Enhancing Text-to-Image Synthesis with an Improved Semi-Supervised Image Generation Model Incorporating N-Gram, Enhanced TF-IDF, and BOW Techniques

Vamsidhar Talasila[1]*, Murali Mohan V[2], Narasingarao M R[3]

Abstract: In order to create photographic images from semantic text descriptions with high resolution and visual integrity, text-to-image synthesis is used. But if the resolution of the images is increased, network complexity and processing needs become more difficult. To generate high-resolution images, existing models use huge parameter networks and computationally expensive methods, which leads to unstable training processes and expensive training. In this research, we propose a novel semi-supervised image generation model (SIGTI) that uses both labelled and unlabeled datasets for text-to-image conversion. The labelled dataset, which contains text and associated images, is used throughout the training phase. From the text data, we extract features including N-grams, better TF-IDFs, and BOWs. These features are then used to train the feature set of the NIC semi-supervised image creation model, which combines an upgraded GAN and Deep CNN. The unlabeled dataset with simply text is used during the testing phase. In order to create the appropriate relevant image, we extract the N-gram, enhanced TF-IDF, and BOW features from the text and compare them with the trained features using the proposed NIC model. To determine how well our suggested model works, we evaluate it thoroughly and compare its performance to other established methods.

1. Introduction

The development of graphic representations based on natural language descriptions has gained popularity recently. For artists and graphic designers, this development has been helpful in their creative endeavours. Computer vision (CV) and natural language processing (NLP) are both included in the field of text-to-image conversion [3] [5]. It has been increasingly popular to create visuals from textual descriptions, which helps graphic designers and artists with their creative processes. This method uses the opposite job from picture captioning, where the objective is to predict the next word based on the content of the image and the words that have come before [1] [8]. However, creating synthetic images necessitates the complex placement of multiple pixels. Deep generative models have made this effort much easier [2] [4], especially with the introduction of Generative Adversarial Networks (GANs). The difficulty of producing natural-looking visual data, particularly in high-dimensional environments, is efficiently addressed by the special neural

network topology of GANs, which enables the synthesis of images with no restrictions on dimensions. As a result, GANs are used in a variety of fields, such as computer-aided design and creative arts [10].

Despite the significant advancements made in this area, there are still issues with present techniques that make it difficult to combine text and images well. There are primarily two unresolved problems that require attention. First off, current methods frequently overlook interactions and thorough knowledge transmission between several stages [6] [12]. The resulting images have different degrees of coarseness and fineness. Second, the images created by these approaches have a coarse-to-fine order because they do not consistently create cross-sample links between images generated at various phases. The sharpness and presence of tiny local characteristics in the photographs are both impacted by this inconsistency. Conventional methods don't provide the stability needed to maintain cross-sample interactions, which can lead to uneven learning within networks [7] [14].

The limited variation observed in the generated images could potentially be attributed to the single-stage generator's focus on incorporating structured and high-dimensional textual background information, while neglecting the random latent coding responsible for introducing diversity [9]. The following list of important obstacles that text-to-image synthesis must overcome. Despite some successes,

Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India[1]*
Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India[2]
Professor, Department of Computer Science and Engineering, GITAM University, Visakhapatnam,
Andhra Pradesh, India[3]

current methods still find it difficult to generate images with high resolution and photorealistic features, which leads to simple generated images [11] [17]. The first difficulty is producing photorealistic images from text descriptions since the universe of likely images that correlate to a given text description is multimodal. Numerous graphic depictions can successfully convey the meaning of the given textual description. The second issue is brought on by the problems that existing language analysis techniques have in extracting important and implicit information from textual content, particularly suggested meaning [13] [15]. Finally, addressing several visual elements throughout several rounds is required to produce lifelike images, including blending, occlusion, lighting, shadow generation, and more [13] [15].

It can be difficult to train an autonomous text-to-image synthesis system [16] without explicitly labelling image-text combinations. Compared to supervised approaches that concentrate on improving visual-textual interaction models to raise the visual fidelity of generated images, unstructured text-to-image synthesis is still in its early stages, which adds additional complexity. [18]. This presents a number of challenging challenges to do. First, creating a trustworthy generative model that functions well without relying on manually labelled image-text pair data. The second step is making sure the images that are produced appropriately reflect the regional visual information that is stored in the input text. Assuring that the generated visuals are logically consistent with the supplied text and aesthetically pleasant [19] [20]. While several deep learning approaches are used to create graphics from text, it is still difficult to produce results that are realistic and semantically consistent. As a result, the SIGTI (Semantic Image Generation with Text Input) approach is suggested in this work for text-to-image synthesis.

The following is the main contribution of this work:

- To effectively reduce the dimensionality of the dataset, we suggest combining improved TF-IDF-based features with N-gram and BOW-based features that were derived from the text data.

- Within the NIC model, we suggest an improved GAN model that incorporates deep CNN and inverse binary cross entropy.

This essay is structured as follows:

A thorough summary of earlier research projects and their contributions to the field of text-to-image synthesis is given in Section 2. Section 3 presents the text-to-image synthesis paradigm's proposed architecture and essential elements. The superior outcomes that were obtained and the related experimental analyses are displayed in Section 4. Finally, Section 5 wraps up the investigation by summarising the results.

2. Literature Review

Zhang et al. [1] introduced a text-to-image synthesis method in their 2021 study that makes use of a multi-generator framework based on a Generative Adversarial Network (GAN). The technique involves combining the verbal description of the original image with limitations, such as a noise vector. To improve the probability distribution over the target distribution, several generators were used in accordance with the Deep Convolutional GAN (DCGAN) principles. During training, multiple synthetic training samples produced by various generators were analysed, and the generators were gradually stopped based on the discriminator's ability to recognise them. The suggested method successfully created images with distinguishing features while also assuring conformity to global restrictions. By utilising several generators and adjusting the discriminator's functionality correspondingly, the GAN demonstrated stability as the training process advanced in higher-dimensional areas.

AFang et al. [2] revealed a thorough text-to-image synthesis pipeline in 2020. In their suggested method, the input image was divided into foreground and background images. In order to detect significant foreground objects, the foreground images were segmented using the Microsoft COCO dataset. Throughout the image synthesis process, great attention was paid to maintaining the scales and placements of these objects. Markov Chain Monte Carlo (MCMC) techniques and an original cost function were also used to handle limited layout challenges. A Poisson-based method and a relighting model were used to ensure flexibility in how the foreground objects and the background image were handled. These methods made it possible to create aesthetically appealing and consistently contextualised images from written descriptions.

A text-to-image synthesis method utilising inter-stage cross-sample equality distillation was introduced in 2021 by Mao et al. [3]. The approach concentrated on improving cross-sample connections through the application of restrictions at various phases of the image production procedure. The proposed methodology successfully produced aesthetically beautiful images and showed increased quantitative performance when compared to existing methods using cross-sample similarity distillation at each level. The study advanced the field of text-to-image synthesis by using these methods.

