Recognizing geographical locations using a GAN-based text-to-image approach

Dina M. Ibrahim, Amal A. Al-Shargabi
Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia

Article Info

Article history:

Received Jun 11, 2024
Revised Sep 13, 2024
Accepted Sep 30, 2024

Keywords:

AttnGAN model
Deep learning
Generative adversarial networks
Location-based services
Road infrastructure
Text reading
Text-to-image

ABSTRACT

Generating photo-realistic images that align with the text descriptions is the goal of the text-to-image generation (T2I) model. They can assist in visualizing the descriptions thanks to advancements in machine learning algorithms. Using text as a source, generative adversarial networks (GANs) can generate a series of pictures that serve as descriptions. Recent GANs have allowed oldest T2I models to achieve remarkable gains. However, they have some limitations. The main target of this study is to address these limitations to enhance the text-to-image generation models to enhance location services. To produce high-quality photos utilizing a multi-step approach, we build an attentional generating network called AttnGAN. The fine-grained image-text matching loss needed to train the AttnGAN's generator is computed using our multimodal similarity model. With an inception score of 4.81 on the PatternNet dataset, our AttnGAN model achieves an impressive R-precision value of 70.61 percent. Because the PatternNet dataset comprises photographs, we've added verbal descriptions to each one to make it a text-based dataset instead. Many experiments have shown that AttnGAN's proposed attention procedures, which are critical for text-to-image production in complex circumstances, are effective.

Corresponding Author:

Dina M. Ibrahim
Department of Information Technology, College of Computer, Qassim University
Buraydah, Saudi Arabia
Email: d.hussein@qu.edu.sa

1. INTRODUCTION

Synthesizing images from text descriptions has lately been a hot topic of research because of the introduction of generative adversarial networks. It's a flexible and easy method for creating conditional images, with considerable improvements in visual realism, diversity, and semantic congruence in recent years. However, the discipline still faces several obstacles that will necessitate additional research, such as permitting the creation of high-resolution photos containing many objects and generating appropriate and trustworthy evaluation criteria that correlate with human judgment.

Deep learning has enabled significant breakthroughs in computer vision applications and image processing techniques in recent years, and these advancements have had a significant impact on the field. For example, the field of picture synthesis, which is the act of creating new images and changing existing images, is one such example. In addition to its many practical uses in areas such as art creation, picture editing, virtual reality (VR), computer-aided design, and video games, image synthesis is an important task to perform.

In their 2014 article "Generative Adversarial Networks," Goodfellow *et al.* [1] detailed the design of the technique and provided the first empirical proof of it. In generative adversarial networks (GANs), the adversarial learning concept trains both the generator and the discriminator. A generator model is employed to produce images by taking input points from a latent space and combining them with a discriminator model that sorts images into two categories: real (from the dataset) and fake (from the generator) [1], [2]. Finding the possible distribution of actual data samples and then using that distribution to create new samples is what GAN is all about. There has been a great deal of research into GANs since its inception because of the vast potential uses for them in fields such as image and vision computing, and voice and language processing.

Due to the advent of GANs, supervised and unsupervised training of generative models for images are now feasible [1], [3]. GANs have aroused a great deal of attention and have resulted in significant advancements in image synthesis research. The picture synthesis challenge was framed as a game with two players between two competing artificial neural networks, according to the researchers. Unlike a discriminator network, a generator network is trained to generate realistic samples, whereas a discriminator network is trained to discern between actual and generated images [4], [5]. A discriminator network is developed to distinguish between genuine and produced images. The generator's training purpose is to deceive the discriminator during the training process. So far, this approach has been successfully used a number of applications, including representation learning, style transfer, image-to-image translation, data augmentation, high-resolution human face synthesis, image super-resolution, and image in-painting [6], [7].

The goal of text-to-image (T2I) synthesis is to generate an image that accurately reflects the meaning of a text description. Some of the most recent studies have concentrated on this topic [8], [9]. As an analogy, T2I can be viewed as the inverse of image captioning, in that it accepts an image as input and produces a text description of that picture as an output. However, while the approaches given in this study can be applied to a wide range of picture domains, the majority of T2I research is focused on methods for creating aesthetically realistic images that are photographic [10]. The goal of a T2I model is to produce images that are photorealistic and semantically compatible with the descriptions of the text [11]. They can assist in visualizing the descriptions thanks to advancements in machine learning algorithms. Using text as a source, GANs can generate a series of pictures that serve as descriptions. Unsupervised machine learning includes generative model methods [12]. New GANs have allowed current T2I models to achieve remarkable results. However, they have some limitations. A significant step toward artificial intelligence similar to that of humans can be achieved by developing a system that comprehends the interplay between visual perception, language, and images, and which can generate visual representations of the meaning conveyed by written descriptions. This study aims to address these limitations. The creation of GAN made unsupervised training of generative models for images possible. GANs have attracted the interest of many researchers and propelled the field's efforts in the area of image synthesis.

This study investigated the effects of an attentional generative adversarial network (AttnGAN) to synthesize fine-grained text into images. While earlier studies have explored the impact of a generative adversarial network, they have not explicitly addressed its influence on a multi-stage process, we build an attentional generating network to enable the AttnGAN to produce high-quality photos. In order to train the AttnGAN's generator, we provide a deep attentional multimodal similarity model that computes the fine-grained image-text matching loss.

2. BACKGROUND AND RELATED WORK IN T2I USING GANS

A recent study proposed in [13] query that GAN is a unique scene retrieval framework that uses AttnGAN images as query images. According to the findings of the experiments, the framework that has been proposed is capable of accurately retrieving scenes and enables users to locate the scenes that they have picked. They used two kinds of data sets. The COCO dataset consists of daily scene images and the annotations that go along with them. There are 82,783 training images in the dataset, each of which has five descriptions. Three movies were selected from the MP-II MD dataset: "BadSanta," As "Good As It Gets," and "Harry Potter and the Prisoner of Azkaban," which had 430, 538 and 592 scenes, respectively, with an average of 100 frames each scene and a total of 153,320 frames. Subjective evaluation results. The average scores from 25 subjects are shown in these results.

A score of 1 represents "not relevant," while a score of 5 represents "relevant," proposed framework (PF): 4.96, BL (baseline method): 4.76, comparative method 1 (CM1) is 3.20, comparative method 2 (CM2) is 4.48, comparative method 3 (CM3) is 4.44, and comparative method 4 (CM4) is 4.88. They used Recall@k for the quantitative evaluation criterion and defined frames included in the target scene as our ground truth.

Recall@k was obtained for the integrated dataset, which included five movies: "Bad Santa", "As Good as it Gets", "Halloween", "Rendezvous mit Joe Black", and "Harry Potter and the Prisoner of Azkaban". In the 60,000 rank, recall@k=.7. Additionally, by demonstrating the effectiveness of the suggested framework, it is possible to confirm the usefulness of the generated images, which are not aesthetically pleasing.

