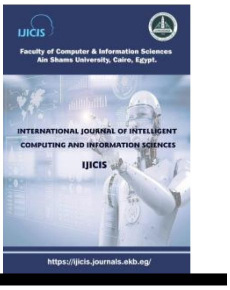
IJICIS, Vol.25, No.2, 63-73
DOI: 10.21608/ijicis.2025.392458.1400

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/


IJICIS
Faculty of Computer & Information Sciences Ain Shams University, Cairo, Egypt.
INTERNATIONAL JOURNAL OF INTELLIGENT COMPUTING AND INFORMATION SCIENCES
IJICIS
https://ijicis.journals.ekb.eg/

ARA-RATGAN FOR ARABIC TEXT TO IMAGE SYNTHESIS

Mostafa Samy Ibrahem*
Scientific Computing,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
mostafasamyibrahem@gmail.com

Mariam Nabil Al-Berry
Scientific Computing,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
mariam_nabil@cis.asu.edu.eg

Sayed Fadel
Scientific Computing,
Faculty of Computer and Information Sciences, Ain Shams University,
Cairo, Egypt
sayedfadel@cis.asu.edu.eg

Received 2025-06-05; Revised 2025-06-05; Accepted 2025-06-29

Abstract: *Current text-to-image systems have revealed outstanding performance in tasks requiring the automated synthesis of realistic generated images from text descriptions. Previous approaches typically employ multiple sperate fusion blocks to adaptively fuse appropriate text information into the generation process, which increases the difficulty of training and conflict with one another. To solve these concerns, we present Arabic Recurrent Affine Transformation (Ara-RATGAN) a novel framework that integrates AraBERT -- a pretrained Arabic BERT that has been trained on billions of Arabic words to generate robust Arabic sentences embedding with Recurrent Affine Transformation (RAT) to generate images with high-quality from Arabic-language text descriptions. Furthermore, a spatial attention model is used in the discriminator to promote semantic coherence between text and synthesized images, which identifies corresponding image areas, and directs the generator to produce more appropriate visual contents according to the Arabic text descriptions. We conducted our extensive experiments on Arabic CUB dataset translated from English to Arabic, which shows a superior performance of our proposed model in comparison to the previous Arabic text-to-image models. Our approach addresses two key challenges: (1) Text-Image Fusion: Unlike traditional methods that use isolated fusion blocks, we employ RAT to model long-term dependencies across layers, ensuring global consistency in text conditioning. (2) Semantic Alignment: A spatial attention mechanism is used in the discriminator to enhance the semantic coherence between the synthetic visuals and Arabic text.*

Keywords: *AraBERT, Generative Adversarial Networks (GANs), Text-to-Image, Feature Fusion.*

1. Introduction

In previous years, Generative Adversarial Networks (GANs) [1], have revealed significant effectiveness in a variety of applications, including enhancing resolution and augmenting data. The top-performing popular text-to-image models employs GANs as the main core of their architectures to produce natural-looking images with high resolution after training with English text captions.

***Corresponding Author**: Mostafa Samy Ibrahem

Scientific Computing Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt

Email address: mostafasamyibrahem@gmail.com

StackGAN [2], for example, used a two phases process to produce genuine images with high resolution by layering many pairs of generators and discriminators in a 2-stage style problem. First it generated low-resolution images from the noise vector and text captions, and then it was trained on generating realistic quality images with the use of another conditioned pair with images from the first stage and the text captions.

Similarly, [3] paid attention to the matching words of the description by taking numerous steps to create fine and smooth features in images. Whereas MirrorGAN [4], presented three models that converted from text to image and from image to text, and the second model additionally employed numerous pairs of discriminators and generators.

Despite prior contributions in these areas, which have made tremendous advances, they are sluggish and unstable in training [3, 4]. Furthermore, they consume a significant amount of time and computing resources to train. GANs often fuse appropriate text information into the synthesizing process adaptively with the use of multi-isolation fusion blocks like Conditional Instance Normalization (CIN) [5], and Conditional Batch Normalization (CBN) [6]. The CIN was first proposed for style-transfer. After that BigGAN [7], and StyleGAN-T [8], used CBN and CIN to generate natural looking images with impressive visual qualities. Also, DF-GAN [9], DT-GAN [10], and SSGAN [11] utilized them to integrate textual data into generated visuals.

