iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 11 **Issue:** I **Month of publication:** January 2023

**DOI:** https://doi.org/10.22214/ijraset.2023.48867

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 11 Issue I Jan 2023- Available at www.ijraset.com*

# A Survey on Machine Learning Techniques for Video Caption Accessibility to Assist Children with Learning Disabilities

Anagha V[1], K.S. Kuppusamy[2]
[1, 2]*Department of Computer Science Pondicherry University Pondicherry, India*

***Abstract:*** ***On the World wide web, videos have become a speedy and effective information distribution method. Caretakers observed that YouTube was the most popular medium for kids during the COVID-19 epidemic, with more than 78% of kids watching it. To make a video accessible to children with learning impairments, captions are considered an assistive technology tool. Those with learning disabilities are not able to comprehend text content quickly within a video timeframe. We explored the temporal dimension of the caption of a frame in the video. The main difficulties with video caption accessibility are that users, within a short amount of time, have to get the gist of the visual content. In addition, they have to understand captions that are timed with frames. Children with learning disabilities have a hard time understanding captions, so this is one of the challenges they face. We presented a holistic view of caption accessibility and the challenges faced by children with learning disabilities in this paper.***

***Keywords: Learning disability, Caption Accessibility, Readability, Text to Image Synthesis Models***

## I. INTRODUCTION

The web is integrated with a lot of multimedia content. The most popular among video-based content. As the use of video-based learning continues to grow in popularity, we must focus on how we can help those people, including but not limited to users who are non-native speakers and students with disabilities. Basically, video captioning is the process of understanding and explaining a video. In order to create descriptions of the movie that are human-like, one must first comprehend the semantics of the film. Collaboration between the study of natural language processing and computer vision areas is necessary [5]. Consequently, subtitles are necessary to ensure that educational movies are accessible to everyone and appreciated by a range of students.The prior studies remarked that about 4 million children and teens suffer from learning difficulties, and many of them battlevarious learning problems. Roughly 10% of school-age children struggle with specific reading comprehension issues. We can able to understand that from their childhood itself because these learning-disabled children will show some intimations.Many children's reading difficulties are first noticed around the age of 7 or 8 if they exhibit any of the following symptoms:

*1)* Difficulties with basic reading skills, such as word recognition.
*2)* Difficulties understanding the key points in reading passages.
*3)* Frequent frustration with reading tasks.
*4)* Minor difficulties reading aloud but may read with minimal tone variation.
*5)* Difficulties remembering key details of what they've read.

These signs show that we can able to find those children with these kinds of learning disabilities in the early stages itself. When children are in primary school, their inability to read, write, or comprehend basic words or phrases is a warning sign thatthey require ongoing attention from parents and teachers. Some studies show that medical science also put forward a decision Support System (DSS) [3] identifying the early stages of dyslexia will aid in improving psychological assessment for detectinglearning disorders like dyslexia.Most researchers have attempted to investigate and discover strategies for enhancing the learning process with the use of multimedia accessibility by offering an accurate interpretation of text and captions in past and current studies .We conducted a literature review for video caption accessibility in this research, which was inspired by contemporary advancements in accessibility and educational technology, such as the popularity of YouTube platforms and virtual learning environments. Thestructure of this survey is as follows. Section 2 explains the Web Content Accessibility Guidelines, a detailed study of various captions that are supported for accessing the video content, and how those captions/texts ensure comprehension of the content.

©IJRASET: All Rights are Reserved 

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 11 Issue I Jan 2023- Available at www.ijraset.com*

Section 3 outlines the survey on General Adversarial Networks (GAN) for Text to Image Synthesis.

## II. BACKGROUND

### *A. Problem Statement*

As it can be challenging to visualize the text when reading it, this can be a problem. Additionally, there are words that can occasionally be misinterpreted. When text is presented in an image format, it is much easier to comprehend. Visual aids can deliver information more directly. Viewers are drawn in and engaged in their interests through visual content. Visual communication is used to some extent in presentations, learning, and other crucial activities.