Arabic text-to-image creation utilising AraBert via a DF-based GAN (Generative Adversarial Network) was proposed by Bahani et al. in 2022 [4]. The project includes the development of a fresh dataset designed exclusively for Arabic text-to-image synthesis, with text generation support provided by DeepL-translator. Utilising AraBert's skills, this method allowed for the strong embedding of the created words. In order to align the text with the suggested

shape of the DF-GAN, the dimension vector size was also decreased. The UP Blocks of the DF-GAN were then fed the shortened embedded texts and trained using two datasets. The study that was presented showed how the whole design worked well, producing high-resolution photos with better performance. Additionally, the discrepancy between the generated graphics and the original Arabic text was successfully minimised.

An effective single-stage architecture for distributed text-to-image synthesis was introduced in 2022 by Zhang et al. [5]. The proposed concept used a multi-level modular structure to solve a significant problem. In this method, two modules—the pixel attention module and the channel attention module—based on word stage attention and utilising DiverGAN's advantages were used. A dual residual structure was used to preserve different source features as they spread via deeper networks, ensuring the retention of a variety of properties. By connecting the completely linked layer within the pipeline, the diversity issue was resolved. Injecting variables into the linear stage also helped to increase quality and performance. The presented technique showed considerable improvements in producing a variety of excellent images from text descriptions.

An unsupervised method for text-to-image synthesis that did not rely on human-labeled data was presented by Dong et al. in 2021 [6]. The suggested approach used visual tricks to blend two different photos into one new image. The model was given pseudo image-text pairs by using a GAN-based approach. Using a discriminative loss function based on the previously discussed visual approaches, the generator and discriminator modules were trained. Then, using the input phrases as input, the pre-trained GAN model was used to create realistic visuals that displayed exceptional global semantic consistency. The suggested process effectively generated crucial and aesthetically pleasing graphics that matched the given text inputs.

A self-attention supported multi-level GAN architecture for text-to-image synthesis was introduced in 2021 by Dong et al. [7]. By including a multi-stage picture dependence and combining multi-level visual semantic information with the input language, this method improves image quality. A multi-level perceptual loss was created to enhance the visual semantics between the images and the text, improving the semantic alignment between actual and synthetic images. The proposed model also featured a mode seeking formulation for enhanced performance. Comparative experimental results showed that the suggested model performed better than existing methods.

Table I. Features and difficulties of current text-to-image synthesis approaches

Source [citation]	Model	Output	Problems
[6]	Visual concept Discrimination method	It produces images with precise visual concepts	Construction of a reliable generative model is still in its early stages
[5]	DrierGAN	In terms of quality and diversity, it does better	It is necessary to consider the creation of believable samples and determine the quantity of subjectivity that is available
[3]	Cross sample similarity distillation method	It increases the diversity and effectiveness of images taken	It does not provide trustworthy semantic images
[1]	DCGAN	One achieves higher resolution image synthesis	More discriminators need to be added for the model to become more stable
[4]	BERT	It has the power to synthesise the image with originality and catch important words	Consistent text descriptions are required to resolve the issue with the incomplete image
[8]	CDRGAN	It improves in terms of both the human rank score and the inception. Additionally, it produces visuals that are very similar to natural ones	Investigating an alternative method of producing graphics-based images is necessary

			The written description does not optimise the discriminator for the suggested approach
[7]	SAMGAN	It increases the variety of images that are produced	The written description does not optimise the discriminator for the suggested approach
[2]	Constrained MCMC	It excels at text-to-image matching to a greater extent	Calculating time complexity is necessary

Chained deep recurrent GANs were used in 2021 by Wang et al. [8] to offer a unique method for text-to-image synthesis. Their approach successfully combined data about local features and global image frames. The authors effectively spread the parameters and handled computational bottlenecks by using modelled chained deep recurrent processes. Additionally, the suggested approach handled the pixel-level logic linkages in the created graphics with intelligence. As a result, the provided strategy performed better than traditional methods using a variety of computational metrics, proving its superiority.

A. Problem Statement

The table above (Table I) gives a thorough analysis of the approaches currently used for text-to-image synthesis, highlighting their advantages and disadvantages. Below, a number of interesting techniques are covered. The DCGAN method was used by the author of paper [1] to create higher-resolution picture synthesis. However, the addition of more discriminators is required to guarantee model stability. The methodology built on Constrained MCMC [2] showed better text-to-image matching. It turned out to be difficult to estimate how complex it was in terms of time. The quality and diversity of the generated images were successfully improved by the Cross sample similarity distillation approach [3]. However, maintaining consistent semantic imagery continues to be difficult.

An Arabic text-to-image production method using AraBert via DF-based GAN was suggested in Paper [4]. With innovation, this technique synthesised visuals and caught important words. Addressing issues with partial images brought on by erratic text descriptions, however, is still a difficult undertaking. Diverse and high-quality images were produced more efficiently using the DiverGAN-based technology [5]. Making realistic samples is still a challenge to test, though. Using the Visual idea discrimination method, the author produced images that accurately represented visual concepts [6]. However, it is still difficult to train a trustworthy generative model using this method. The diversity of the created images improved as a result of the SAMGAN method [7] that was provided in the research.

Uncertainty persists on how to best optimise the suggested method's discriminator using the text descriptor. The CDRGAN-based algorithm [8] produced images that were nearly natural and improved in both inception and human rank ratings. Exploring different avenues for image production in computer graphics is still a challenging task.

3. Suggested Text-to-Image Synthesis Based on an Improved Gan Through an Improved Tf-Idf

A. Text-to-Image Synthesis Proposed

While the automatic creation of useful visuals from text has a lot of potential and appeal, existing AI systems have a long way to go before they can accomplish this. The creation of efficient and adaptable RNN architectures that can learn discriminative representations of textual features, however, is the result of recent developments. Additionally, deep convolutional GANs have produced highly captivating images of particular categories, such as interiors of homes, album covers, and portraits, with promising results. In this paper, we offer a SIGTI (Synthesis of Images from Text Input) framework that makes use of two datasets: an unlabeled dataset that only contains text and a labelled dataset that includes both text and images.

- Training Phase: Each image in the labelled dataset is linked with its appropriate text during the training phase. Various techniques, including N-gram, Bag-of-Words (BOW), and improved TF-IDF based features, are used to extract features from the text. These extracted characteristics are then utilised to train a deep convolutional captioning network (DCCN) and a generative adversarial network (GAN)-based semi-supervised picture creation model. The NIC model learns to match the features with the relevant images under the guidance of the extracted features, producing aesthetically coherent images from the input text.