Despite their ubiquity, CIN and CBN have a significant disadvantage, in that they are segregated at various levels, ignoring the assignment of textual information globally, which is fused into multiple layers. Furthermore, because of the use of isolated fusion blocks, which do not interact with one another, they are difficult to optimize and train sufficiently. In contrast, RATGAN [12] was a proposed framework, which reliably regulated all fusion blocks. Also, to provide unified control of multiple layers, RAT described the output of distinct levels using standard context vectors of the same shape. The contextual vectors were then linked together using Recurrent Neural Networks (RNNs) to determine extended sequential relationships. Besides the fact that fusion blocks with skip connections in RNNs to maintain consistency not just between adjacent blocks, but it also lowers training complexity. Furthermore, it used a spatial attention mechanism within the discriminator to promote text-image semantic alignment in generated output images. Text descriptions, which were aware of the match of image regions content, directed the generator to generate more appropriate image content. With spatial attention, its discriminator may direct its attention to regions of the images relevant to the textual description, allowing it to monitor the generator more effectively. Afterall, it needs a necessary pre-trained model to generate sentence embeddings.

CUB [13] dataset has 11, 788 images with 5 captions for each image containing train and test data. As a result, these instances are insufficient for model training phase, to provide robust and resilient sentence representations. Modern text-to-image systems heavily depend on text encoders, which are used to extract embeddings from the input text, and to establish cross-modal alignment between linguistic and visual representations.

Classical approaches usually use LSTMs [14], or GRUs [15], which struggle to encode the Arabic text properly, due to its morphological complexity and limited contextual awareness of the model. After the transformers [16] architecture has been proposed, the efficiency of the text encoder has been greatly impacted. For Example, AraBERT [17], a dedicated Arabic BERT [18] variant, that underwent extensive pre-training phase on billions of Arabic-language tokens, and captured the morphological pattern through tokenization, and bidirectional attention, achieving a state-of-the-art result in the Arabic-language tasks. Multilingual text encoders like mT5 [19] offer a broader language coverage but often underperform on Arabic due to diluted token embeddings. Recent hybrid approaches (e.g., LSTM-Transformers cascaded [20] attempted to balance the sequential processing and attention mechanisms, but at the cost of heavy computations. To address these constraints, our work uniquely adapted

AraBERT's text encoder strength to image synthesis, demonstrating that language-specific text encoders outperform generalized models for Arabic text to image generation.

Our key contributions can be summarized in the following points:

1. AraBERT integration: we leverage AraBERT, a BERT-based Arabic language model, to generate robust sentence embeddings.
2. Creating new data for Arabic text to image synthesis tasks by utilizing Helsinki-NLP opus-mt-en-ar [21]
3. Recurrent Affine Transformation (RAT): replacing the isolated fusion blocks with RNN-connected RAT layers, ensuring global text conditioning.
4. Our experimental results on the challenging CUB dataset have proved the ability of the pipeline to produce unprecedented realism in the generated images from Arabic text captions.

2. Related Work

Text-to-Image generation is considered one of the tasks in conditional image generation. Since the development of GANs, conditional GANs have performed exceptionally well on this task of image generation. The conditional variant of GAN, which simply combines the noise vector and conditional feature vector, was initially proposed in [22]. Nevertheless, text to image models built around this research concatenated the text features with the noise vector to fuse the text information. A more sophisticated fusion approach (i.e., CIN), that employs variance and adaptive mean to control the style of the image was developed in [23]. In contemporary work, CIN and its variations were often employed. For instance, to obtain impressive results in visual quality perspective on ImageNet, BigGAN and StyleGAN-T employed CBN and CIN sequentially. Furthermore, CBN and CIN have recently been used for including text information into synthesis. SSGAN has suggested Semantic Spatial Aware CBN as a technique to make it aware of spatial regions which are important to text description. For better fusion of text information into the generation process, DF-GAN, has been proposed in order to deep fuse the text information better with the use of multiple affine layers within each block. In contrast to earlier research, DF-GAN, foregoes normalization operations without affecting performance, reducing the computational load and large batch size restrictions. Our model discards normalization operations in a manner similar to DF-GAN since normalization has no impact on performance.