Zhang *et al.* [14], both rendering graphics from text descriptions and stylizing them based on a given style image were the goals of the authors. GAN systems can generate a stylized picture directly from an image and text description, rather than generating the image first and then applying style without having any prior knowledge of the text description. When the task was broken down into two halves, StackGAN was able to generate more realistic, higher-resolution images. In stage I GAN, low-quality visuals are generated based on the text to approximate the meaning (like the methods described above). In stage II GAN, the text and output from stage I are used to build a higher-resolution rendition with additional details. Data pairs in the type of text description, stylized image are required for their project's requirements. The "image captioning" data sets, which comprise data pairings in the form "text description, image," were not available, thus we used neural style transfer to generate styled images for these datasets.

Many image captioning databases are available, such as these: COCO has more than 200,000 images covering a wide range of subjects, each with five accompanying captions. CUB-200-2011: 6,033 images of birds from 200 different species. To train COCO, we needed a large number of photographs, and in our instance, we needed to generate stylized versions of all of these photos, possibly several times, in order to handle each of COCO's many styles. This is why we opted to not only use a smaller dataset, but also limit the number of species in the dataset, so that we could reduce training and data production time. In order to verify that each group had enough samples, we separated the data by species, not as a random sampling. For the first 451 photos, we created stylized images in two styles, which related to eight bird species (larger than average categories for the CUB dataset). Because it takes too long to collect enough stylised training data and execute a hyperparameter search with enough iterations each time to determine the ideal settings, no relevant quantitative results were obtained.

Similarly, attention-driven multi-stage refinement for fine-grained text-to-image generation is proposed by the authors in their work on the AttnGAN network, in [15]. By focusing on the important words in the textual description, the AttnGAN, a new attentional generative network, is able to generate fine-grained features at various picture sub-regions. This paper has used AttGAN for COCO dataset: train 8,855 images with 10 caption/image and CUB dataset: train 80K images with 5 caption/images. The result of the AttnGAN 4.36 ± .03 inception score and 67.82 ± 4.43 R-precision for CUB dataset and 25.89 ± .47 inception score and 85.47 ± 3.69 R-precision for COCO dataset. By improving the best reported inception score by 14.14% on the CUB dataset and 170.25% on the more difficult COCO dataset, our AttnGAN far surpasses state-of-the-art GAN models. Particularly important for complicated scene text-to-image generation, extensive experimental data show that the AttnGAN's postulated attention processes work.

An adversarial network (ACGAN) is designed to produce high-quality images with a resolution of 1024x1024 in [16]. The initial proposal is a multi-tiered cascade architecture for text-to-image synthesis. In order to train the model to produce images with photorealistic detail, we progressively add more layers, with each layer using the outputs and word vectors of the preceding layer as inputs. Second, to train the generation, we insert a deep attentional multimodal similarity model into the network and match word vectors and images in a common semantic space to calculate a fine-grained matching loss. Results show that the suggested model outperforms AttenGAN on the CUB dataset with an inception score of 4.48 and the Oxford-102 dataset with an inception score of 6.42 percent, respectively. When it comes to text-generated images, the ACGAN model produces better results and more accurate results.

They can assist in visualizing the descriptions thanks to advancements in machine learning algorithms. In order to generate a series of visual representations of textual descriptions, GANs can be employed [17]. A subset of unsupervised machine learning is generative model algorithms. The versatility and ease of use of GANs make them ideal for a wide range of tasks, including model creation and visualization based on user input. They have applications in house design and architectural planning. They can also be utilized to create animations, but they can speed up the process, making them ideal for use in virtual games. The advertising and clothing industries are two more that can benefit from GANs.

Comparing the suggested model's output to that of mirror GANs, the results are mixed. The proposed model appears to produce unique images that closely resemble birds, as indicated by the inception score of 5.089. The proposed model appears to produce unique images that closely resemble birds, as indicated by the inception score of 5.089. The generated images' variability increases as the inception score rises. In a similar vein, a high inception score indicates that the Google inception model made a precise classification.

Every machine learning problem core of every dataset. The three sets of extensively utilized datasets in T2I research include Oxford-120 Flowers [18], CUB-200 Birds [19], and COCO [20]. With each dataset consisting of approximately 10,000 photos, Oxford-102 Flowers [18] and CUB-200 Birds [19] are on the smaller side. Each image depicts a single object, and each image has 10 single sentence captions that have been collected using Amazon Mechanical Turk (AMT) workers in the US [21]. All of the employees were instructed to describe themselves in detail. In contrast, COCO [20] consists of approximately 123,000 photos, each of which has five captions that were generated by humans and collected via AMT. Every employee was given specific instructions to describe the scene in detail while avoiding certain terms and phrases. These included not using the word ”there” in the first line, not discussing anything that could happen in the past or future, not using formal names, and using a minimum of 8 words. The official 2014 COCO split is used in most of T2I works. Chen *et al.* [22] provide more details on the captions collection process. Unlike the Oxford-102 Flowers and CUB-200 Birds datasets, COCO images usually contain many, often interacting objects. Table 1 shows a list of the most common datasets used for T2I illustrating the total number of training, testing, captions, and object categories per each dataset.

Table 1. A list of the most common datasets for T2I approach

Dataset	Trained images	Tested images	Total images	Captions of image	Object category
Oxford-102 Flowers	7,034	1,155	8,189	10	102
CUB-200 Birds	8,855	2,933	11,788	10	200
COCO	82,783	40,504	123,287	5	80

For our proposed work model, input noisy images with suitable captions, generator network, discriminator network, and training methods are essential parts that should be focused on. In this study, our research method consists of six steps which are illustrated in Figure 1.

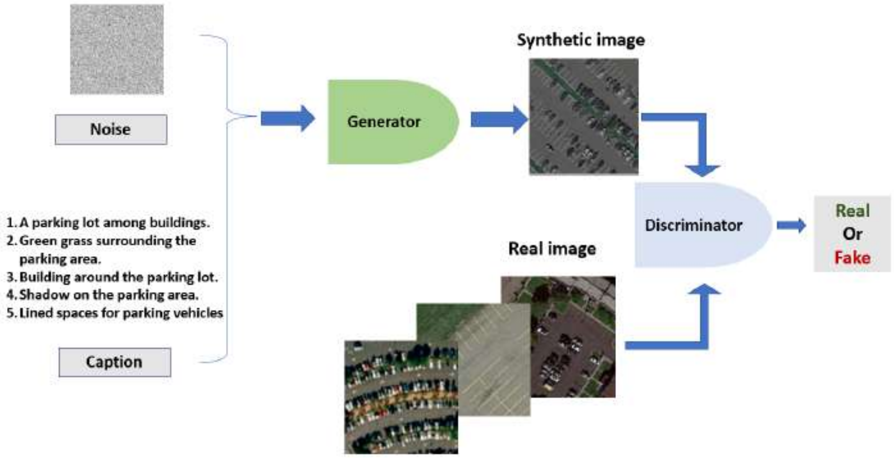
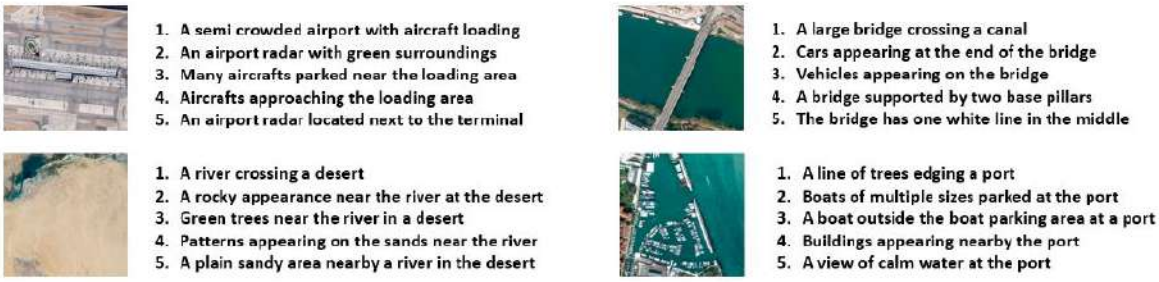