- Testing Phase: In the testing phase, features from the provided text are extracted, including N-grams, an enhanced TF-IDF, and BOW. These extracted features are then added to the NIC model, which consists of a GAN and DCNN that have been improved. The retrieved features and trained features are contrasted within the NIC model. By utilising extracted features that share traits with the training features, the NIC model creates an image that is contextually relevant for the input text. Figure 1 shows a general representation of the suggested text-to-image synthesis that incorporates an upgraded GAN model and better TF-IDF.
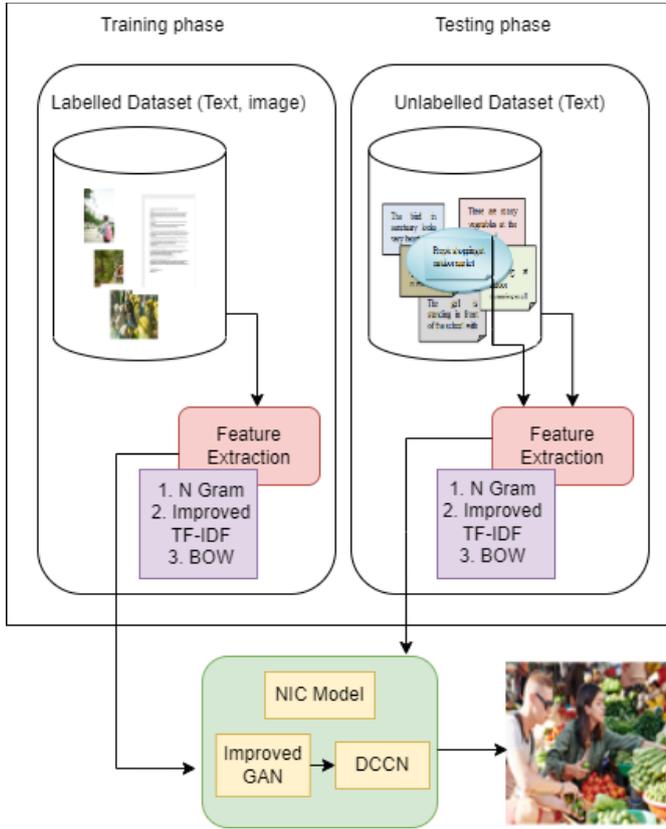
$$P(Z_1, Z_2, ...Z_n) = P(Z_1)P(Z_2 | Z_1).....P(Z_n | Z_1...Z_{n-1})$$
$$(1)$$

The size of the estimated formulation must be kept to a minimum because the word currently in use only includes a small number of words, as can be illustrated by Eq. (2). N in this case denotes a smaller value.

$$P(Z_i | Z_1...Z_{i-1}) = P(Z_i | Z_{i-N+1}...Z_{i-1}) \tag{2}$$

The likelihood of a word occurring in a word series based on unigrams can then be written as in Eq. (3).

$$P(Z_1, Z_2, ...Z_n) = \prod_{i=1}^{n} P(Z_i) \tag{3}$$

Additionally, Eq. (4) can be used to describe the bigram-based likelihood of a word occurring in a word series.

$$P(Z_1, Z_2, ...Z_n) = \prod_{i=1}^{n} P(Z_i | Z_{i-1}) \tag{4}$$

Following that, Eq. (5) can be used to indicate the trigram-based likelihood of a word occurring in a word series.

$$P(Z_1, Z_2, ...Z_n) = \prod_{i=1}^{n} P(Z_i | Z_{i-2}, Z_{i-1}) \tag{5}$$

The probability of maximum likelihood technique is used in the N-gram model to balance the probability value for each word. All values that can be expressed as in Equation (6) are continued with this guess. The full collection of word frequencies, the total number of words in the corpus, and the occurrences of the two words occurring together in the text corpus are all explicitly defined in this context.

$$P(Z_i) = \frac{O(Z_i)}{W}$$

$$P(Z_i | Z_{i-1}) = \frac{O(Z_{i-1}Z_i)}{O(Z_i)}$$

$$P(Z_i | Z_{i-N+1}, ...Z_{i-1}) = \frac{O(Z_{i-N+1}, ...Z_i)}{O(Z_{i-N+1}, ...Z_{i-1})} \tag{6}$$

- Enhanced TF-IDF: To assess the importance of terms within a document set, the TF-IDF approach [22] is used. It provides a thorough evaluation by combining word frequency and inverse document frequency.

- Term Frequency tracks how frequently a word appears in a corpus of text. The frequency data are initially scaled and normalised using the corpus's overall word set. The number of words and the size of the text corpus are taken into account when calculating term frequency. Equation (7) illustrates how term frequency can be stated generally across all text collections.

Fig. 1. The overall structure of the proposed SIGTI is better TF-IDF based features and an upgraded GAN model

B. Overview of N-gram, enhanced TF-IDF, and BOW-based features for feature extraction

- N-gram: In natural language processing, the N-gram technique [21] is frequently used to examine the relationships between nearby words and choose appropriate labels based on previously acquired data. In this study, text from the labelled dataset, D1, is processed to extract N-gram-based characteristics. For instance, N-gram analysis is used to identify the correct noun to complete the word sequence "a man sits under the," such as a chair, bus, house, etc. The N-gram technique also takes into account word relationships and context. In the series given, for instance, the words "tree" and "sun" may be determined to be pertinent and added after "a man sits under the." However, using the word "school" in this sentence would be inappropriate because it would result in an absurd sentence like "A man sits under the school." The N-gram technique chooses words with higher N-gram values to generate accurate sentences by using past knowledge and looking for relevant sentences in a corpus. The association between phrases and words is described by the numerical measure known as the N-gram. Assume that Z is a sentence made up of n word series, and define n as the number of word series. The probability of a word, w, appearing in the word series can be represented as P(Z) in Eq. (1) if w is a member of the word series. $P(Z) = P(Z_1, Z_2, ...Z_n)$

$$tf(t,d) = \frac{count\ of\ word\ in\ document}{No.of\ words\ in\ document} \qquad (7)$$

- Similar to term frequency, inverse document frequency tracks how frequently a word appears in the corpus. Inverse document frequency determines the document frequency whereas term frequency quantifies the frequency count. As illustrated in Equation (8), the formulation for inverse document frequency can be expressed.

$$df(t) = Occurrences\ of\ word\ in\ document \qquad (8)$$

In addition, inverse document frequency serves as document frequency's reciprocal, expressing a word's informational value. Equation (9) provides an expression for the inverse document frequency formula. N stands for the document size in this equation.

$$Idf(t) = \frac{N}{df(t)} \qquad (9)$$

The combined TF-IDF can then be written as Eq. (10). Here, Nt describes the word length in the document.

$$W(t,d) = tf(t,d) * \log\left(\frac{N}{N_t}\right) \qquad (10)$$

The correctness of the aforementioned equation, Eq. (10), is increased by imposing particular constraints, such as adjusted IDF and improved normalisation length. As seen in Eq. (11), the revised formulation can be expressed. Here, [missing text] stands in for [missing context], Pt stands for the word class weight coefficient, and N is the enhanced normalisation length

$$b_t = \frac{\sum_{i=1}^{C_t} k_{t_i}}{C_t}$$

$$W(t,d) = \sqrt{tf(t,d)} * ModifiedIDF(i) * INL * P_t * b_t \qquad (11)$$

It is possible to formulate the constrained modified IDF as in Eq. (12).

$$ModifiedIDF(i) = \log_{10}\left(\frac{No.of\ document\ in\ p+1}{Document\ frequency\ of\ term\ i}\right) \qquad (12)$$

In this study, the text documents' dimensionality is reduced using an upgraded normalisation technique. Eq. (13) illustrates the enhanced normalisation. The value of k in this equation is set to 0.5, and represents the length of the document.