### *B. Guidelines for Accessibility*

In order to enable accessibility, WCAG [31] specified a number of principles, guidelines, success criteria, benefits, and examples These recommendations cover a wide range of methods for making web material usable by people with disabilities as well as everyone else. Accessibility aims to offer a solution to the difficulties people may have in obtaining whatever details. For those with mobility limitations, giving them keyboard access would be an example. For those with vision problems, adding alternate text to the photographs would be another. The primary medium for communicating information online is video. Due to its attraction, users are drawn to watching movies. A strategic consulting and digital business consultancy estimate that each day, more than 500 million hours of videos are viewed on YouTube [30]. To ensure barrier-free access, accessibility elements must be offered to people who, because of their own limitations, lag behind video content. Making the video material accessible to everyone is the aim of several video accessibility regulations.

The W3C specifies that audio and video content that has been developed or published must be accessible at the time of publication. People with impairments may find that transcripts, captions, audio descriptions, and sign language interpretation suit their basic accessibility needs [32]. The WCAG's Guideline 1.2 covers time-based media accessibility guidelines [29]. Pre-recorded audio or video should be offered as an option. It is required that voiceovers and synced captions accompany the pre-recorded video. It is required that voiceovers and synced captions accompany the pre-recorded video. Synchronized captions to audio, video and multimedia content are suggested as a requirement by the US Government under Section 508 [27] to improve accessibility. Most guidelines suggested that people with hearing impairments and learning disabilities would interact with media content more effectively if closed captions were included in the audio information.

### *C. Captions*

Universal internet accessibility aims to make digital information available to everyone, regardless of their location, circumstance, or culture. Due to their impairment, people who have hearing loss are unable to understand auditory information. Early research indicates that captions can enhance EFL/ESL learners' vocabulary development, word recognition, and listening and reading comprehension [7]. In addition, Assistive technology may be used to get around the difficulty of obtaining digital content by providing subtitles for the video.

This development has allowed those who have hearing impairments to participate in online classes and movies. As a result, producing captions quickly will make them more difficult to access [11][28]. The readability of each caption frame depends on how frequently it is presented on the video. Thus, some user groups can access videos that do not adhere to the universal accessibility requirements since they are not affected by the limitations present in these movies.

#### *1) Closed captions*

People with disabilities frequently utilize closed-captioning technology, which has been around for years, to access video-based content. In American higher education environments, instructors frequently use videos. Closed captions may not be educationally advantageous for many pupils, though. When viewing television, for instance, closed captions can be "on" todisplay a visual text translation of what is being spoken on the screen at the same time.

Closed captions often show up as one to three lines of white text on a black backdrop at the bottom of a viewing screen [9]. The technology needs to ensure that all essential video-based information is represented in text for viewers with hearing difficulties and we can give a visual representation for those people with learning disabilities for better understanding. A video with closed captions is shown in Figure 1.

©IJRASET: All Rights are Reserved 

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 11 Issue I Jan 2023- Available at www.ijraset.com*

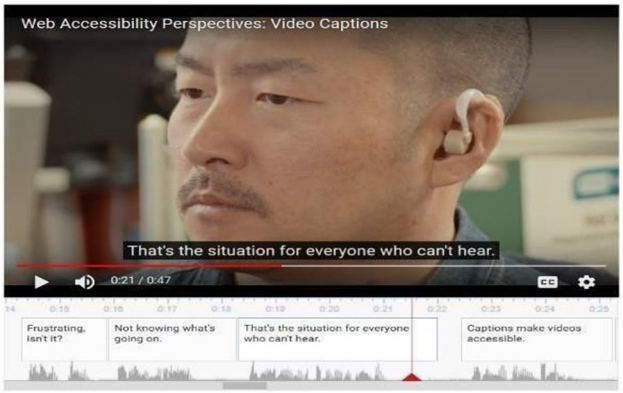



Fig. 1. Example of Closed Caption

*2) Fast caption*

With the aid of closed captions, users may watch videos more rapidly thanks to Fast Captions. A temporal summation of the captions will aid in making the rapid caption more understandable and speed up comprehension of the video's content. The author of the article [10] came up with a method for extracting a group of summarized phrases from closed captions using a technique called extractive summarization. Figures 2 and 3 illustrate our experiments with our fast-captioning technique usingNPTEL movies with caption files (Krishnan, 2009) for better understanding. Figure 2 depicts a video with captions that are playing at 1 speed; ordinarily, the captions are displayed at the bottom of the screen by all players. Figure 3 displays the movieat 6x speed along with a transcript of the subtitles. They asserted that this is a well-established technique for extracting phrases from videos and that the number of summary sentences relies on the level of the playback speed.