Noise
1. A parking lot among buildings.
2. Green grass surrounding the parking area.
3. Building around the parking lot.
4. Shadow on the parking area.
5. Lined spaces for parking vehicles
Caption
Generator
Synthetic image
Real image
Discriminator
Real
Or
Fake

Figure 1. Our proposed research method

- Step 1: input a randomly generated noise signal and text to the generator. The generator will generate random images that do not have any relevance to the real world.
- Step 2: now, we are going to start training the discriminator. We input the random images generated by the generator with a sample of the real images of the dataset.
- Step 3: the discriminator gives by then the probabilities of all the inputted images being real.
- Step 4: compare with the actual labels (1 for real images and 0 for fake images). Calculate the loss error and back-propagate it to update the weights of the discriminator.
- Step 5: run the generated images through the discriminator, but this time without the real images. The discriminator will most probably output probabilities of the images being fake.

– Step 6: we now compare the probabilities in the output with the expected label not the actual. Then, calculate the loss error and backpropagate it to update the weights of the generator.

Figure 2 shows samples of our expected proposed model. After training, the generator is able to generate realistic images that are independent from the real images (not exact, but close), and the discriminator thinks they're correct!



1. A semi crowded airport with aircraft loading
2. An airport radar with green surroundings
3. Many aircrafts parked near the loading area
4. Aircrafts approaching the loading area
5. An airport radar located next to the terminal

1. A large bridge crossing a canal
2. Cars appearing at the end of the bridge
3. Vehicles appearing on the bridge
4. A bridge supported by two base pillars
5. The bridge has one white line in the middle

1. A river crossing a desert
2. A rocky appearance near the river at the desert
3. Green trees near the river in a desert
4. Patterns appearing on the sands near the river
5. A plain sandy area nearby a river in the desert

1. A line of trees edging a port
2. Boats of multiple sizes parked at the port
3. A boat outside the boat parking area at a port
4. Buildings appearing nearby the port
5. A view of calm water at the port

Figure 2. Examples of our predicted output images

The performance of the T2I models based on GAN techniques is evaluated based on several metrics: inception score (IS), Fréchet inception distance (FID), R-precision, and finally human rank. Score for IS: generative adversarial network models produce synthetic images, and the IS [23] is a measure that objectively assesses their quality. Two characteristics of a set of produced images—their diversity and their quality—are aimed at by the score.

The FID [24] measures how far apart the distributions of real and fake images are for features that a network that has already undergone training pulls out. When comparing GANs, the FID is more reliable than the IS and can detect a wider variety of disruptions. The visual-semantic similarity between text descriptions and generated images is measured by R-precision [15], which ranks retrieval results based on extracted picture and text attributes.

Human rank: for qualitative evaluation, human rank is used. In the CUB and Oxford test sets, 50 text descriptions were chosen at random, and the created model generated 5 images for each sentence [16]. The five images and accompanying sentences are explained to several people, who then rank the image quality in various ways before calculating an average ranking to assess the quality and diversity of the created images. Table 2 presents a summary of the GAN models' performance metrics.

Table 2. A summary of the GAN models performance metrics

Model	FID	IS	R-precision	Human rank	Dataset
GAN-INT-CLS	68.79	2.88+-.04	-	-	CUB
	60.62	7.88+-.07	-	-	COCO
GAWWN	-	3.62+-.07	-	-	CUB
StackGAN	51.89	3.70+-.40	-	-	CUB
	74.05	8.45+-.03	-	-	COCO
StackGAN++	15.30	4.04+-05	-	-	CUB
	81.59	8.30+-.10	-	-	COCO
HDGAN	-	4.15+-.05	-	-	CUB
	-	11.86+-.18	-	-	COCO
AttnGAN	25.72	4.36+-.03	67.82+-4.43	-	CUB
	35.49	25.89+-.47	85.47+-3.69	-	COCO
MirrorGAN	26.80	4.56+-.05	86.0+-.55	-	CUB
	37.86	26.47+-.41	86.0+-1.05	-	COCO
DM-GAN	16.09	4.75+-.07	72.31+-.91	-	CUB
	32.64	30.49+-.57	88.56+-.28	-	COCO
Obj-GAN	-	-	-	-	CUB
	21.21	32.79+-.21	93.39+-2.08	-	COCO
GAN-INT-CLS	-	2.88 ± .04	-	2.81 ± .03	CUB
	-	2.66 ± .03	-	1.87 ± .03	Oxford-102
	-	7.88 ± .07	-	1.89 ± .04	ms COCO
GAWWN	-	3.62 ± .07	-	1.99 ± .04	CUB

3. MATERIALS AND METHODS

3.1. Datasets for the study

In this work, We will apply T2I-AttnGAN to the satellite data set, to get the needed datasets for this study we used two well-known different scene datasets, the PatternNet dataset, and the Merced dataset and we proposed this dataset to enhance map services.

3.1.1. The PatternNet dataset

Image retrieval using remote sensing was the purpose of collecting the large-scale, high-resolution remote sensing dataset known as PatternNet [25]. There are 38 different categories, and each one has 800 images that are 256x256 pixels in size. Images used by PatternNet for several US cities were sourced from Google Earth via the Google Maps API. Figure 3 shows some examples of each class's images.

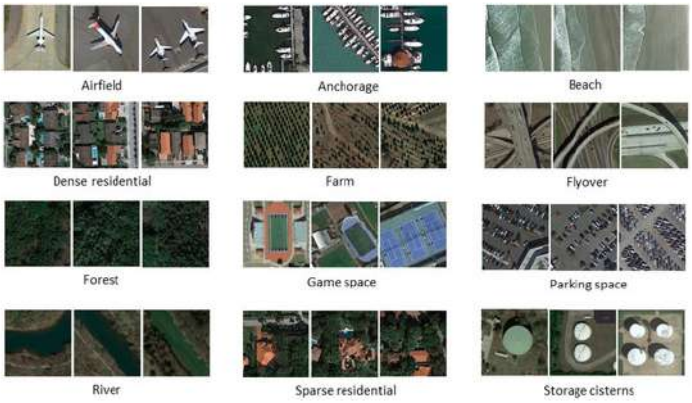


Figure 3. Sample of the PatternNet dataset

3.1.2. The Merced dataset

The UC-Merced land use dataset, which is famous for its 2100 high-resolution remote sensing photos organized into 21 classes is the first [26]. Figure 4 shows some examples of each class's images. This dataset is difficult to classify since several classifications, such as ”dense,” ”medium,” and ”sparse” residential, have a lot of overlap. This dataset has been used extensively to test techniques for classifying aerial scenes.