$$INL = \log\left[\frac{1}{2}\left\{\tanh\left(k\left(\frac{L - mean(TF(i,j))}{std(TF(i,j))}\right)\right) + 1\right\}\right] \qquad (13)$$

BOW: In this work, we transform the text data using the bag-of-words (BOW) method [23]. Based on how frequently a word appears in the document, it is transformed into a BOW representation. For jobs involving document classification, this BOW technique is frequently used. Based on the presence or frequency of terms in the document, characteristics are extracted using this model. The feature matrix is then developed while taking into account. Here, denotes the number of sentences, and denotes the number of texts in the corpus by the letter n. In order to train the classifier, the retrieved features are then used in the NIC model. The entire collection of retrieved features includes the extracted BOW-based features, which are designated as BOWF.

$$Ext^{Fset} = \left[P\left(E_i \mid E_{i-N+1},...,E_{i-1}\right), W(t,d), Bow^F\right]$$

C. *Proposed a NIC model for semi-supervised image creation*

A semi-supervised image creation model called NIC (Neural Image Captioning) uses the feature set collected from the provided text as its input. The NIC model includes a Deep CNN (Convolutional Neural Network) and an improved GAN (Generative Adversarial Network). In the augmented GAN model, inverse binary cross entropy is used to improve the loss function that is used during training. The NIC model thus outputs acceptable and believable visuals. Figure 2 shows the text-to-image production process based on the NIC model, as well as the improved GAN model and Deep CNN.
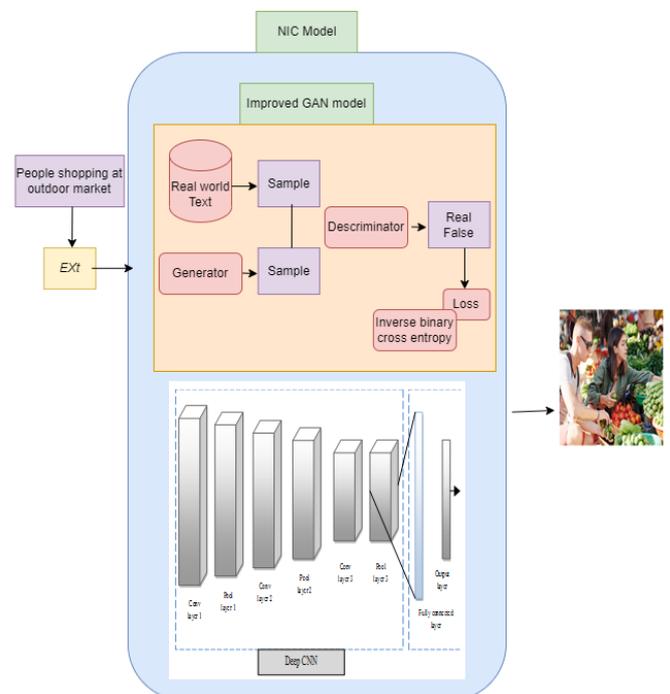

Fig. 2. Text-to-image generation using the Deep CNN and upgraded GAN models, based on the NIC model

- Enhanced GAN: The GAN [24] model makes use of two networks: the discriminator and the generator.

These networks function non-cooperatively, which means they compete with one another. All alterations are made within the parameter space of the parameterized neural networks that serve as the generator and discriminator. While the discriminator's primary goal is to categorise the samples, the generator seeks to produce samples that resemble those generated from the actual data distribution. The samples, which comprise both those produced by the generator and those drawn from the actual data distribution, are categorised by the discriminator. It gives the generator-generated and genuine data samples, respectively, lower and greater probability. Data and a noise vector derived from a Gaussian distribution with a zero mean and unit variance are used by both the generator and the real data distribution. The generator creates a suitable image by putting the retrieved features through the noise vector.

- Loss function: There are two loss functions in the fundamental GAN model. When doing binary classification, the discriminator seeks to reduce the negative likelihood, but the generator seeks to increase the possibility that the generated samples will be recognised as original. The discriminator parameter is defined as maximising the loss function between false and real samples when training the samples. The generator and discriminator are both trained during the training phase using a minmax loss function, which may be written as given in Equation (14). In this equation, G(Z) stands for the noise vector input for the generator, D stands for the discriminator network, G stands for the generator network, x stands for the real data retrieved from the real data distribution, and pdata stands for the real data distribution.

$$mim \max B(G,D) = \min \max E_{x-pdata}\left[\log D(x)\right] + E_{x-pdata}\left[\log\left(1 - D(G(Z))\right)\right]$$

(14)

Equation (15) illustrates how the standard loss function for the discriminator can be represented. G stands for the generator network in this equation, while Z stands for the noise vector input.

$$\nabla_{\theta_d} = \frac{1}{m}\left[\log G(x^{(i)}) + \log\left(1 - G(E(Z^{(i)}))\right)\right]$$

(15)

Inverse binary cross entropy, which is used to assess the accuracy of loss errors between actual and anticipated data, is added to the aforementioned Equation (15). Equation (16) expresses the increased GAN that results from this enhancement. In this equation, N stands for the total number of data points, and represents the actual value, and represents the anticipated value.

$$L(\hat{y},y) = \frac{1}{\left[-\frac{1}{N}\sum_{i}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]\right]}$$

(16)

The updated discriminator loss can thus be expressed as Eq. (17).

$$\nabla_{\theta_d} = \frac{1}{m}\sum_{i=1}^{m}\left[\log G(x^{(i)}) + \log\left(1 - G(E(Z^{(i)}))\right)\right] + L(\hat{y},y)$$

(17)

- Deep CNN: A deep CNN [25] is used in this study. When compared to conventional fully connected neural networks, the deep CNN is renowned for its capacity to learn a smaller number of parameters. This is accomplished via local receptive regions, distributed weights, and spatial sampling. Convolutional layers are specialised layers that make up the CNN; each layer is made up of tiny kernels that extract high-level characteristics. The fully connected layer receives its input from the last convolutional layer. An input layer, several convolutional layers, subsampling layers, and non-linear layers make up the overall structure.

- Layer Input: The features are received by the network's input layer as a multidimensional array of features. Its main job is to transfer the information to the variable convolutional layer, which is the next layer.

- Variable Convolutional Layer: The convolutional layer, which pulls features from the layer before it, is a crucial part of the deep CNN. While succeeding convolutional layers extract higher-level characteristics, the first layer catches low-level features. The dimensions of the kernels are determined by the numerous sets of filters that make up each layer. Across many levels of the kernel map, the kernel size defines the receptive field of the filter. A single unit is processed at a time throughout this operation, which is carried out using a sliding method. The DCNN model relies heavily on padding since it enables input padding to be zero.

- Non-linear layer: To improve feature detection, the non-linear layer of the network combines activation functions that are linear and non-linear. Non-linear activation functions allow the network to model more complicated patterns even when linear activation functions alone may not be able to capture complex linkages. There are many different non-linear activation functions, but sigmoid () is frequently utilised since it is straightforward and computationally effective. The Rectified Linear Unit (ReLU), which enhances pertinent features by using a non-linear function, is another well-liked option. The

receptive fields of convolutional layers are preserved throughout the whole network thanks to this activation function, which is significant.

