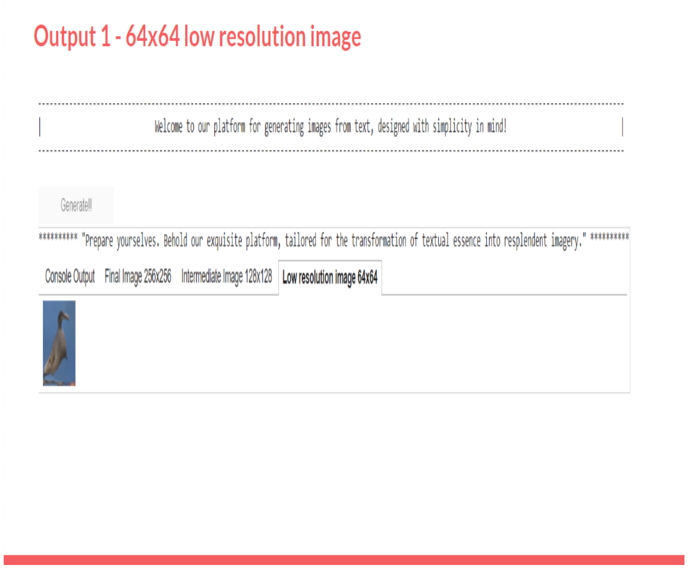
Our experimental results have demonstrated the generalization ability of the AttnGAN by generating images from unseen text descriptions. Further testing examines the sensitivity of the model's outputs to changes in the input sentences by modifying some of the most attended words in the text descriptions. The resulting images accurately reflect subtle semantic differences in the text description, showcasing the model's capability to capture nuanced variations. Additionally, the AttnGAN is capable of generating images to reflect the semantic meaning of descriptions of novel scenarios that are not likely to occur in the real world. However, we also observe instances where the AttnGAN generates images that are sharp and detailed but lack realistic plausibility. For example, birds with multiple heads, eyes, or tails are generated, which are more reminiscent of creatures from fairy tales than real-world birds. This suggests that there is room for improvement in capturing global coherent structures and ensuring the realism of generated images.

Observations from the extensive experimentation further validate the generalization ability of the AttnGAN while also highlighting areas for potential improvement in its performance.

## X. Input Output Examples

In this section, we present input-output examples to demonstrate the capabilities of the proposed AttnGAN model in generating images from textual descriptions. Each example consists of a textual prompt and the corresponding image generated by the AttnGAN model.

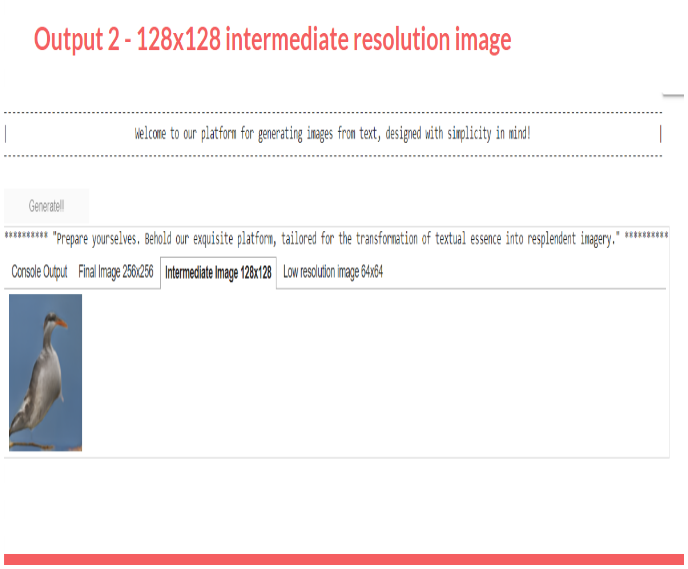
Example:
Input: ”A white color bird with a red long neck.”
Output:


Output 1 - 64x64 low resolution image

Welcome to our platform for generating images from text, designed with simplicity in mind!

********** "Prepare yourselves. Behold our exquisite platform, tailored for the transformation of textual essence into resplendent imagery." **********

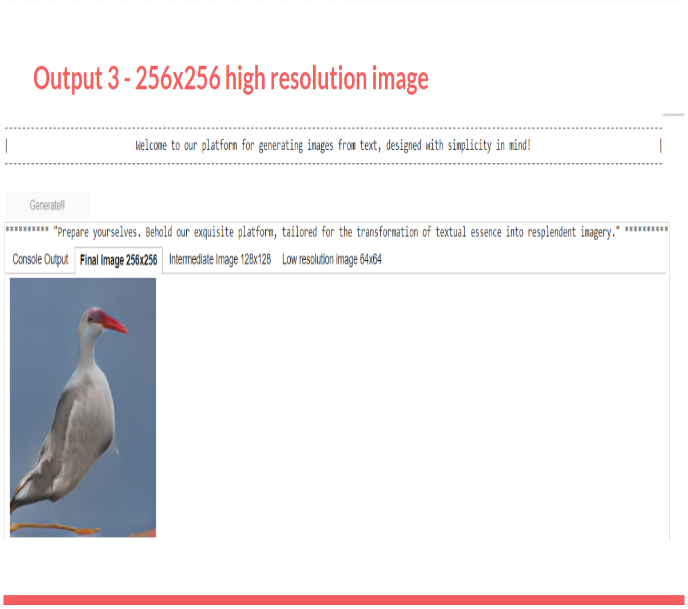
Fig. 3. 64X64 resolution image


Output 2 - 128x128 intermediate resolution image

Welcome to our platform for generating images from text, designed with simplicity in mind!