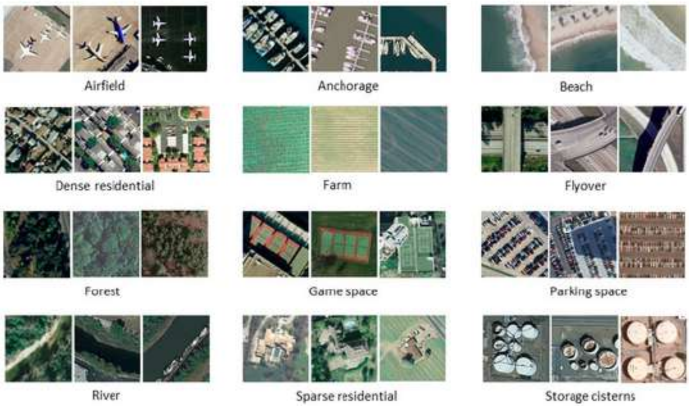


Figure 4. Sample of the Merced dataset

The datasets used in this study are the satellite datasets to enhance map services. During the annotation, a few guidelines that needed to be adhered to when the sentences were being generated were considered:

- Pay attention to the most prominent items; little ones might not be worth your time.

- Refrain from drawing attention to things' hues (like blue cars) and instead focus on their density and presence. To differentiate it from any generic parking lot (downtown, for example), it is necessary to add the word ”airport” when referring to a parking lot located in an airport.
- Stay away from conjunctions and punctuation.

3.2. Our AttnGAN-based T2I proposed model

Figure 5 shows the name of the GAN model employed in the suggested system, which is AttnGAN. The AttnGAN and the deep attentional multimodal similarity model (DAMSM) are the two main parts of this model. The planned AttnGAN's architecture is shown in the Figure 5. The DAMSM supplies the generative network with fine-grained picture-text matching loss, while each attention model autonomously collects the criteria (i.e., the most relevant word vectors) for creating distinct image sub-regions.

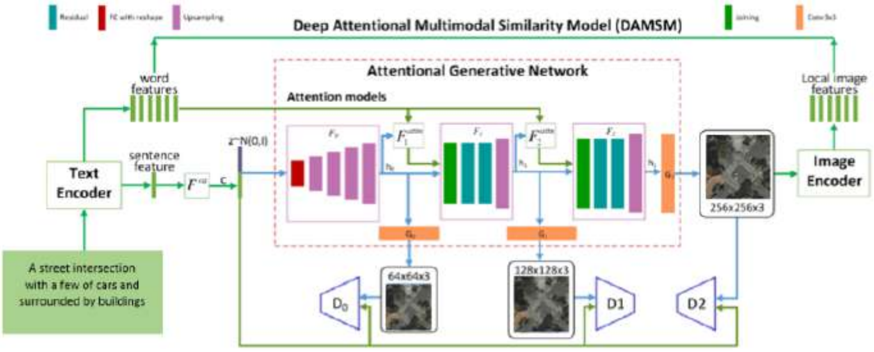


Figure 5. The architecture of the proposed AttnGAN model

3.3. Attentional generative adversarial networks

GAN form the basis of the most current text-to-image synthesis algorithms. The criterion for GAN-based picture generation is often the complete text description encoded into a global phrase vector. Despite the remarkable outcomes, producing high-quality images is hindered by conditioning GAN solely on the global sentence vector, which does not provide crucial word-level fine-grained information [15]. An adversarial attention-generating network to generate images from text with a finer level of detail, AttnGAN enables attention-driven multi-stage refining. There are two new parts to the model. As a first step, we build an attentional generative network, which uses a pre-existing attention mechanism to selectively generate visual regions by highlighting words with high contextual relevance. More precisely, every word in the sentence is encoded into its own word vector in addition to the global sentence vector that contains the natural language description.

Starting with a low-resolution image, the generative network uses the global sentence vector. Afterwards, it forms a word-context vector by querying word vectors using an attention layer with the picture vector in each sub-region [15]. The model then uses a multi-modal context vector that is formed by combining the regional image vector with the matching word-context vector to produce additional image features in the nearby sub-regions. A higher-resolution image with more information at each level is efficiently produced by this. DAMSMs are the other parts of the AttnGAN. An attention method allows the DAMSM to use both coarse-grained word-level information and global sentence-level information to calculate the sentence-to-generated-image similarity. To train the generator, the DAMSM adds a fine-grained image-text matching loss. Attentional Model: In order for the attentional generative network to produce realistic visuals with both sentence-level and word-level requirements, the final goal function is defined as:

$$L = L_G + \lambda L_{DAMSM}, \; whrer \; L_G = \sum_{i=0}^{m-1} L_{G_i} \tag{1}$$

Here, the hyperparameter λ is used to make sure that the two parts of the equation are balanced. The first term is the GAN loss that jointly approximates conditional and unconditional distributions [15]. Generator model: the AttnGAN's generator Gi is paired with its discriminator Di during the i-th step. Model for Discriminator: In contrast to training Gi, each discriminator Di is taught to minimize the cross-entropy loss in order to identify inputs as either real or fake.

3.4. Deep attentional multimodal similarity model

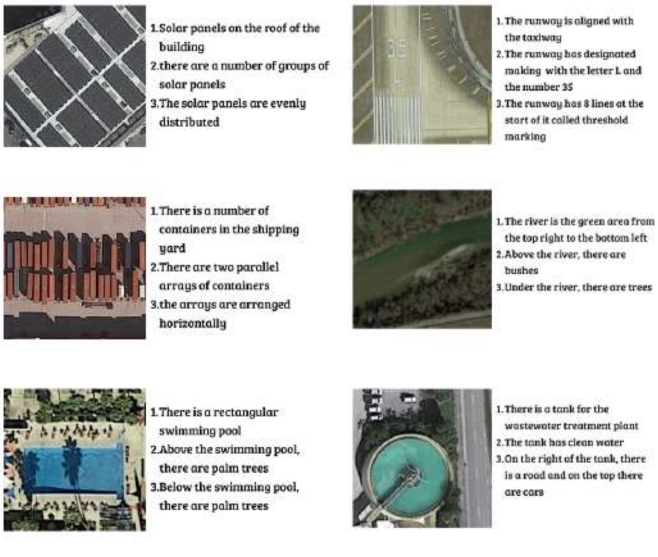
Using a common semantic space, the DAMSM learns to associate image and text elements with one another using two neural networks. As a result, it is able to generate images with a finer grain of loss by determining the degree of similarity between the image and text at the word level. By utilizing bidirectional long short-term memory (LSTM), the text encoder is able to extract semantic vectors from the text description. Two hidden states, one for each direction, are associated with each word in the bidirectional LSTM. Afterwards, we build an attention model that combines the two hidden states (query) of each word to produce a region-context vector [15]. In order to convert pictures into meaningful vectors, the picture encoder uses a convolutional neural network (CNN). Various sub-regions of the image are learned by the CNN's intermediate layers, while the image's global features are learned by the subsequent layers.