Previous research [24, 25] demonstrated that adding more spatial information may significantly enhance image quality. Bounding-boxes and key-points are used in the approach of GANTCLS, to decide the locations and types of objects to draw in the image. In order to construct the feature map, they used the essential locations and bounding boxes as the generator's input. However, key-points are very unreliable, and identifying key-points and bounding-boxes requires a lot of human involvement and effort. Key-points are still required during training, despite the fact that they also attempted to produce them using an additional GAN.

Wang and Gupta [26], created images based on structure maps that might affect produce crude drawings of the source images, similar to GANTCLS. Additionally, they added another GAN to create structure maps. The two models mentioned before developed handcrafted annotations as they learn to produce images from scratch with a clear understanding of spatial structure. Nevertheless, for each of them to train their model, they required more manual annotations.

The novel architecture put forward by AttGAN, synthesized fine-grained informational details in images by giving more attention to the descriptions of associated words. In the initial stage, AttGAN generated images with low-resolution utilizing overall sentence and noise vectors. In the following stage, the previously hidden features of the image and features of the words are used to create additional hidden features of images, and so on, until an image with high-resolution is generated.

To generate images from text description in MirrorGAN [4], the authors proposed an architecture that is composed of three models. They suggested using embedding of word sentence average, in order to ensure global semantic congruence between the synthesized images and text descriptions. The idea of learning redescription to learn text-image generation was also utilized.

DMGAN [27], tried to improve the initially generated image content by including a dynamic memory module. Using a memory writing gate, this approach picked the required text information, while keeping the initial image content into consideration. It also employed a response gate to combine the textual information received from the memory gate with the image characteristics.

In TextControlGAN [28], a Regressor model in Neural Network style was proposed. The model was able to learn the appropriate features from the conditional text. A global-local aware word judgement method was utilized in MJ-GAN [29], thus constructing a multiple level judgment to judge the authenticity of the generated image, while sentence-level judgment judge on the semantic consistency between the image regions' content and information extracted from words describe the image. In order to capture the image scene's complexity with a single stage generator, HDGAN [30] introduced an objective regularized representation of the mid-level features and generator training assistance. These all systems rely on English text descriptions processed through various advanced technologies. Our work incorporates AraBERT [17], which follows BERT [18] architecture and has demonstrated significant success in Arabic language comprehension. We introduce a novel framework that combines AraBERT's effectiveness as a pre-trained model with RATGAN to generate images from Arabic Text. To evaluate our architecture's performance and the quality of the generated images, we employ standard metrics used in previous works: Inception Score, and Fréchet Inception Distance.

3. Proposed Model

3.1. AraBERT

In 2018 a Bidirectional Encoder Representation from Transformers model, shortly known as BERT [18] was proposed by Google. It was built using the Transformers architecture, which represent a breakthrough deep learning model architecture that has made substantial progress in Natural Language Processing (NLP) applications. In order to understand the context and relations between single words in a sentence, both the words that come before and after were taken into account. BERT can build more accurate representation of words and sentences by considering context from both sides. To learn broader language representation and better understanding, BERT was trained on an extensive corpus of unlabeled text, such as Wikipedia articles and books. This was accomplished by anticipating masked words in a sentence and assessing if two sentences in the original text are sequential. BERT could also capture a comprehensive grasp of language patterns and semantics, thanks to the pre-trained procedure.

BERT was then finetuned on a certain downstream task after pre-training, such as Named Entity Recognition (NER), Question Answering, or Text Classification/Categorization. Furthermore, BERT was trained on labelled data for a required task during fine-tuning, and accordingly the model's parameters were updated to generate the correct prediction, which finally resulted in increasing the performance and accuracy of several NLP tasks.

Similar to BERT, AraBERT is the Arabic version of BERT following the same architecture design but trained on Arabic text from the internet. Handling a rich and morphological language like Arabic language with the lack of resources compared to English Language. AraBERT was trained with a large-scale corpus of Arabic data from new articles, websites and books in different Arab regions.

This included for example, Arabic corpus which was derived from 10 major new sources covering 8 nations with size of 1.5 billion words, and the publicly available International Arabic News corpus, which was formed using 31 news sources in 24 Arabs countries. It has over three million articles (near to one billion tokens). The dataset contains 70M+ sentences. Motivated by these results, we used AraBERT to generate robust Arabic word embeddings to train our RATGAN model to synthesize images from Arabic text.