However, audiences may access videos more rapidly with the use of quick subtitles. This strategy is also acceptable for viewers with hearing difficulties because subtitles enable them to infer the video's content. Closed captioning for videos has become more popular as the importance of online accessibility and universal design principles has grown. The suggested Fast Captioning method can be used in the majority of these instances as more videos now provide closed captions.

We performed a study on how to represent the captions or text in an efficient way for people who face difficulties. Current research is focused on Automatic image synthesis from the corresponding text/captions by using various machine learning, such as supervised learning model (SL), General Adversarial Networks (GAN), and Deep Learning models for how to visualize or represent the text content which is present in a video, a detailed study was in table1. However, those models put forward a method such as text to image synthesis.

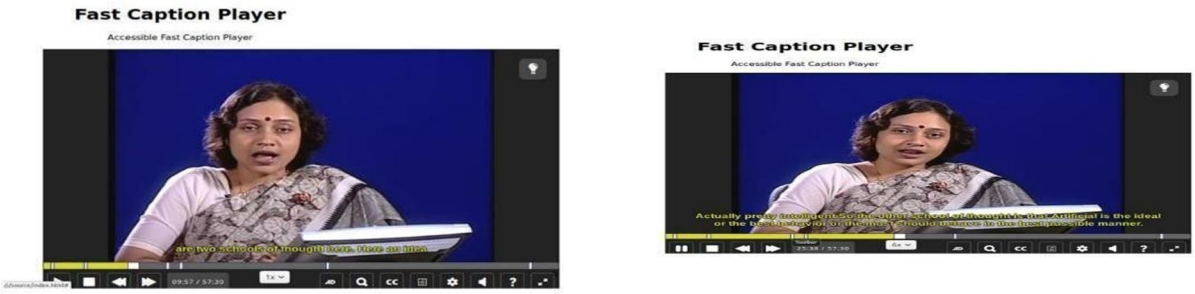



Fig. 2. Captions for video playing with speed

From the above studies, we explored the two types of caption representations such as closed captions and fast captions, wherewe come to the point that the majority of the videos that contain closed captions will make it possible to access the video content easily. And making those video captions/content more accessible for those users with various disabilities we perform another study on how to represent those contents in an efficient way in the following Section.

## III. SURVEY ON TEXT TO IMAGE SYNTHESIS METHODS

A picture is worth a thousand words! The written text offers efficient, effective, and brief forms of communication, whereas visual content, such as images, is a more thorough, accurate, and understandable mode of information sharing and understanding.

©IJRASET: All Rights are Reserved 

Nowadays, investigating text-to-picture synthesis is a stimulating and worthwhile endeavor. The goal is to generate images from words or captions, which entails using written descriptions as input to generate associated images [24]. Recent evidence suggests that the representation of text/ captions as an image will increase the understandability of word or content accessibility rather than reading those texts for those who suffer from various disabilities such as learning disabilities and audio impairments. However, when we explored on latest research the author [20] offered a method for automatically translating short, straightforward tales for young children written in current standard Arabic into the most appropriate visuals that effectively convey the stories' meanings. It is similar to imitating children's imaginative processes when they read stories, but it would be incredibly difficult for a machine to accomplish. We use a variety of ways to locate the images for problems involving simplification, and we dynamically link those images to relevant phrases. Recent research it's observed that nowadays, GAN models are widely used for better results.

*A. Traditional Learning-Based Text-To-Image Synthesis*

According to early research [34], a search and supervised learning combined method was predominantly used for text-to-image synthesis (Figure 4). The relationship between keywords (or key phrases) and images can be used to spot text units that are instructive and "picturable." Then, based on the text, these units will look for the most likely image portions, finally optimizing the picture arrangement based on both the text and the image parts. Several key AI components, including machine learning, computer vision, natural language processing, and graphics, were commonly used in these methodologies.