We present the attentional generative network (AttnGAN) and the DAMSM, two new attention models with distinct functions; i) the AttnGAN is able to autonomously generate various picture sub-regions by using the attention mechanism in the generative network to choose word-level conditions and ii) applying an attention mechanism enables the DAMSM to calculate the LDAMSM, a loss function for fine-grained text-image matching. Note that LDAMSM is only applied to the Gm-1 output since the final purpose of the AttnGAN is to produce big images from that generator. We made an effort to use LDAMSM on all resolution photos produced by (G0, G1,..., Gm-1). The computational cost went up, nevertheless, and performance stayed the same.

3.5. Our dataset PatternNet

Several guidelines for the construction of the sentences were considered during the annotating process: the first rule of focus is to ignore minor details in favor of the major, dominating items. Second, focus on describing what is actually there in the situation rather than what is not. Thirdly, instead of fixating on the quantity of objects, try using more general terms like "many", "few", and "many". Fourth, focus on the presence and density of objects rather than their color (like blue cars). Fifth, the word "airport" must be included when referring to a parking lot within an airport in order to differentiate it from any other type of parking lot, such as one located downtown. Then sixth, avoid using conjunctions or punctuation. In Figures 6 and 7, readers will observe a few results from our dataset.

In our study, we have only used images from PatternNet datasets from [25] and we added a textual description for each image to transform it into a text-based dataset. There are a total of 38 classes in PatternNet, with 800 images of 256 × 256 pixels each. The total number of photos we used was 4,800, as shown in Table 3, because we picked 24 classes and 200 images from each class of these datasets. We created two folders (PatternNet) containing 24 classes for the images used and another folder (PatternNet.text) containing 24 classes for the description in which we annotated the images. In the 200 images we chose from 24 different classes we annotated each image in a separate text file with three different sentences; 14,400 sentences were handed down as a result.



1.Solar panels on the roof of the building
2.there are a number of groups of solar panels
3.The solar panels are evenly distributed

1.The runway is aligned with the taxiway
2.The runway has designated making with the letter L and the number 35
3.The runway has 8 lines at the start of it called threshold marking

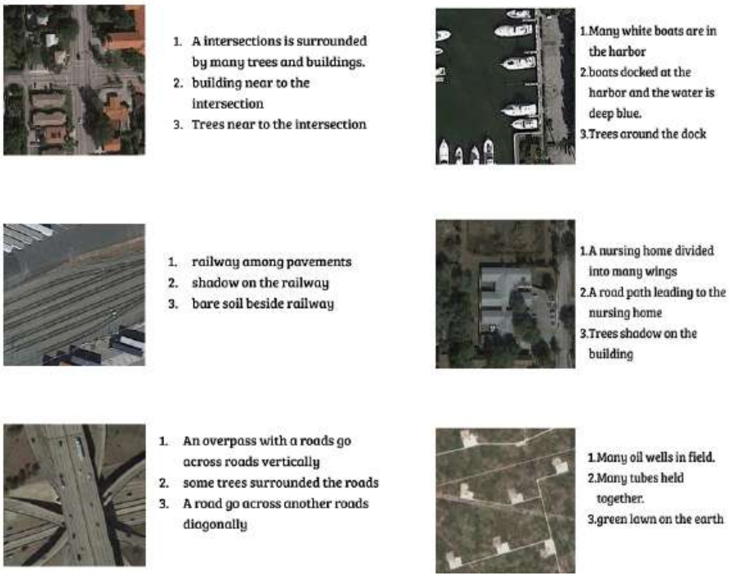
1.There is a number of containers in the shipping yard
2.There are two parallel arrays of containers
3.the arrays are arranged horizontally

1.The river is the green area from the top right to the bottom left
2.Above the river, there are bushes
3.Under the river, there are trees

1.There is a rectangular swimming pool
2.Above the swimming pool, there are palm trees
3.Below the swimming pool, there are palm trees

1.There is a tank for the wastewater treatment plant
2.The tank has clean water
3.On the right of the tank, there is a road and on the top there are cars

Figure 6. Example for images with three sentences for each image



1. A intersections is surrounded by many trees and buildings.
2. building near to the intersection
3. Trees near to the intersection

1.Many white boats are in the harbor
2.boats docked at the harbor and the water is deep blue.
3.Trees around the dock

1. railway among pavements
2. shadow on the railway
3. bare soil beside railway

1.A nursing home divided into many wings
2.A road path leading to the nursing home
3.Trees shadow on the building

1. An overpass with a roads go across roads vertically
2. some trees surrounded the roads
3. A road go across another roads diagonally

1.Many oil wells in field.
2.Many tubes held together.
3.green lawn on the earth

Figure 7. Another example for images with three sentences for each image

Table 3. Dataset classes in our study

38 original classes	24 classes we chose
Beach	Airplane
Chaparral	Baseball field
costal mansion	Basketball court
forest	Closed road
cemetery	Crosswalk
Golfe course	Dense residential
Oil well	Ferry terminal
Bridge	Football field
Parking space	Harbor
Freeway	Intersection
Runway marking	Mobile home park
tennis court	Nursing home
storge tank	Oil gas field
Football field	Overpass
Ferry terminal	Parking lot
Dense residential	Railway
Crosswalk	River
Closed road	Runway
Basketball court	Shipping yard
Baseball field	Solar panel
Airplane	Sparse residential
Harbor	Swimming pool
Intersection	Transformer station
Mobile home park	Wastewater treatment plant
Nursing home	-
Oil gas field	-
Overpass	-
Parking lot	-
River	-
Runway	-
Shipping yard	-
Solar panel	-
Sparse residential	-
Swimming pool	-
Transformer station	-
Wastewater treatment plant	-

3.6. The experiment of our proposed model

To assess the planned AttnGAN, a large amount of experimentation is conducted. We begin by delving into the AttnGAN's foundational elements, which encompass the attentional generating network and the DAMSM. Afterwards, we test our AttnGAN on the PetternNet dataset, which converts text to images.

Our dataset contains 24 classes each class have 200 images, so the total number of images are 4,800 images. Each image has its own description in a separate text file with three sentences. We divided the dataset into 80% for training and 20% for testing. This division is executed in each class from the 24 classes where from each class 160 images with their text used for training and 40 images with their text used for testing. So, the total number of the training images is 3,840 images with their text and the total number of the testing images is 960 images with their text. Table 4 lists the statistics of datasets.