3.2. The Distinction between RATGAN's Text Encoder and AraBERT

RATGAN's textual encoder employs a bidirectional LSTM architecture, used to generate words and sentence vectors with high semantic embeddings by encoding the text inputs. In order to preserve inter-word semantic relationships, each word is represented within the hidden states as an encoded form of the input information. The hidden states are then concatenated to generate text embeddings semantics. (D×T) is the text dimension incorporating the words. D is the vectors length, and total number of words T, taken from the overall text descriptions of the dataset text. In RATGAN experimental work, D = 256, which corresponds to RATGAN's input dimension. Furthermore, using a weak supervision technique the text encoder was trained entirely from scratch, using the CUB train dataset which is composed of 9414 text description. The Arabic language is the richest language in vocab. So, utilizing a tiny text encoder and starting training it from scratch may result in poor semantic meaning learning. As a result, rather than training it from scratch on tiny datasets, we present a pretrained skilled model that has been trained on thousands of millions Arabic words and has shown a significant performance in NLP. It finetunes the AraBERT model to generate proper semantic text and words embeddings. Despite the diversity of the Arabic Language terms, this model demonstrated the capacity and ability of a pretrained model to synthesize images from Arabic-language text. Figure 1 illustrates the architecture of the Ara-RATGAN model.

3.3. RATGAN

In order to improve the consistency of different layers' fusion blocks, we employed Recurrent Affine Transformation in the generator. RAT first performs scaling operations on the channels of image features using the scaling parameter, followed by shifting operation on the channels of the image features, explained in "Equation 1". The same affine transformation operations are done for each image feature vector in the image features maps. A RAT block is formed by stacking numerous RNNs, RATs, and convolutions to increase the depth of the network and its nonlinearity. Where h_t is the RNN's hidden state, and β, γ are the parameters predictions by two of conditioned MLPs on h_t as demonstrated in "Equation 1".

$$\text{Affine}\,(c \mid h_t) = \gamma_i \odot c + \beta_i \qquad (1)$$

For assigning textual information in a global manner, we use LSTM instead of Vanilla RNN inside the RAT blocks, to simulate the temporal structure inside the RAT blocks. Unlike Conditional Instance/Batch Normalization and deep conditional fusion, our approach does not deal with affine transformation as separate modules.

Conversely, we use RNN to describe the dependency of long-term between fusion blocks, that besides compels the fusion blocks to be consistent with each other, it makes the training with skip connections easier. A one stage Generator was used which is 6 up sample blocks to produce fake images using the sentence vectors. Furthermore, a Gaussian distribution was used to sample the noise vector used to be fed to the generator at the start of training. To regulate the generated image contents, a RAT block comes before each up-sample block. A tangent hyperbolic function transforms the feature map into a fake image.