- The sub-sampling layer, sometimes referred to as the pooling layer, downscales the complexity of earlier kernel maps. The entire network becomes better capable of maintaining learned features as a result. While reducing the dimensionality of the kernel map, the sub-sampling layer makes sure that noticeable patterns are kept. The convolutional layer, pooling layer, and activation layer are the three parts that make up each layer in the DCCN architecture. A collection of q kernels of dimension are convolved with the input feature map's dimensions. Convolution results in one kernel map for each kernel. The kernel moves to the right as it moves downhill. The input feature map, kernel, and convolution output interact in each convolutional layer. The identical filter is connected to each unit in the kernel map. In Figure 3, the interactions between the input feature map, kernel, and convolutional layer results are shown.

- The fully connected layer, which incorporates multiple layers of generic networks, is the top layer of the DCCN topology. All the characteristics and activations of earlier levels were combined in this layer. Additionally, it extracts the features and makes use of a variety of loss functions to train the classifier. As a result, the suggested method...... produces visuals that are pertinent to the content.
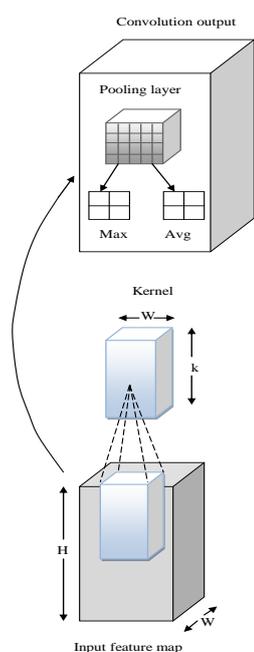

Fig. 3. Kernel, convolution output, and input feature map performance

4. Result and Discussion

A. Methodology for simulation

The text-to-image synthesis paradigm proposed in this study was implemented using MATLAB. Two different datasets, namely the CUB dataset [26] and the Flower Dataset [27], were utilized for experimentation. To evaluate the performance of the suggested SIGTI model, several standard techniques such as, Cycle GAN, GAN + DC-WO [29], DCGAN, SGAN, GAN + CMFA [28], SI-SSD [30], SRGAN and CGAN were employed for comparison. The fidelity and quality of the generated images were assessed using evaluation metrics such as IS and FID.

B. Synthesis of test image analysis for the CUB and Flower datasets

Based on the text descriptions provided, we compare the images produced using the upgraded GAN model with a number of standard approaches. Figures 4 and 5 respectively exhibit the graphic results displaying the images produced from the CUB and Flower datasets. The updated GAN model performs better, as can be shown by comparing the generated pictures to their associated text descriptions. When compared to the traditional approaches, the upgraded GAN generates images that are more realistic and organic. The improved feature extraction method based on TF-IDF is responsible for this improvement. The upgraded GAN produces images with better results and a higher degree of fidelity and realism.

C. Using the CUB and Flower dataset, ISanalysis on SIGTI and the traditional methods

The Inception Score (IS) is a commonly used metric for evaluating the quality of images generated by generative image models, including generative adversarial networks (GANs). In this study, traditional approaches were employed to compare the SIGTI model using IS analysis. The evaluation was conducted with the help of the flower and CUB datasets, and the outcomes are shown in Tables II and III.

Higher IS values, which indicate better ability in producing high-quality images, are anticipated from a dependable system. The SIGTI model demonstrated the greatest IS values and outperformed the conventional systems in both the floral and CUB datasets. The SIGTI model also obtained the highest IS value (5.3826 1.2692) for the CUB dataset when compared to other approaches.

| Text Description | "Bird with gradient brown patterns and a short black beak that is almost entirely black.

This bird is smooth, black, and has grey eyes and a pointed bill.

A medium-sized bird with dark black primaries and a short, pointed beak that is light black overall

This medium-sized bird has bright grey eyes and a sharp, short beak in addition to being mostly all black.

This little bird has some brown feathers on its belly and a black body, wings, and tail.

A little black bird with a short tail and a razor-sharp beak.

The beak of this little bird is fairly large and round, and it is dark brown in colour.

This bird's wings are black, and its eye is white. This bird's breast and belly are dark grey, while its bill and wings are black.

This black bird with a white eyering has a round body, a dark brown belly, and a pointed black bill. | This bird's body is black with little white patches on it, and its head is brilliantly golden with a black eyering.

The bill of this bird is bent, it has brown primaries, and its crown is yellow.

This bird has a yellow head with black eyes, a short black pointed beak, a black body and tail.

This bird has a yellow head and white patches on its wings. This bird has a grey body, a head that is yellow, and black wings.

A bird that is all black with a yellow head and beak.

This bird's head is yellow, while its wings are black.

With the exception of the bright yellow crown, throat, and breast, this huge bird's body is mostly dark grey and white and has long black tarsus and feet..

This specific bird has a yellow head and a black belly. | A little orange beak and a white patch are on the black bird's body.

This bird is black with an orange beak and a white nape.

With the exception of his white nape and white streaks on its covert, the bird is mostly black in colour.

A white speck may be seen on the back of this little, black bird's head. Sharp claws and a relatively large, multicoloured beak are features of the black bird.

A bird of black colour with white wing bars and a small, pointed bill.

The beak of this black bird is large, short, and pointed, and its nape is white.

Black wings and a large bill are features of this bird.

This bird's wings are black, and its nape is white.

This bird's black | The tiny bird has a white short beak, a bright blue body, and grey tail tips.

With the exception of its white bill, white and black primaries, and grey rectrices, this bird is almost entirely a bright blue colour.

A little blue bird with short grey bill, grey rectrices, and black and white wings.

The short, pointed beak of this bird is white and black with a white and vivid blue main colour.

A little bird that is extremely blue with black talons, a short, slightly curved white beak, and grey feathers on the end of its tail and wings.

The large, pointed beak of this blue and black bird is distinctive.

This bird has a very short beak and is blue with black feathers.

This bird has a long, pointed beak and is green and black in colour. A tiny blue bird with | This bird has a brown belly, a blue tail, and a blue head.

It has a blue head, a tummy that is brown and white, and multicoloured wings. This bird is smaller and has many various colours.

This medium-sized multicoloured bird features a multicoloured tail, an orange breast, and black eyes.

This tiny bird has a brilliant blue head and an orange breast.

This bird has a blue crown, brown breast, white wing bars, and a white belly.

A little bird with a blue crest and an orange breast.

This blue bird has a small beak in comparison to the rest of its body, an orange belly, and colourful, striped wings.

A tiny bird with a small black beak and a blue head. black legs, a lengthy tail with many colours. |
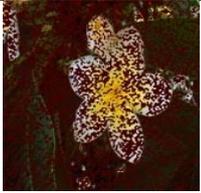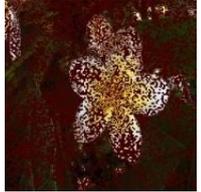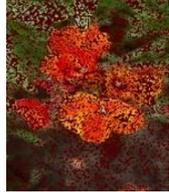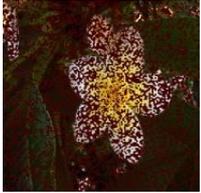|---|---|---|---|---|---|

| | | This bird's head is yellow, while its wings are black. | wings and white crest contrast with its plumage. | a black tarsus and a pointed white and black bill. | This bird's blue head, brown breast, white abdomen, and blue tail are all distinctive features.