Table 4. PatternNet dataset statistics

Dataset	Training	Testing
For each class from the 24 classes	160 (image + text)	40 (image + text)
Total	3840 (image + text)	960 (image + text)

Here are some examples of the outcomes obtained by modifying the text descriptions using our AttnGAN model that was trained on PatternNet. We put examples from four classes: airplane class, dense_residential, harbor class, swimming_pool. Figure 8 shows examples of our AttnGAN model trained on PatternNet from the airplane class using the sentences:

- Airplane on runway.
- White plane.
- Red wings.

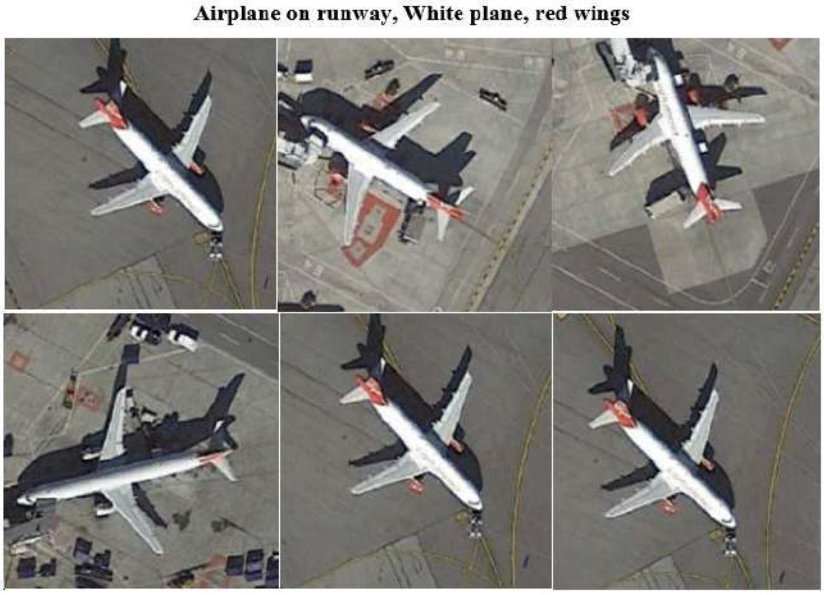


Figure 8. Results obtained by modifying the text descriptions of the airplane class using our AttnGAN model trained on PatternNet as an example

As in Figure 9, examples of our AttnGAN model trained on PatternNet from the dense_residential class are represented using the sentences:

- Buildings same in shape.
- Water pools appearing beside the buildings.
- Trees among the buildings.

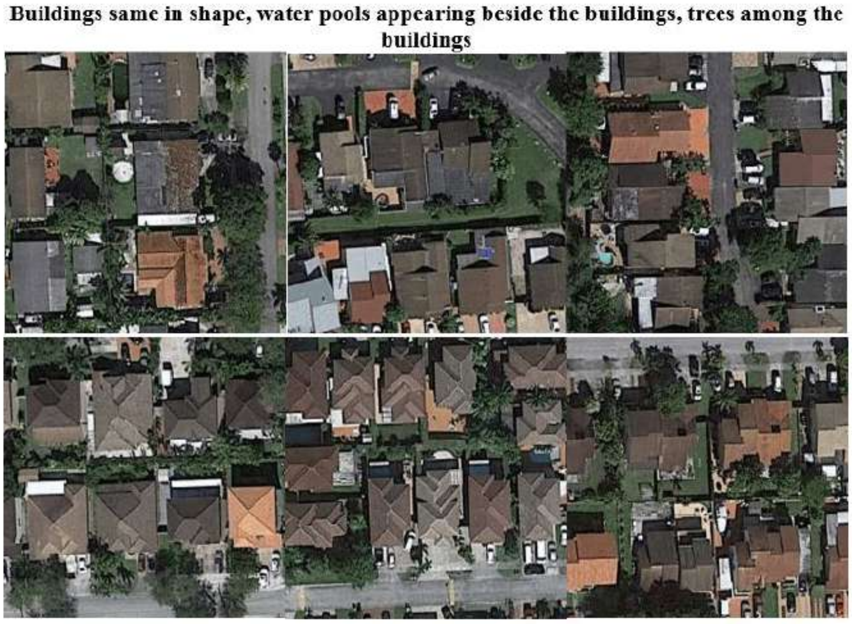

Buildings same in shape, water pools appearing beside the buildings, trees among the buildings

Figure 9. The dense residential class was used to train our AttnGAN model on PatternNet. We then changed a few of the most attended words in the text descriptions to see the outcomes

Figure 10 demonstrates examples of our AttnGAN model trained on PatternNet from the harbor class are represented using the sentences:

- Many white boats are in a harbor.
- A large number of boats moored at the harbor.
- The harbor is teeming with boats, and the water is a stunning shade of blue.

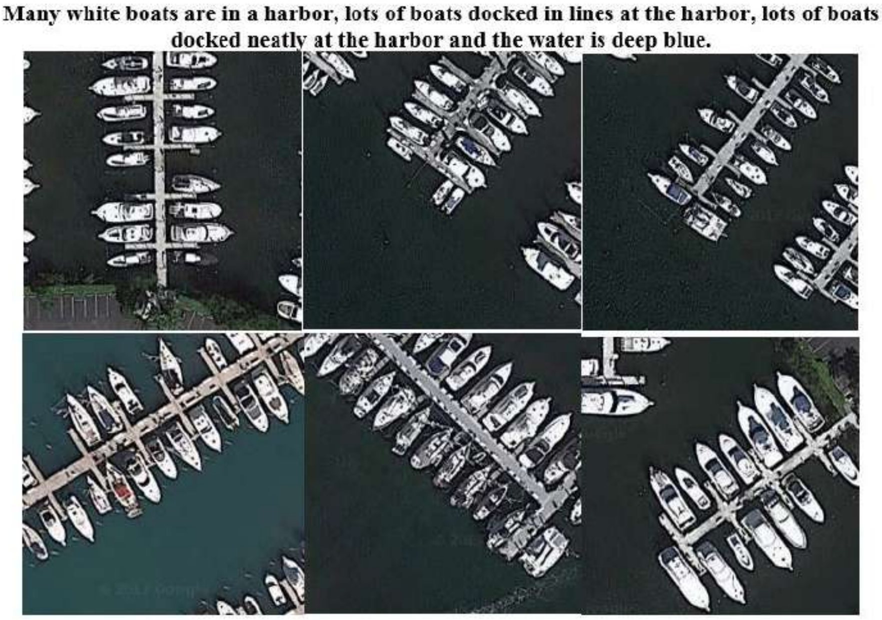
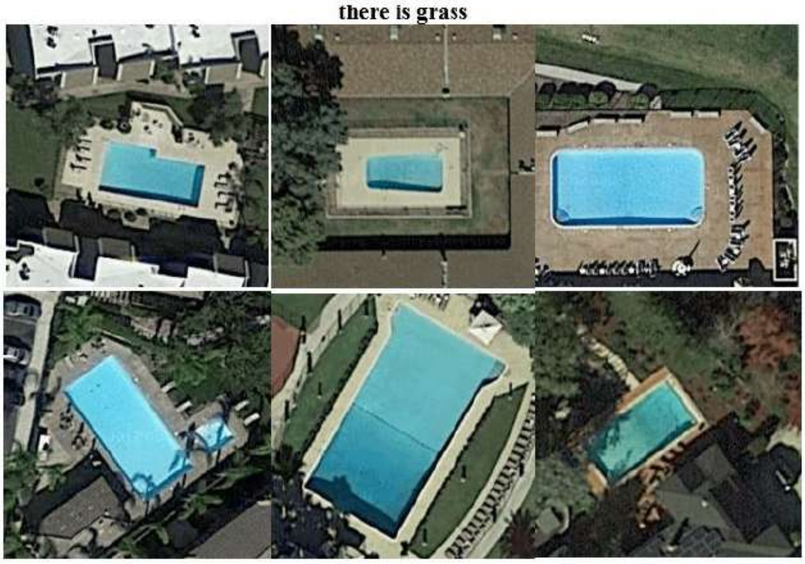

Many white boats are in a harbor, lots of boats docked in lines at the harbor, lots of boats docked neatly at the harbor and the water is deep blue.