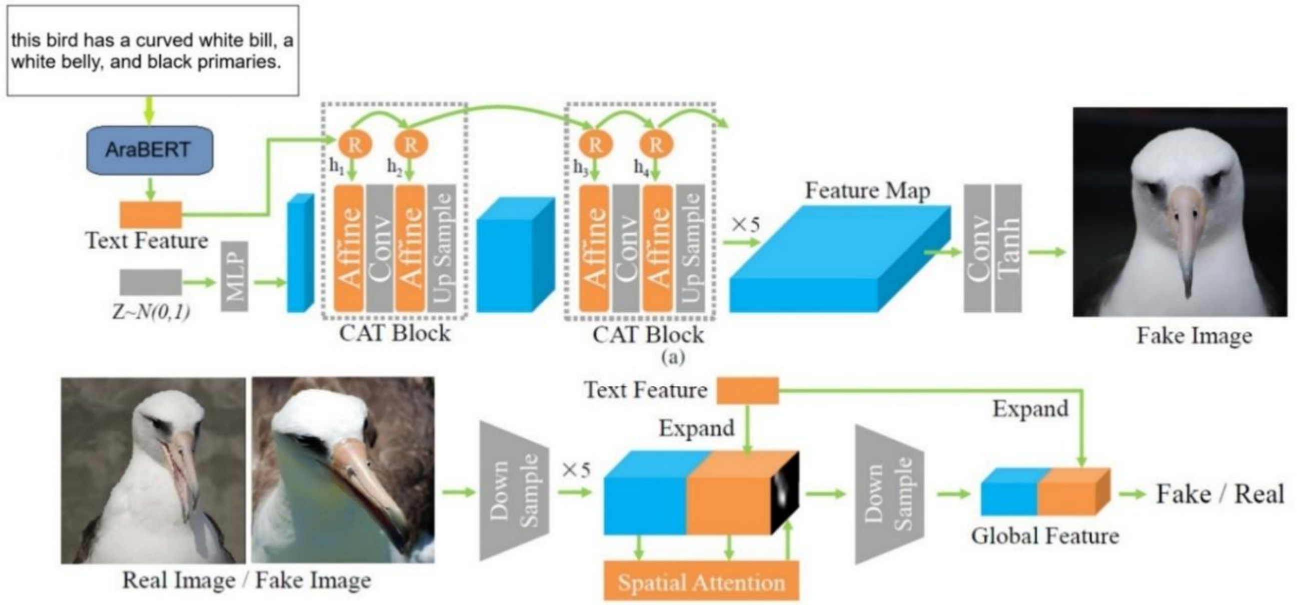


Figure 1: The architecture of the Ara-RATGAN model

3.4. RATGAN's Discriminator

Spatial attention approach was added into our discriminator to improve the semantic coherence between the synthesized images and text descriptions. To be aware of matching text descriptions and image content, it directs the generator to generate more appropriate real images. Figure [1] shows how the image is encoded into feature maps using numerous down-sample blocks. Spatial Attention generates an attention map that surpasses the representation provided by sentence vectors for unimportant regions by combining the information in the image feature map and sentence vector. To determine the energy value, the feature channels are sent into a Multi-Layer Perception (MLP) consisting of a single hidden layer. The energy value is then turned into probabilities of attention. At the end, the feature channels are combined with the images feature maps and treated as input to the subsequent down-sampling blocks to synthesize a comprehensive global feature representation. The function of the soft threshold then predicts attention map probability to stabilize the RATGAN training process.

Before normalization, a logistic function is used to compress the negative energy values in range (0,1). The popular softmax function wasn't the one used, since it maximizes the largest probability while suppressing other probabilities to be close to zero. The exceptionally low probabilities impede gradient backpropagation, exacerbating the challenges in stabilizing GAN training. The soft threshold function, on the other hand, keeps attention probabilities from approaching zero and boosts the effectiveness of backpropagation. More text features are assigned to relevant image areas by the spatial attention module, which aids the discriminator in determining if the text image couple matches. The more powerful the discriminator in adversarial training encourages the generator to generate more realistic resolution with meaningful content. The final equation of the Discriminator is demonstrated in "Equation 2", where p is the probability value of x at k-th element which is the input value or score, j is the index variable used for summation K represents the total number of elements being considered.

$$p(x_k) = \frac{\frac{1}{1+e^{-x_k}}}{\sum_{j=1}^{K} \frac{1}{1+e^{-x_j}}} \qquad (2)$$

3.5. Fusion of AraBERT in RATGAN

We have used AraBERT, a strong architecture that has been pre-trained on billions of Arabic words, to generate words/sentence embeddings of the text descriptions without the need for training it. We minimized the size of the output sentence vector in order to maintain the shape consistency of the model architecture shape. As shown in Figure (1), we injected the sentences vector into the LSTM modules which describes the extended-range dependencies and relationships among fusion blocks, and was injected into the Discriminator, which has spatial attention to enhance the alignment of meaning between generated images and their corresponding textual descriptions, and surpasses sentence vectors for unimportant regions which led to enhance the generated images.

4. Results and Discussion

This section presents an overview of the dataset, training settings, and evaluation metrics used to evaluate our model, then our proposed framework in a qualitative and quantitative manner.

Datasets: Our experimental results were conducted on CUB dataset. CUB has a wide range of 200 different birds' categories with a total of nearly 12k images. The data was partitioned into training and testing sets. With 150 train classes and 50 classes used for testing. It has 10 captions/text description for each image.