The bird possesses an exquisite blue head and back, complemented by a brown breast. |
|---|---|---|---|---|---|
| Original Image | | | | | |
| CGAN | | | | | |
| Cycle GAN | | | | | |
| DCGAN | | | | | |
| GAN + CMFA | | | | | |
| GAN+DC-WO | | | | | |

	(a)	(b)	(c)	(d)	(e)
SI-SSD					
SGAN					
SRGAN					
Improved GAN					

Fig. 4. The task was completed by combining the upgraded GAN model with traditional methods, both the classic methods and the enhanced GAN used the CUB dataset to produce images.

| Text description | "Petal tips on this flower's yellow petals are white. Five white petals are arranged in a straight row on this flower, and at the end of the row are five petals that are yellow with white fringes.

This flower has oval-shaped petals and is | The flower features red, yellow, or orange petals.

The vivid colour has a nice appearance and makes me think of a sunny summer day.

This flower has little veins in its petals and is orange and yellow | The petals of this flower are orange, and there are a few white anthers in the centre.

This flower has orange petals with ruffled edges, and it is in bloom right now. | The primary characteristics of this flower are its brilliant pink petals and green pedicel.

The petals of this pink flower are arranged together before wrapping around the centre. | The pink flower's petals are arranged in tight layers in the middle before loosening up gradually as they move outward.

The pink, overlapping petals that make up the flower are what give |
|---|---|---|---|---|---|

coloured white and yellow. The flower in the illustration has white petals with a golden centre. Five smooth, white petals with yellow centres and rounded tips make up this flower. Five circular petals with white edges and yellowy orange centres make up this flower. A medium-sized flower with five white petals that start to turn yellow towards the centre. Smooth, white petals with orange centres make up this flower. The five oval-shaped petals of the flower are bright yellow in the centre and turn white at the edges.	in colour. Orange-tinged yellow petals that are grouped together and have white anthers. This flower features yellow petals with numerous white stamens. It also has an orange blossom with a structure that resembles a berry in the centre. Round petals and white stamen can be found on this yellow and red flower. This flower has four rounded, smooth, and yellow petals. Yellow and orange flowers in abundance, together with a grey stigma.	This flower features orange petals with white stamens The flower contains several of the elongated, orange petals with what appear to be scarlet veins. This flower's orange pedals have a lengthy stigma. The petals of this particular bloom are orange and then yellow. A flower with orange petals that are short and rounded. The petals of this flower are circular and are orange in colour. The stamen of this flower are short and white.	The layers of this flower's brilliant pink petals are closed together and have a pink colour. The petals of this crimson flower are grouped closely towards the centre and overlap one another. It is supported by a green pedicel. The petals of this red flower encircle the centre, giving it a red appearance. This flower has red petals that are grouped together. The petals of this flower have what appear to be jagged edges and have bright red blossoms that twist around one another.	it its pink colour. This flower contains clusters of curled, vivid pink petals that surround the receptacle on a sturdy, thorny stalk. Having long, curled petals and a green pedicel, this pink bloom is also a flower. The petals of this pink flower are tightly curled around the centre and have a pink colour. This flower's layers of pink petals are joined by a green pedicel. Pink petals are grouped together on this flower's petals.
Original Image				
CGAN				

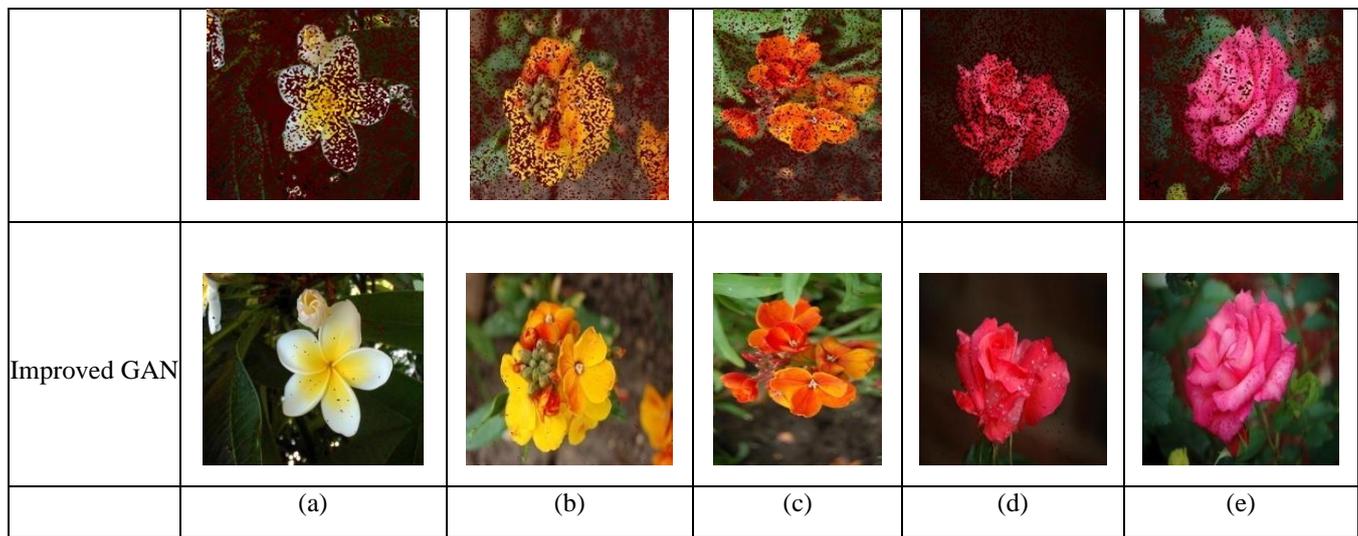|					
Cycle GAN					
DCGAN					
GAN + CMFA					
GAN+DC-WO					
SI-SSD					
SGAN					
SRGAN					
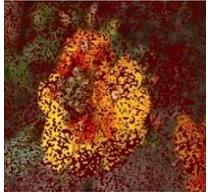
Fig. 5. With the use of the Flower dataset and both traditional and advanced GAN algorithms, text to picture synthesis was completed.

	a	b	c	d	e
Improved GAN					
	(a)	(b)	(c)	(d)	(e)

TABLE II. USE OF THE CUB DATASET TO ANALYSE THE IS ON SIGTI AND THE CONVENTIONAL TECHNIQUES

	IS
CGAN	4.6041 ±1.4481
GAN+DC-WO	4.8779 ±1.0207
DCGAN	4.3191 ±1.1733
SIGTI	5.5576 ±1.3953
Cycle GAN	2.0037 ±0.95308
SI-SSD	5.293 ±1.3288
SGAN	2.0899 ±0.99408
GAN + CMFA	4.5359 ±1.0998
SRGAN	3.8863 ±1.2334