Figure 10. AttnGAN model trained on PatternNet on harbor class with some most attended words changed in text descriptions

Finally, Figure 11 shows examples of our AttnGAN model trained on PatternNet from the swimming_pool class are represented using the sentences:

- There is a rectangular swimming pool.
- There is water in the pool.
- Right and left of the pool there is grass.



There is a rectangular swimming pool, there is water in the pool, right and left of the pool there is grass

Figure 11. Results of our PatternNet-trained AttnGAN model on the swimming pool class, adjusting the most frequently used terms in the text descriptions

4. RESULTS AND DISCUSSION

In order to quantify the evaluation, we utilize the inception score [15]. We propose adding R-precision, a popular metric for evaluating retrieval results, as an additional metric for the text-to-image synthesis job, as the inception score does not reveal how well the produced picture corresponds to the provided text description. By definition, the R-precision is r/R if there are R-relevant documents for a query and we look at the top R-ranked retrieval results of a system and discover that r is relevant.

Our study demonstrates that the incorporation of DAMSM, which enhances the model's performance, is the key differentiator of our model. We experiment with several values of λ in order to evaluate the suggested LDAMSM. To ensure the quality of the images, we determine the optimal value of λ for every dataset by gradually raising it until the overall inception score begins to decrease on a held-out validation set. Table 5 shows that there is a significant increase in both the inception score and R-precision with the value of λ.

Table 5. All AttnGAN models' top inception scores and R-precision rates on PatternNet datasets

Method	Inception score	R-precision(%)
AttnGAN, no DAMSM	4.01	15.24
AttnGAN, λ= 0.1	4.21	18.95
AttnGAN, λ= 1	4.54	40.63
AttnGAN, λ= 5	4.57	62.25
AttnGAN, λ= 10	4.81	70.61
AttnGAN, λ= 15	4.51	63.41

Figure 12 illustrates the inception score results of our AttnGAN model trained on PatternNet by applying different values of λ. As shown in the figure, the highest value of the inception score is 4.81, obtained when we adjust the value of λ = 10. When we try to increase the value of λ to 15 the model inception score decreases. The R-precision results are demonstrated in Figure 13. The highest value of the inception score is 70.61 obtained when we adjust the value of λ = 10. When we try to increase the value of λ to 15 the R-precision value is decreased.

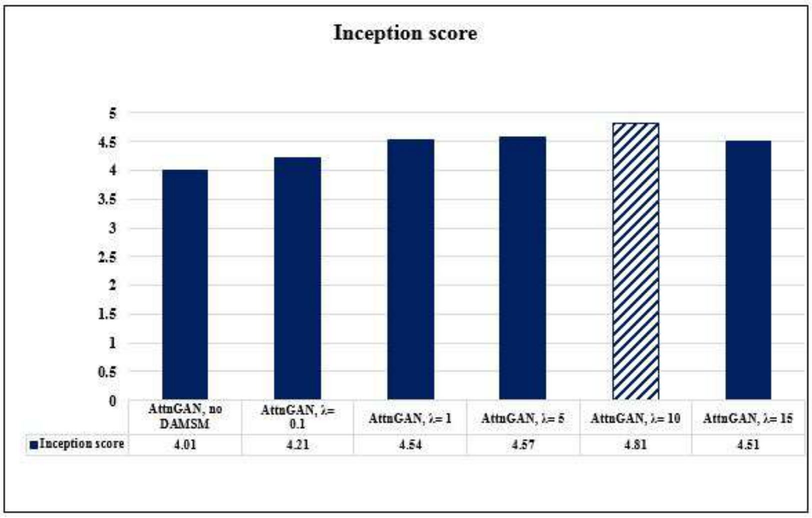


Figure 12. Inception score results of our AttnGAN model trained on PatternNet by applying different values of λ

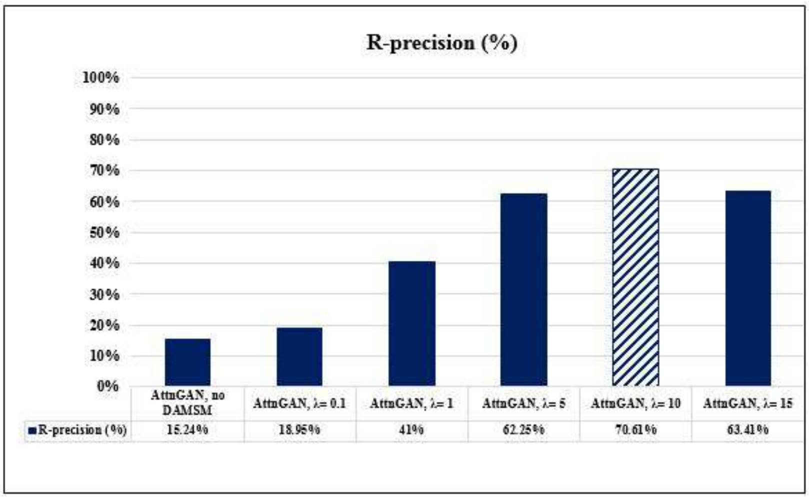


Figure 13. R-precision results of our AttnGAN model trained on PatternNet by applying different values of λ

5. CONCLUSIONS AND FUTURE WORK

This research suggests using an AttnGAN to synthesize fine-grained text into images. Using a multi-stage process, we build an attentional generating network to enable the AttnGAN to produce high-quality photos. In order to train the AttnGAN's generator, we provide a deep attentional multimodal similarity model that computes the fine-grained image-text matching loss. Our AttnGAN obtains an inception score of 4.81 on the PatternNet dataset and the R-precision value is 70.61%, which is a notable achievement. The PatternNet dataset consists solely of photos; thus, we add a textual description for each image to transform it into a text-based dataset. Extensive experimental results indicate the efficacy of the suggested attention processes in the AttnGAN, which are particularly important for text-to-image generation for complicated scenarios.

We believe more opportunities in the future must be taken to enhance our research. Future studies may explore the impact of using another architecture of the generative adversarial network which is StyleGAN and compare it with the AttnGAN.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Qassim University, represented by the deanship of scientific research, on the financial support for this research under the number (2023-SDG-1-BSRC36794) during the academic year 1445 AH / 2023 AD.