Training & Evaluation details: During Ara-RATGAN training, the AraBERT text encoder model weights were frozen at an output size of 256px. 100-dimensional Gaussian distribution was used to sample the random noise vector. To ensure stability during the training process and optimize the network, AdamW optimizer was employed. Learning rates of 0.009 and 0.00045 were used for the generator and discriminator respectively. CUB dataset was used to train the Arabic RATGAN model for 700 epochs using a batch-size of 16. The training phase took around 3 days on a single 3080Ti GPU. The evaluation metrics used were Inception Score (IS), and Fréchet Inception Distance (FID), to measure the quantitative results. A high IS suggests that the synthesized images are of excellent quality. It's calculated using Eq 3. Where $(\mathrm{p}(\mathrm{y}|\mathrm{x})$ is the conditional probability distribution of labels y given an image x, p(y) the marginal probability distribution of labels across the generated images, the Kullback-Leibler divergence $\mathrm{DKL}(\mathrm{p}(\mathrm{y}|\mathrm{x}) \ || \ \mathrm{p}(\mathrm{y})))$between these distributions, N is the number of generated images being evaluated.

$$\mathrm{IS} = \exp(\mathrm{Ex}\ \mathrm{DKL}(\mathrm{p}(\mathrm{y}|\mathrm{x}) \ || \ \mathrm{p}(\mathrm{y}))) \qquad (3)$$

The FID computes the Fréchet Distance between the synthesized images and the real ones' feature distributions. A lower FID denotes that the synthesized images are more similar to the original real images. It's calculated using Eq 4:

$$\mathrm{F}(\mathrm{r},\mathrm{g}) = ||\mu_\mathrm{r} - \mu_\mathrm{g}||^2 + \mathrm{tr}(\Sigma_\mathrm{r} + \Sigma_\mathrm{g} - 2(\Sigma_\mathrm{r}\ \Sigma_\mathrm{g}\ )^{1/2})) \qquad (4)$$

μ_r is the feature vector mean, μ_g is the mean of all generated images' feature vector, Σ_r, Σ_g is the covariance matrix of feature vector of real and generated images respectively.

4.2. Qualitative Results

In this part, we demonstrate the output visualization of our framework on Arabic input captions from CUB dataset as demonstrated in Figure 2. As we can see our model has a realistic image generation, with more pertinent content; as a result of using fusion blocks, which operate as a single entity, what allows Arabic RATGAN to realize more intricate control over synthesized images from Arabic Text captions.

4.1. Quantitative Results

In this part, we assess our model performance using IS and FID of previous work of Arabic DF-GAN [31]. Our model uses Arabic Text description as a generation input caption. IS and FID attained on CUB dataset using previous work and our Ara-RATGAN model in are listed in Table 1. We achieved 4.25 on IS and 27.32 on FID, meanwhile Arabic DF-GAN achieved 3.51 on IS and 55.96 on FID. Our model has superior performance compared with DF-GAN. Also, RATGAN generates high quality and realistic images from Arabic text captions.

Table 1: Performance results of FID and IS of our model and previous English and Arabic text-to-image models

Model	Language	IS ↑	IS ↑	FID ↓
StackGAN [2]	English	3.7	3.7	51.89
AttGAN [3]		4.36	4.36	23.98
DF-GAN [9]		**5.10**	**5.10**	14.81
DMGAN [27]		4.75	4.75	16.09
MirrorGAN [4]		4.56	4.56	18.34
TextControlGAN [28]		4.41	4.41	57.92
MJ-GAN [29]		4.62	4.62	10.38
HDGAN [30]		4.15	4.15	--
Arabic DF-GAN [31]	Arabic	3.51	3.51	55.96
Ara-RATGAN *(Ours)*		**4.25**	**4.25**	**27.32**