The major flaw in traditional learning-based text-to-picture synthesis techniques is that they can only alter the characteristics of existing or training images, not produce new image content. Recent years have seen great progress in the study of generative models, which can learn from training photos and generate new visual content. For instance, the modeling tool Attribute2Image models each image as a composite of the foreground and background [35]. In order to create visual material, a layered generative model with decoupled latent variables is also developed using a variational autoencoder. Because the learning is customized/conditioned by provided attributes, the Attribute2Image generative models may produce images with respect to numerous qualities, such as gender, hair color, age, and other factors, as shown in Figure 5.




Fig. 4 The correlation between keywords (or key phrases) and images wereused in early text-to-image synthesis studies.

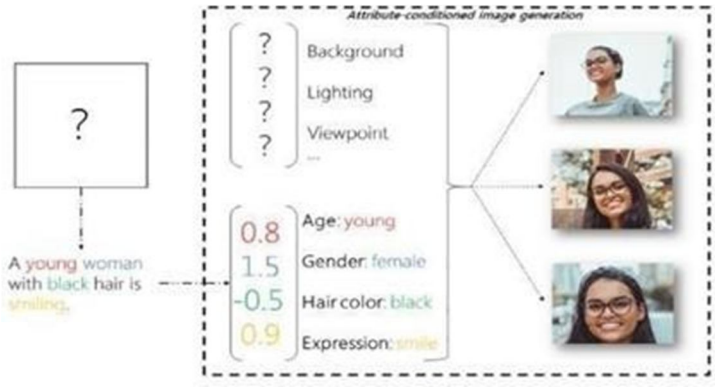



Fig.5. With reference to various features, supervised learning-based text-to-image synthesis

*B. Generative Adversarial Networks*

Generative Adversarial Networks (GAN) is a GANs are contemporary approach, that is applicable to both unsupervised and semi-supervised learning. They accomplish it by implicitly modeling data distributions with high dimensions Neural networksare used by Generative Adversarial Networks to create brand-new instances of data, like voice or images. Learning a generativemodel and training it using neural networks are referred to as "generative adversarial networks." The representations that GANscan learn can be used for a variety of applications, including classification, image super-resolution, style transfer, semantic image editing, and picture synthesis. GAN has two sub-parts (generator and discriminator):

1) *Generator:* A mapping from one representation space (latent space) to another (actual data) can be seen as the function of a generator network in picture synthesis [36]. Every image in the data space is distributed in some way in a very complex, high-dimensional feature space when it comes to image synthesis. GANs train a generator to create synthetic images from a much simpler feature space, known as the latent space, which is often random noise because sampling from such a complicated field is extremely difficult. It creates new data instances, the majority of which are bogus samples, and sends them to the discriminator in an effort to trick the discriminator.
2) *Discriminator:* Discriminator networks, which frequently consist of several convolutional and/or fully connected layers in deep neural networks, can be thought of as mappings from image data to the likelihood that the image comes from the true data space. The discriminator, however, employs down-sampling as opposed to up-sampling. It is trained using gradient descent, just like the generator, but its goal is to update the weights to improve the likelihood that it will correctly distinguish between real and falseimages.

GANs are newly proposed models and they are made up of two neural networks: Generator G, which creates new data instances using noise variables as input, and Discriminator D, which determines whether or not each instance of data is part ofthe training data set. A two-player minimax game with the following objective function is played by D and G.

LGAN = Epx[log[D(x)] + Epz[log[1 − D(G(z))] (1) where,

1) L is the loss function.
2) E is the empirical estimation.
3) x is the training data with the true data distribution.
4) z represents the noise variable sampled from the distribution.
5) G(z) represents the generated data instances.

Deep neural networks make up Generator and Discriminator. While the Discriminator's objective is to identify accurate data, the Generator's objective is to deceive the Discriminator. The discriminator is opposed to Generator, and they are both in conflict. The Generator makes every attempt to convince the discriminator that the fake instances it generates are real samples of data while simultaneously increasing the likelihood of mistakes while the discriminator is able to recognize the real ones. As a result, these stages are repeated several times, greatly improving the training of both sub-models.