TABLE III. SIGTI AND CONVENTIONAL TECHNIQUES IS ANALYSIS USING DATASET CUB

	IS
GAN+DC-WO	4.8801 ±1.0552
CGAN	3.8841 ±1.0521
DCGAN	2.9339 ±1.0261
SI-SSD	5.1263 ±1.2087
SRGAN	4.3006 ±1.1025
SGAN	3.9883 ±1.1482

SIGTI	5.3826 ±1.2692
Cycle GAN	4.075 ±1.1732
GAN + CMFA	4.6039 ±0.016701

D. Using the CUB Dataset, an analysis of the IS on SIGTI and the traditional methods

The FID is a commonly used metric in generative image models to measure the dissimilarity between feature vectors of real and generated images. In this study, FID evaluations of the SIGTI model were conducted using the flower and CUB datasets for comparison. It is desirable to achieve a lower FID value for better performance. Notably, the SIGTI model outperformed the other techniques, as evidenced by its FID value of 14.209 for the flower dataset. Furthermore, Table V clearly demonstrates that the SIGTI model achieved the lowest FID value among the compared methods. These results highlight the superior performance of the SIGTI model compared to traditional approaches due to its modifications.

TABLE IV. USING SIGTI'S FID ANALYSIS AND THE MORE CONVENTIONAL TECHNIQUES WITH DATASET FLOWER

	FID
DCGAN	23.568
SRGAN	25.637
GAN+DC-WO	19.418
SIGTI	14.209
GAN + CMFA	21.756
SI-SSD	16.717
SGAN	29.651

CGAN	23.057
Cycle GAN	22.837

TABLE V. ANALYSIS OF FID USING SIGTI AND CONVENTIONAL TECHNIQUES USING CUB DATASET

	FID
DCGAN	28.834
GAN+DC-WO	14.564
Cycle GAN	24.705
SRGAN	22.938
SI-SSD	13.924
SGAN	25.509
GAN + CMFA	16.428
SIGTI	11.836
CGAN	24.942

E. An ablation study was conducted on the SIGTI, TF-IDF, GAN model, with IS and FID metrics using datasets of Flower and CUB.

The performance of the SIGTI model was assessed and compared against a model employing conventional TF-IDF and a model utilizing conventional GAN. This evaluation was conducted using the Flower and CUB datasets to measure their respective performances. Table VI and Table VII present the results of the performance evaluation, including IS and FID analyses. It is clear that SIGTI has produced better results thanks to enhanced TF-IDF and improved GAN. Additionally, SIGTI scored 11.836 when evaluating the CUB dataset in terms of FID, compared to the models using conventional TF-IDF and traditional GAN, which both scored 24.23 and 24.11 respectively. These results confirm that SIGTI outperforms the traditional models in terms of FID as well, thanks to enhanced TF-IDF and improved GAN.Table VI shows that SIGTI obtained an IS score of 5.5576 1.3953 for the flower dataset, compared to scores of 4.7001 1.4783 for the traditional TF-IDF model and 3.9681 1.2593 for the conventional GAN model. These outcomes unequivocally show the benefit of utilising SIGTI with enhanced TF-IDF.

TABLE VI. PERFROMANCE EVALUATION OF SIGTI, TF-IDF, AND GAN MODEL THROUGH LOWER AND CUB DATASET WITH IS METRIC

Flower Dataset	IS
Conventional TF-IDF model	4.7001 ±1.4783
SIGTI	5.5576 ±1.3953
Conventional GAN model	3.9681 ±1.2593

CUB Dataset	IS
Model with conventional GAN	4.4379 ±1.1377
Conventional TF-IDF model	4.093 ±1.1087
SIGTI	5.3826 ±1.2692

TABLE VII PERFROMANCE EVALUATION OF SIGTI, TF-IDF, AND GAN MODEL THROUGH LOWER AND CUB DATASET WITH FID METRIC

Dataset (Flower)	FID
Conventional TF-IDF model	22.398
Conventional GAN model	22.288
SIGTI	14.209
Dataset (CUB)	**FID**
Model with conventional GAN	24.111
Model with conventional TF-IDF	24.23
SIGTI	11.836

5. Conclusion

This research introduces an innovative framework for text-to-image synthesis that leverages both labeled and unlabeled datasets. The unlabeled dataset contains text information exclusively, while the labeled dataset comprises pairs of text and their corresponding images. During the training phase, the labeled dataset is employed to establish connections between the provided text descriptions and their associated images. Various features, including N-grams, enhanced TF-IDF, and BOW-based features, are extracted from the text data. These extracted features are utilized to train the NIC model, a semi-supervised image generation model that combines GAN and DCNN components. The NIC model learns to match the extracted features with the given text descriptions and generate suitable images accordingly.

During the testing phase, the unlabeled text-based dataset is utilized. N-grams, enhanced TF-IDF, and BOW-based features are extracted from the text data in this dataset. These extracted features are then inputted into the NIC model, which consists of enhanced GAN and DCNN components. Inside the NIC model, a comparison is made between the extracted features and the trained features. If the extracted features demonstrate similar characteristics to the trained features, the NIC model produces a pertinent image that aligns with the given text description.

References

[1] Zhang, M., Li, C. & Zhou, Z, "Text to image synthesis using multi-generator text conditioned

generative adversarial networks", Multimed Tools Appl., vol. 80, pp. 7789–7803, 2021.

[2] https://doi.org/10.1007/s11042-020-09965-5

[3] Fang, F., Luo, F., Zhang, HP., Hua-Jian Zhou, Alix L. H. Chow & Chun-Xia Xiao,"A Comprehensive Pipeline for Complex Text-to-Image Synthesis", J. Comput. Sci. Technol., vol. 35, pp. 522–537, 2020. https://doi.org/10.1007/s11390-020-0305-9

[4] Mao, F., Ma, B., Chang, H., Shiguang Shan & Xilin Chen, "Learning efficient text-to-image synthesis via interstage cross-sample similarity distillation", Sci. China Inf. Sci., vol. 64, Article num. 120102, 2021. https://doi.org/10.1007/s11432-020-2900-x

[5] Mourad Bahani, Aziza El Ouaazizi,and Khalil Maalmi, "AraBERT and DF-GAN fusion for Arabic text-to-image generation", Array, vol.16, 2022.

[6] https://doi.org/10.1016/j.array.2022.100260

[7] Zhenxing Zhang , and Lambert Schomaker, "DiverGAN: An Efficient and Effective Single-Stage Framework for Diverse Text-to-Image Generation", Neurocomputing, vol. 473, pp. 182-198, 2022. https://doi.org/10.1016/j.neucom.2021.12.005

[8] Yanlong Dong, Ying Zhang, Lin Ma, Zhi Wang, and Jiebo Luo, "Unsupervised text-to-image synthesis", Pattern Recognition, vol.110, 2021.

[9] https://doi.org/10.1016/j.patcog.2020.107573

[10] Dunlu Peng, Wuchen Yang, Cong Liu, and Shuairui Lu, "SAM-GAN: Self-Attention supporting Multi-stage Generative Adversarial Networks for text-to-image synthesis", Neural Networks, vol. 138, pp. 57-67, 2021.