REFERENCES

[1] I. Goodfellow *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139-144, 2020. doi: 10.1145/3422622.

[2] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F. Y. Wang, "Generative adversarial networks: introduction and outlook," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, p. 588–598, 2017. doi: 10.1109/JAS.2017.7510583.

[3] S. Tyagi and D. Yadav, "A comprehensive review on image synthesis with adversarial networks: Theory literature and applications," *Arch. Comput. Methods Eng.*, vol. 29, no. 5, pp. 2685-2705, Aug. 2022. doi: 10.1007/s11831-021-09672-w.

[4] R. Zhou, C. Jiang and Q. Xu, "A survey on generative adversarial network-based text-to-image synthesis," Neurocomputing, vol. 451, pp. 316-336, Sep. 2021. doi: 10.1016/j.neucom.2021.04.069.

[5] Y. X. Tan, C. P. Lee, M. Neo, K. M. Lim, J. Y. Lim and A. Alqahtani, "Recent advances in text-to-image synthesis: Approaches datasets and future research prospects," *IEEE Access*, vol. 11, pp. 88099-88115, 2023. doi: 10.1109/ACCESS.2023.3306422.

[6] H. Cao *et al.*, "A survey on generative diffusion model," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 2814-2830, 2024, doi: 10.1109/TKDE.2024.3361474.02646.

[7] C. Zhang, C. Zhang, M. Zhang, and I. So Kweon, "Text-to-image diffusion models in generative AI: A survey," *arXiv:2303.07909v2*, 2023. doi: 10.48550/arXiv.2303.07909.

[8] C. Schuhmann *et al.*, "LAION-5B: An open large-scale dataset for training next generation image-text models," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022. doi: 10.48550/arXiv.2210.08402.

[9] M. Tao, H. Tang, F. Wu, X. Jing, B.-K. Bao and C. Xu, "DF-GAN: A simple and effective baseline for text-to-image synthesis," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 16494-16504, Jun. 2022. doi: 10.1109/CVPR52688.2022.01602.

[10] S. Frolov, T. Hinz, F. Raue, J. Hees, and A. Dengel, "Adversarial text-to-image synthesis: A review," *Neural Networks*, vol. 144, p. 187–209, 2021. doi: 10.1016/j.neunet.2021.07.019.

[11] S. K. Alhabeeb and A. A. Al-Shargabi, "Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction," *IEEE Access*, vol. 12, pp. 24412-24427, 2024, doi: 10.1109/ACCESS.2024.3365043.

[12] J. Agnese, J. Herrera, H. Tao and X. Zhu, "A survey and taxonomy of adversarial neural networks for text-to-image synthesis", *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 4, Jul. 2020. doi: 10.1002/widm.1345

[13] R. Yanagi, R. Togo, T. Ogawa, and M. Haseyama, "Query is GAN: scene retrieval with attentional text-to-image generative adversarial network," *IEEE Access*, vol. 7, 153183-153193, 2019. doi: 10.1109/ACCESS.2019.2947409.

[14] H. Zhang, S. Jiang, and Y. Fu, "Stylized Text-to-Fashion Image Generation," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 2021, pp. 1–8, doi: 10.1109/FG52635.2021.9667042.

[15] T. Xu *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, pp. 1316–1324, doi: 10.1109/CVPR.2018.00143.

[16] L. Li, Y. Sun, F. Hu, T. Zhou, X. Xi, and J. Ren, "Text to realistic image generation with attentional concatenation generative adversarial networks," *Discrete Dynamics in Nature and Society*, 2020, doi: 10.1155/2020/6452536.

[17] B. Bordia, S. Patel, G. Supreeth, and B. R. Mohan, "Text to Image Generation using Hybrid Attention Generative Adversarial Network," *Journal of Critical Reviews*, vol. 7, no. 15, p. 6068-6075, 2020, doi: 10.31838/jcr.07.15.776.

[18] M. E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," *In 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, IEEE, 2008. doi: 10.1109/ICVGIP.2008.47

[19] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," *California Institute of Technology*, 2011. [Online]. Available: https://authors.library.caltech.edu/records/cyyh7-dkg06.

[20] T. Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740-755. doi: 10.1007/978-3-319-10602-1_48.

[21] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *In International conference on machine learning*, PMLR, 2016, June, pp. 1060-1069, doi: 10.48550/arXiv.1605.05396.

[22] X. Chen *et al.*, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015, doi: 10.48550/arXiv.1504.00325.

[23] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X, Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, p. 2234-2242, 2016, doi: 10.48550/arXiv.1606.03498.

[24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017, doi: 10.48550/arXiv.1706.08500.

[25] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, p. 197-209, 2008. doi: 10.1016/j.isprsjprs.2018.01.004.

[26] Y. Yu and F. Liu, "A two-stream deep fusion framework for high-resolution aerial scene classification," *Computational intelligence and neuroscience*, vol. 2018, 2018, doi: 10.1155/2018%2F8639367.

BIOGRAPHIES OF AUTHORS

Dina M. Ibrahim has been an associate professor at the Information Technology Department, College of Computer, Qassim University, KSA, since 2015. In addition, Dina works as an assistant professor at the Computers and Control Engineering Department, Faculty of Engineering, Tanta University, Egypt. She was born in the United Arab Emirates, and her B.Sc., M.Sc., and Ph.D. degrees were obtained from the Computers and Control Engineering Department, Faculty of Engineering, Tanta University, in 2002, 2008, and 2014, respectively. Dina worked as a consultant engineer, database administrator, and vice manager on the Management Information Systems (MIS) Project, at Tanta University, Egypt, from 2008 until 2014. Her research interests include networking, wireless communications, machine learning, security, and the Internet of Things. Dina has published about 70 articles in various refereed international journals and conferences. She has been serving as a reviewer in the Wireless Network (WINE) Journal since 2015. Dina has also served as a co-chair of the International Technical Committee for the Middle East Region of the ICCMIT conference since 2020. She can be contacted at email: d.hussein@qu.edu.sa and dina.mahmoud@f-eng.tanta.edu.eg.

Amal A. Al-Shargabi is an associate professor at the College of Computer, Qassim University. She received her Ph.D. degree in Information Technology and Quantitative Sciences from Universiti Teknologi MARA (UiTM), Malaysia. She earned her Master's degree, with an award of excellence, in Computer Science from the same university. Her research interests include empirical program comprehension, empirical software engineering, and machine learning. She has attended and participated in various international conferences such as the IEEE//ACM International Conference on Program Comprehension, the International Conference on Soft Computing in Data Science, and the IEEE Conference on e-Learning, e-Management, and e-Services. She has received several awards, including the Invention, Innovation and Design Exhibition (IIDX'16) Award, Malaysia, and the Three Minute Thesis (3MT'16) Award, Malaysia. She can be contacted at email: amal.alshargabi@dmu.ac.uk and a.alshargabi@qu.edu.sa.