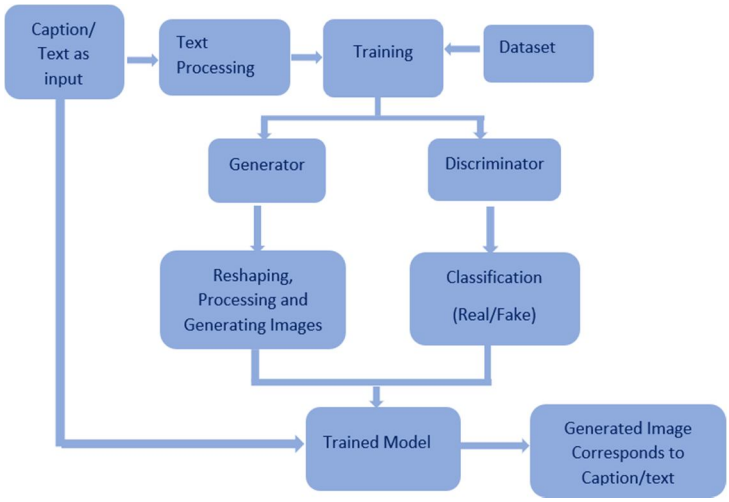



Fig.4. Proposed System for Caption/Text to Image generation using GAN

The proposed system is based on GAN for the generation of images from caption/text which is shown in Fig. 4. Where the text which is extracted from the video used for giving a visual representation of video content to enhance video accessibility for persons with learning disabilities and visual impairments. Initially, the caption/text of the video has been given as input to the system. Next, the input has been passed for tokenization where main textual features are extracted. Then Images from the dataset as well as the textual features along with noise are combined and further passed towards the training stage. The generator and discriminator here handle the additional processing, with the generator generating the images and the discriminator discriminating them. This process keeps going until the generator can produce images using the given data. Finally, a trained model that can produce images based on the provided textual input is obtained. So, from the above studies and literature review Table.1, we come to know that the GAN model can able to give better performance for reducing the comprehension of the video content by giving visual representation corresponding to the particular caption/text.

TABLE 1. LITERATURE REVIEW

| S. No | Authors | Year | Methodology | Inference | Metrics and Remarks |
|---|---|---|---|---|---|
| 1 | Rozita Ismail & Azizah Jaafar [1] | 2011 | Educational multimedia courseware | The author talked about how dyslexic children have a favorable attitude toward using computers and multimedia educational programs. | More than 60% multi- media encourages learning for dyslexic children |
| 2 | Luz Rello & Ricardo Baeza-Yates [2] | 2015 | Text customization, Readability | People with dyslexia may benefit from particular text presentation factors that make the text on a screen more readable, according to the author's analysis. | Font-size: 17inch screen, Char space: 14%, Type- face: arial give more ac- curacy |
| 3 | Nur Izzah Rahmiah Ab Llah & Kha- dijah Hanis Ahmad Firdaus [3] | 2018 | Business Model Canvas | This study into the need for decision support systems (DSS) to improve psychological testing for dyslexia diagnosis | Not much accurate to diagnose dyslexic children because its basic model |
| 4 | Juanita Rodriguez & Maria Victoria Diaz [4] | 2017 | Captions, Audio De- scription, Video Description | Author discussed the advantages of adopting accessible media to improve students' reading ability who have sensory issues and are from his background in panic. | CC& DV gives 60% ac- curacy to guide children classrooms |
| 5 | Bhavya Bhavya& Si Chen Zhilin Zhang [5] | 2022 | Caption transcription, Collaborative editing | Discussed how captions enable video-based learning by collaborative caption editing | LR method gives 93.8% |
| 6 | Haojin Yang & Christoph Meinel [6] | 2014 | Automatic video index- ing | The author described a method for indexing and retrieving content-based lecture videos in sizable lecture video archives. | Keyword & video task gives 99% accuracy |
| 7 | Habib Gowhary & Zeinab Pourhalashi [7] | 2014 | English proficiency test- TOEFL | Discussed how captions help to im- prove Iranian EFL learner's listening comprehension & affect learners' level of proficiency | Post text score of significant level is 0.018 |