[11] https://doi.org/10.1016/j.neunet.2021.01.023

[12] Min Wang, Congyan Lang, Songhe Feng, Tao Wang, Yi Jin, and Yidong Li, "Text to photo-realistic image synthesis via chained deep recurrent generative adversarial network", Journal of Visual Communication and Image Representation, Vol. 74, 2021, 102955.

[13] https://doi.org/10.1016/j.jvcir.2020.102955

[14] Lianli Gao, Daiyuan Chen, Zhou Zhao, Jie Shao, and Heng Tao Shen, "Lightweight dynamic conditional GAN with pyramid attention for text-to-image synthesis", Pattern Recognition, vol.110, 2021. https://doi.org/10.1016/j.patcog.2020.107384

[15] S. Naveen, M. S. S Ram Kiran, M. Indupriya, T.V. Manikanta, and P.V. Sudeep, "Transformer models for enhancing AttnGAN based text to image generation", Image and Vision Computing, vol.115, 2021. https://doi.org/10.1016/j.imavis.2021.104284

[16] Ailin Li, Lei Zhao, Zhiwen Zuo, Zhizhong Wang, Haibo Chen, Dongming Lu, and Wei Xing, "Diversified text-to-image generation via deep mutual information estimation", Computer Vision and Image Understanding, vol.211, 2021.

[17] https://doi.org/10.1016/j.cviu.2021.103259

[18] Sharad Pande, Srishti Chouhan, Ritesh Sonavane, Rahee Walambe, George Ghinea, and Ketan Kotecha, "Development and deployment of a generative model-based framework for text to photorealistic image generation", Neurocomputing, vol.463, 2021.

[19] https://doi.org/10.1016/j.neucom.2021.08.055

[20] Zhongjian Qi, Jun Sun, Jinzhao Qian, Jiajia Xu, and Shu Zhan, "PCCM-GAN: Photographic Text-to-Image Generation with Pyramid Contrastive Consistency Model", Neurocomputing, vol. 449, 2021. https://doi.org/10.1016/j.neucom.2021.03.059

[21] Qingrong Cheng, and Xiaodong Gu, "Cross-modal Feature Alignment based Hybrid Attentional Generative Adversarial Networks for text-to-image synthesis", Digital Signal Processing, vol. 107, 2020. https://doi.org/10.1016/j.dsp.2020.102866

[22] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga and M. Bennamoun, "Text to Image Synthesis for Improved Image Captioning," in IEEE Access, vol. 9, pp. 64918-64928, 2021,

[23] doi: 10.1109/ACCESS.2021.3075579.

[24] Xue, Y., Guo, YC., Zhang, H., Tao Xu, Song-Hai Zhang & Xiaolei Huang, "Deep image synthesis from intuitive user input: A review and perspectives", Comp. Visual Media, vol. 8, pp. 3–31, 2022. https://doi.org/10.1007/s41095-021-0234-8

[25] Zakraoui, J., Saleh, M., Al-Maadeed, S., & Jihad Mohammed Jaam, "Improving text-to-image generation with object layout guidance", Multimed Tools Appl., vol. 80, pp. 27423–27443, 2021. https://doi.org/10.1007/s11042-021-11038-0

[26] Goar, D. V. . (2021). Biometric Image Analysis in Enhancing Security Based on Cloud IOT Module in Classification Using Deep Learning- Techniques. Research Journal of Computer Systems and Engineering, 2(1), 01:05. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/9

[27] J. Ni, S. Zhang, Z. Zhou, J. Hou and F. Gao, "Instance Mask Embedding and Attribute-Adaptive Generative Adversarial Network for Text-to-Image Synthesis," in IEEE Access, vol. 8, pp. 37697-37711, 2020,

[28] doi: 10.1109/ACCESS.2020.2975841.

[29] M. Yuan and Y. Peng, "Bridge-GAN: Interpretable Representation Learning for Text-to-Image Synthesis," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 11, pp. 4258-4268, Nov. 2020,

[30] doi: 10.1109/TCSVT.2019.2953753.

[31] M. Yuan and Y. Peng, "CKD: Cross-Task Knowledge Distillation for Text-to-Image Synthesis," in IEEE Transactions on Multimedia, vol. 22, no. 8, pp. 1955-1968, Aug. 2020,

[32] doi: 10.1109/TMM.2019.2951463.

[33] Cheng Cheng, Chunping Li, Youfang Han, and Yan Zhu,"A semi-supervised deep learning image caption model based on Pseudo Label and N-gram", International Journal of Approximate Reasoning, vol. 131, 2021. https://doi.org/10.1016/j.ijar.2020.12.016

[34] R. K. Roul, J. K. Sahoo and K. Arora, "Modified TF-IDF Term Weighting Strategies for Text Categorization," 2017 14th IEEE India Council International Conference (INDICON), Roorkee, India, 2017, pp. 1-6, doi: 10.1109/INDICON.2017.8487593.

[35] Ashok Kumar, L. ., Jebarani, M. R. E. ., & Gokula Krishnan, V. . (2023). Optimized Deep Belief Neural Network for Semantic Change Detection in Multi-Temporal Image. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2), 86–93. https://doi.org/10.17762/ijritcc.v11i2.6132

[36] Soumya S., and Pramod K.V., "Sentiment analysis of malayalam tweets using machine learning techniques", ICT Express, vol. 6, 2020. https://doi.org/10.1016/j.icte.2020.04.003

[37] Divya Saxena, and Jiannong Cao, "Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions", 2020.

[38] Ivars Namatevs, "Deep Convolutional Neural Networks: Structure, Feature Extraction and Training ", Information Technology and Management Science, vol. 20, pp. 40-47, 2017.

[39] doi: 10.1515/itms-2017-0007

[40] https://neptune.ai/blog/gan-loss-functions

[41] https://nextjournal.com/jbowles/n-gram-models-part 1#:~:text=An%20n %2Dgram%20model%20is,rich%20pattern%20disco very%20in%20text.&text=In%20other%20words%2 C%20it%20tries,or%20words%20near%20each%20o ther).

[42] Vamsidhar Talasila and M. R. Narasingarao, "BI-LSTM Based Encoding and GAN for Text-to-Image Synthesis", Sensing and Imaging, vol.23, 2022.

[43] Vamsidhar Talasila, Narasingarao M R , Murali Mohan V, "Optimized GAN for Text-to-Image Synthesis: Hybrid Whale Optimization Algorithm and Dragonfly Algorithm", Advances in Engineering Software, vol.173, 2022.

[44] Ghazaly, N. M. . (2020). Secure Internet of Things Environment Based Blockchain Analysis. Research Journal of Computer Systems and Engineering, 1(2), 26:30. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal /article/view/8

[45] Vamsidhar Talasila, M. R. Narasingarao, and V. Murali Mohan, "Modified GAN with Proposed Feature Set for Text-to-Image Synthesis", International Journal of Pattern Recognition and Artificial Intelligence, 2023.