********** "Prepare yourselves. Behold our exquisite platform, tailored for the transformation of textual essence into resplendent imagery." **********

Fig. 4. 128X128 resolution image


Output 3 - 256x256 high resolution image

Welcome to our platform for generating images from text, designed with simplicity in mind!

********** "Prepare yourselves. Behold our exquisite platform, tailored for the transformation of textual essence into resplendent imagery." **********

Fig. 5. 256X256 resolution image

## XI. Conclusion

In this paper, we introduce AttnGAN, an Attentional Generative Adversarial Network designed for fine-grained textto-image synthesis. Our approach employs a novel attentional generative network to facilitate the multi-stage process of image generation, ensuring the production of high-quality images that faithfully represent textual descriptions. Extensive experimentation validates the effectiveness of the proposed attention mechanism within AttnGAN, particularly in the context of generating images for complex scenes.

Our attentional generative network allows AttnGAN to attend to relevant words in the textual description at different stages of the image generation process. This attention mechanism enables the model to capture fine-grained details and intricate visual features, resulting in more realistic and visually coherent images. By iteratively refining the generated images based on attentional cues from the text, AttnGAN demonstrates superior performance in handling complex scenes compared to previous approaches.

Through comprehensive experimental evaluation on datasets such as the Caltech-UCSD Birds-200-2011 (CUB) dataset, we showcase the efficacy of AttnGAN in generating high-quality images from textual descriptions. The attention mechanism plays a crucial role in capturing the nuances of the input text, allowing the model to produce images that accurately reflect the semantics and details described in the text.

The success of AttnGAN in generating fine-grained images underscores the importance of attention mechanisms in textto-image synthesis tasks. Our approach not only enhances the quality of generated images but also improves the model's ability to handle diverse and complex scenes effectively. The proposed attentional generative network opens up new possibilities for advancing the state-of-the-art in text-to-image generation and holds promise for a wide range of applications in art, design, virtual reality, and beyond.

In conclusion, AttnGAN represents a significant advancement in the field of text-to-image synthesis, offering a powerful framework for generating high-quality images from textual descriptions. The attention mechanism integrated into AttnGAN enables the model to effectively leverage the semantics of the input text, resulting in visually compelling images that faithfully represent the described scenes. Our work paves the way for future research in multimodal learning and inference, driving progress towards more intelligent and human-like artificial systems.

## XII. References

[1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. VQA: visual question answering. IJCV, 123(1):4–31, 2017.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473, 2014.

[3] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In NIPS, 2015.

[4] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollar, ´ J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In CVPR, 2015.

[5] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng. Semantic compositional networks for visual captioning. In CVPR, 2017.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In NIPS, 2014.

[7] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. DRAW: A recurrent neural network for image generation. In ICML, 2015.

[8] X. He, L. Deng, and W. Chou. Discriminative learning in sequential pattern recognition. IEEE Signal Processing Magazine, 25(5):14–36, 2008.

[9] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning deep structured semantic models for web search using clickthrough data. In CIKM, 2013.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In CVPR, 2017.

[11] B.-H. Juang, W. Chou, and C.-H. Lee. Minimum classification error rate methods for speech recognition. IEEE Transactions on Speech and Audio Processing, 5(3):257–265, 1997.

[12] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In ICLR, 2014.

[13] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image superresolution using a generative adversarial network. In CVPR, 2017.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.

[15] E. Mansimov, E. Parisotto, L. J. Ba, and R. Salakhutdinov. Generating images from captions with attention. In ICLR, 2016.

[16] A. Nguyen, J. Yosinski, Y. Bengio, A. Dosovitskiy, and J. Clune. Plug and play generative networks: Conditional iterative generation of images in latent space. In CVPR, 2017.

[17] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In ICLR, 2016.

[18] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In NIPS, 2016.

[19] S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. In CVPR, 2016.

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

[21] S. E. Reed, A. van den Oord, N. Kalchbrenner, S. G. Colmenarejo, Z. Wang, Y. Chen, D. Belov, and N. de Freitas. Parallel multiscale autoregressive density estimation. In ICML, 2017.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. IJCV, 115(3):211–252, 2015.

[23] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In NIPS, 2016.

[24] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. IEEE Trans. Signal Processing, 45(11):2673–2681, 1997.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In CVPR, 2016.

[26] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. Conditional image generation with pixelcnn decoders. In NIPS, 2016.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. arXiv:1706.03762, 2017.

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR- 2011-001, California Institute of Technology, 2011.

[29] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, 2015.

[30] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola. Stacked attention networks for image question answering. In CVPR, 2016.

[31] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic

image synthesis with stacked generative adversarial networks. In ICCV, 2017.

[32] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. arXiv: 1710.10916, 2017.