| | | | | | |
|---|---|---|---|---|---|
| 8 | Mingxiao Li [8] | 2022 | CNN-LSTM | In order to improve the performance of an autoregressive teacher model with strong generalization performance, the author presented a knowledge distillation architecture employing captions. | BLEU Score is 60% after knowledge Distillation |
| 9 | Muralidhar Pantula & K.S. Kuppusamy [9] | 2019 | LSS Algorithm | Study on the readability of closed captions for those with limited reading skills, with the goal of improving comprehension by utilizing the context's lexical and semantic ambiguity | LSS algorithm gives92.6% accuracy on Readability |
| 10 | Muralidhar Pantula & K.S. Kuppusamy [10] | 2018 | Caption Summarization | Introducing a fast-Captioning method that condenses the closed captions based on word frequency so that those with hearing impairments can use it to deduce the content of the video from the captions. | 74% rating on readability using fast captioning |
| 11 | Muralidhar Pantula & K.S. Kuppusamy [11] | 2018 | TempRead readability Technique | Proposing a model to enhance the readability of caption based on time dimension. | Readability 80% in cation frame with respect to no of word and frame time |
| 12 | Muralidhar Pantula1 & K. S. Kuppusamy [12] | 2019 | AuDIVA | Introducing an AuDIVA tool will help people with visual impairments better understand the scene's context and be in a better position to understand the contents. | 92.6% accuracy on JavaScript files generated by AuDIVA |
| 13 | Junhao Liu & Min Yang [13] | 2020 | TGCR method | In order to improve the effectiveness of a "Two- Teacher One Student" (TTOS) knowledge distillation framework, the author makes use of image-to-text and text-to-image generating assignments. | The model gives 85% accuracy on Flicker 25 dataset |
| 14 | MuralidharPantula & K. S. Kuppusamy [14] | 2019 | CORDIF Model | Introducing the CORDIF model for text/caption simplification, which focuses on intra-word characteristics that divide the target word into classes for difficult and simple words. | 83.26% accuracy on ten-fold cross-validation |
| 15 | Harshit Parikh & Harsh Sawant [15] | 2020 | CNN, NLP | Discussed how an encoder processes the images before the Gated Recurrent Unit generates pertinent captions. | 73.8% accuracy on BLEU 1 score |
| 16 | Monika Singh & Amanpreet Kaur [16] | 2015 | DWT, Laplacian of Gaussian filter | Author proposed a hybrid robust method to efficiently extract text/caption localization from keyframes of the videos | 100% accuracy on cartoon video, 90% accuracy on educational video |
| 17 | Srinandan Komanduri & Y. Mohana Roopa [17] | 2019 | OCR (Optical Character Recognition) | This study examines how OCR works to find and detect text in electronic documents and can quickly transform that material into human-readable language. | Using a database to recognize characters & make every effort possible to solve it accurately. |

©IJRASET: All Rights are Reserved 

| | | | | | |
|---|---|---|---|---|---|
| 18 | Xinsheng Wang & Tingting Qiao [18] | 2021 | S2IGAN | introduced a GAN technique to produce high-quality images based on the voice embeddings that the speech encoder in the speech embedding network had extracted. | S2IGAN gives FID value 48.64 on oxford dataset shows the highest accuracy |
| 19 | Maofu Liu & Huijun Hu [19] | 2020 | NICVATP2L model with visual attention and topic modelling | The neural image caption (NIC) model's performance and limitations in applying deep neural networks for image captioning were discussed. | 98% accuracy in generating sentences |
| 20 | Jezia Zekraouii & Samir Elloumi [20] | 2019 | MT tool, CNN/LSTM image captioning model | The author described a method for automatically translating short, straightforward modern standard Arabic children's stories into the best possible visuals that effectively convey the words' meanings. | 49% mean average precision for English keywords instead of Arabic |
| 21 | Fengnan Quan & Bo Lang [21] | 2022 | Text-to-image synthesis GAN, | Developed high-resolution, photorealistic images that correspond to the text in order to introduce an ARRPNGAN model for text-to- image synthesis. | 91.86% accuracy on COCO dataset |
| 22 | Vamsidhar Talasila & Narasingarao M R [22] | 2022 | BI-LSTM Optimized GAN DC-WO Model | To make it easier to understand the text, the author created an improved GAN model that generates related images from texts or captions. | Shows 77% accuracy on CUB dataset |
| 23 | Zhanzhan Cheng & Jing Lu [23] | 2021 | Spatial-temporal video text detector, text recommender | The localized text stream is only recognized once, as opposed to frame-by-frame detection, in the fast and reliable FREE end-to-end video text spotting framework. | 90.30% accuracy on YVT dataset |
| 24 | Zhongyi Ji & Wenmin Wang [24] | 2020 | Semi-supervised Learning | Author introduced a semi-supervised approach is presented to generate images from text descriptions by using Modality-invariant Semantic-consistent Module | 90.72% accuracy on Un-labeled rate 0.2 |
| 25 | ingh & Ritika Shenoy[25] | 2021 | GAN | Author discussed GAN model for generating images from the input text/captions automatically for easier acknowledge | 10 caption per image gives accurate output |
| 26 | Tobias Hinz & Stefan Heinrich [26] | 2022 | SOA | Author introduced a new model for generating the images from the captions and evaluation of images given an image caption. | 74.97% accuracyon SOA-C model |
| 27 | Xiaojin Zhu & Andrew B. Goldberg[34] | 2007 | Text - to -Picture system (TTP) | Synthesized pictures convey information about children’s stories to enhance comprehension by using the TTP model. | 84% accuracy on MNIST dataset |
| 28 | Xinchen Yan & Jimei Yang | 2016 | Conditional Variational AutoEncoder (CVAE) | Study on generating images from visual attributes using the CAVE model and | disCVAE gives 13% on R- precision CUB |