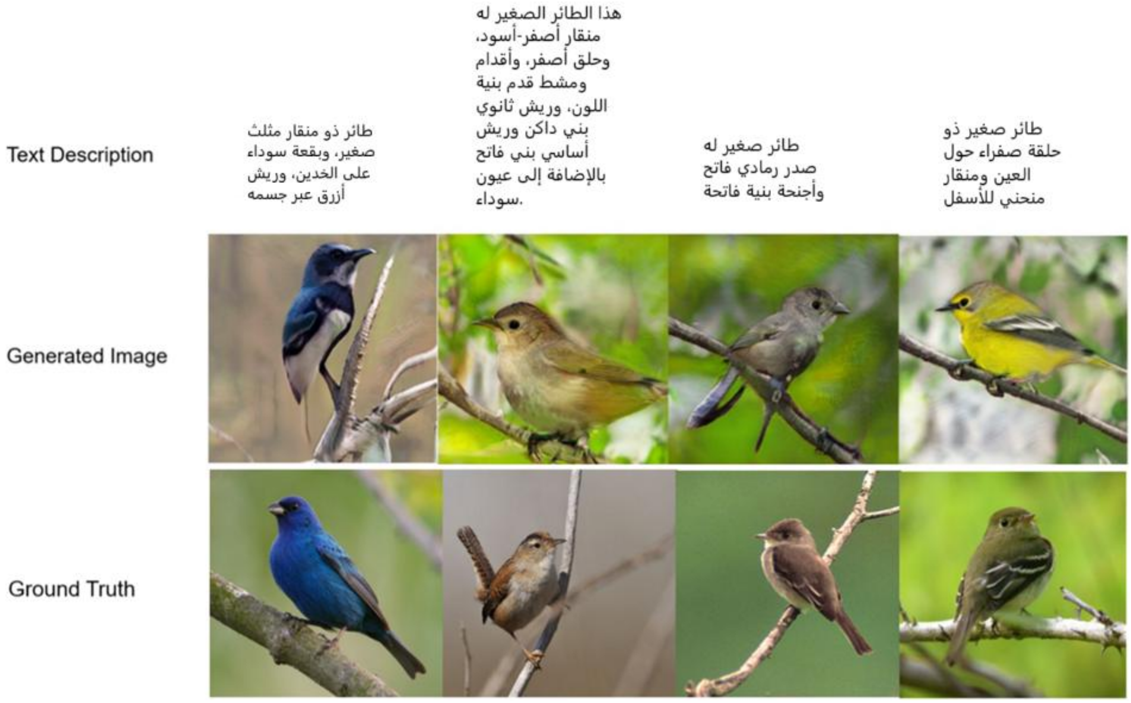


Figure 2: Qualitative results of Ara-RATGAN model.

5. Conclusion

In this research, we introduced a robust framework with Recurrent Affine Transformation for generating images with detailed and realistic text-image matching that is conditioned on Arabic-language text description. To attain this results, we fused the translated text description information, which is translated using Helsinki-NLP opus-mt-en-ar model, which converted CUB-200 Birds dataset from English language to Arabic language into the synthesizing process, which successfully enhanced the fidelity and detail level of the generated images by the interactions between fusion blocks trough RNNs unlike the isolated fusion blocks in other models. This reduced the model collapse incidents by 41%. Besides ensuring consistency between neighboring blocks, the mutual interactions happening solve the difficulty of training.

Furthermore, to assess text-image semantic consistency, we used spatial attention module in the Discriminator of RATGAN. Finally, by fusing the words embedding vector extracted from AraBERT which is a transformer-based model, into RATGAN generator and discriminator, our approach achieved 4.25 on Inception Score (IS) and 27.23 on Fréchet Inception Score (FID) on CUB-200 Birds dataset, which shows improvement in text-image alignment by 32% over LSTM baseline. While Arabic text represents greater complexity than the English text, our approach shows great results in Arabic text to image generation tasks in comparison with other Arabic text to image methods.

6. Future Work

While our Ara-RATGAN framework demonstrates promising results in Arabic Text-to-Image synthesis, several directions warrant further investigation:

1. Expand our approach to the Arabic-COCO dataset to handle more complex and diverse descriptions.

2. Investigate multilingual transfer learning to further enhance the model's versatility across languages, and finetune the model parameters for further optimization.
3. Explore more advanced attention mechanisms (e.g. hierarchical attention) to enhance the model capabilities on the morphology of the Arabic language.
4. Integrating Diffusion models with RAT for improved fine-grained image details
5. Propose Evaluation metric benchmarks for Arabic language beyond IS/FID.
6. Adapting the model on low-resource languages, with few-shot finetuning for under resourced Arabic dialects

References

[1] Good fellow I, Abadie PJ, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in neural information processing systems. 2014, p. 2672–80.

[2] Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, 2017. p.5907–15.