| | | | | | |
|---|---|---|---|---|---|
| | | | | optimization-based approach to posterior inference for evaluation | dataset |

## IV. ANALYSIS OF DATASETS AND EVALUATION

### A. Analysis of Dataset

The results reveal the widespread usage of publicly available datasets in literature for developing text-to-image synthesis models. Though the digital world is full of pictures, there is a scarcity of captioned datasets needed for training. Researchers working on text-to-picture synthesis frequently use two popular datasets, such as the Oxford-102 flowers dataset [40] and the Caltech CUB-200 birds' dataset [20]. Oxford-102 has 8,192 photos of flowers in 102 different categories. There are 11,788 photos of 200 different bird species in the CUB-200 dataset. The description is not included in the datasets; it consists of only photos. Another often-used dataset is Microsoft COCO Common Objects in Context. It includes 123,287 images of intricatelydetailed scenes from everyday life that show typical objects in their natural settings. Five captions are annotated next to each image. The collection contains images of 91 different kinds of items. In 328k images, there are a total of 2.5 million labeled instances. Recently, Text to image synthesis using the Imagen Sandford university dataset [25] has also been employed to provide better results.

### B. Evaluation Metrics

The evaluation of generative models is an emerging research area. Considering that most generative models increase the likelihoodof the data, the average loglikelihood is used as a metric to evaluate models. Human annotators can assess the visual quality ofan image to obtain the performance metrics used in the previous method for testing GAN models. The metric fluctuates depending on how the work is arranged and how motivated the annotators are, which is a disadvantage of utilizing human annotators. We also see that when we provide annotators with feedback regarding their errors, the results significantly alter: Annotators get better at pointing out the shortcomings in created images and offer a more pessimistic quality assessment after receiving such input. The evaluation measures that are frequently used to assess GANs are listed below.

#### 1) Quantitative metric of image quality: Inception Score (IS)

To evaluate the GAN model, we are using the Inception network to that generate a distribution PY |X where PY |X(y|x) is the probability assigned to image x to belong to class y. This is done by treating the images as a random vector X and the image labels as a random variable Y. To generate these probabilities for the classes from the test dataset the GAN will be assessed on, an Inception network is trained. This presupposes a dataset with classes. The evaluated model's output images are then given classifications by the trained network. The distribution of the anticipated classes affects the score. There are two main objectives:

*a)* The object in any image x must be clearly discernible, and hence, the conditional distribution PY |X must have lowentropy.

*b)* As many different kinds of images as possible should be generated. That means all of the classes in the dataset shouldinclude the images rather than just a small subset of classes. In other words, the PY distribution ought to have high entropy.

These two factors encourage the Inception Score's form from Eq. (2) because, if they hold, the KL divergence between the two distributions is large. Only for aesthetic purposes, the exponential function is employed to increase the range of values for thescore.

$$IS(G) = \exp\left(\mathbb{E}_{\mathbf{X}\sim P_g}\left[KL\left(P_{Y|\mathbf{X}}(y|\mathbf{x}) \parallel P_Y(y)\right)\right]\right) \tag{2}$$

It has been demonstrated that the Inception score and subjective assessments of image quality correlate well [37]. We are not using a human assessment to assess the models due to this reason.