[3] Xu T, Zhang P, Huang Q, Zhang H, Gan Z, Huang X, et al. 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. p. 1316–24.

[4] Qiao T, Zhang J, Xu D, Tao D. Mirrorgan: Learning text- to-image generation by redescription. In: Proceedings of the IEEE conference on computer vision an pattern recognition, 2019. p. 1505–14.

[5] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In ICLR. OpenReview.net, 2017.

[6] Harm de Vries, Florian Strub, Jéremie Mary, Hugo Larochelle, Olivier Pietquin, & Aaron C. Courville (2017). Modulating early visual processing by language. CoRR, abs/1707.00683.

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In ICLR, 2019.

[8] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, & Timo Aila. (2023). StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis.

[9] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Fei Wu, & Xiao-Yuan Jing (2020). DF-GAN: Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis. CoRR, abs/2008.05865.

[10] Zhenxing Zhang and Lambert Schomaker. DTGAN: dual attention generative adversarial networks for text to-image generation. In IJCNN, pages 1–8. IEEE, 2021.

[11] Kai Hu, Wentong Liao, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware GAN. CoRR, abs/2104.00567, 2021.

[12] Senmao Ye, Fei Liu, & Minkui Tan. (2022). Recurrent Affine Transformation for Text-to-image Synthesis.

[13] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset. California Institute of Technology.

[14] Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-Term Memory. Neural Computation. 9. 1735-1780. 10.1162/neco.1997.9.8.1735.

[15] Nicolas Zucchet, Seijin Kobayashi, Yassir Akram, Johannes von Oswald, Maxime Larcher, Angelika Steger, & João Sacramento. (2024). Gated recurrent neural networks discover attention.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, & Illia Polosukhin (2017). Attention Is All You Need. CoRR, abs/1706.03762.
[17] Wissam Antoun, Fady Baly, & Hazem M. Hajj (2020). AraBERT: Transformer-based Model for Arabic Language Understanding. CoRR, abs/2003.00104.
[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805.
[19] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, & Colin Raffel (2020). mT5: A massively multilingual pre-trained text-to-text transformer. CoRR, abs/2010.11934.
[20] Chen, M., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Schuster, M., Shazeer, N., Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Chen, Z., Wu, Y., & Hughes, M. (2018). The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 76–86). Association for Computational Linguistics.
[21] Tiedemann, J., Aulamo, M., Bakshandaeva, D., Boggia, M., Grönroos, S.A., Nieminen, T., Raganato, A., Scherrer, Y., Vazquez, R., & Virpioja, S. (2023). Democratizing neural machine translation with OPUS-MT. Language Resources and Evaluation(58), 713–755
[22] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. CoRR, abs/1411.1784, 2014.
[23] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In ICLR. OpenReview.net, 2017
[24] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, & Dimitris N. Metaxas (2017). StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. CoRR, abs/1710.10916.
[25] Cristian Bodnar (2018). Text to Image Synthesis Using Generative Adversarial Networks. CoRR, abs/1805.00676.
[26] Xiaolong Wang, & Abhinav Gupta (2016). Generative Image Modeling using Style and Structure Adversarial Networks. CoRR, abs/1603.05631.
[27] Minfeng Zhu, Pingbo Pan, Wei Chen, & Yi Yang (2019). DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis. CoRR, abs/1904.01310.
[28] Ku H, Lee M. TextControlGAN: Text-to-Image Synthesis with Controllable Generative Adversarial Networks. Applied Sciences. 2023; 13(8):5098. https://doi.org/10.3390/app13085098
[29] Zhang, Zhiqiang and Gao, Yufei and Yu, Wenxin and Zhou, Jinjia, MJ-GAN: Multilevel Judgment Generative Adversarial Networks for Text-to-Image Synthesis. Available at SSRN: http://dx.doi.org/10.2139/ssrn.4431885
[30] Zizhao Zhang, Yuanpu Xie, & Lin Yang (2018). Photographic Text-to-Image Synthesis with a Hierarchically-nested Adversarial Network. CoRR, abs/1802.09178.
[31] Bahani, Mourad & El Ouaazizi, Aziza & Maalmi, Khalil. (2022). AraBERT and DF-GAN fusion for Arabic text-to-image generation. Array. 16. 100260. 10.1016/j.array.2022.100260.