#### 2) Quantitative metric of image quality: Fréchet Inception Distance (FID)

The Fréchet inception distance is used to compare Inception activation responses between real and generated images proposed by the author in the paper [38]. Using features retrieved from the pre-trained Inception-V3 network, it determines the Fréchet distance between the generated image and the real image;

$$d^2((m,C),(m_w,C_w)) = \|m - m_w\|_2^2 + Tr(C + C_w - 2(CC_w)^{\frac{1}{2}}) \tag{3}$$

The generated image distribution and the real image distribution are more similar to one another as the FID decreases.

*3) Text-Image Semantic Consistency Metric: R-Precision*

R-precision is a popular evaluation statistic for sorting information retrieval outcomes. R-precision, where R is the number of pertinent documents for a particular query topic Q, is the precision at R. If among the top-R recovered documents there are r relevant documents, then R-precision is:

$$R - precision = r/R \quad (4)$$

## V. RESULT & DISCUSSION

In this section, we are examining how well various approaches perform in relation to the benchmark dataset and their accompanying performance indicators, such as IS, FID, R, and F1 Score. From the paper [34] text to image synthesis is evaluated by F1 score ie, the value is 84. Result from paper

[35] shows only 13% in precision. So, we can't able to exactly get performance accuracy by using precision and F1 score. So that recent studies focus on Inception Score as a standard method for evaluating GANs. The results in Table 2 reveal that StackGAN++ only slightly outperformed its predecessor, StackGAN [40], for text-to-image synthesis when measured by IS, which is the metric that was used to compare most models, with the exception of DC-GAN [39]. In FID DC GAN has more than 60% than StackGAN.

The performance comparison of various GAN models in the following graph (Figure 5) is shown in relation to the scores at which they were first developed (ISs). The highest IS is attained by ARRPNGAN on both the CUB and COCO datasets. When compared to the most recent DM-GAN [41], ARRPNGAN increases CUB scores from 4.82 to 5.22 and COCO scores from

31.06 to 32.98. Table 2 displays the ARRPNGAN Fréchet inception distance. Additionally, R- precision has the best impact on the semantic consistency of text-image pairing, and the FID of ARRPNGAN is the lowest, indicating that the generated picture is closer to the genuine image. In addition to this recent GAN, model parti will show 75 % in FID score indicating that it also gives better accuracy on Imagen Dataset. The designs and improvement in accuracy in ARRPNGAN [21] and Parti-GAN [42] show that text/caption-to-image synthesis research is progressing to place more attention on the picture details and text semantics for improved comprehension and perception.

TABLE 2. Performance evaluations of various approaches using dataset and evaluation metrics

| | | Datasets and metric | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | | CUB | | | COCO | | | Oxford | | | MNIST | Imagen |
| | | IS | FID | R | IS | FID | R | IS | FID | R | F1 SCORE | FID |
| raditional models | TTP [34] | - | - | - | | - | - | - | - | - | 84 | - |
| | disCVAE [35] | - | - | 13 | - | - | - | - | - | - | - | - |
| DifferentGAN models | DCGAN [39] | 2.88 | 68.79 | - | 7.88 | 60.82 | - | 2.66 | 79.55 | - | - | - |
| | StackGAN [40] | 3.74 | 51.89 | - | 8.60 | 74.05 | - | 3.21 | 55.28 | - | - | - |
| | StackGAN++ [40] | 4.09 | 15.30 | - | 8.40 | 81.59 | - | 3.27 | 48.68 | - | - | - |
| | AttnGAN [20] | 4.39 | 23.98 | 72.25 | 26.36 | 35.49 | 89.16 | - | - | - | - | - |
| | DM-GAN [41] | 4.82 | 16.09 | 73.22 | 31.06 | 32.64 | 88.84 | - | - | - | - | - |
| | ARRPNGAN [20] | 5.22 | 14.28 | 77.31 | 32.98 | 29.14 | 91.86 | - | - | - | - | - |
| | ViT-VQGAN(Parti) [42] | - | - | - | - | - | - | - | - | - | - | 75.5 |

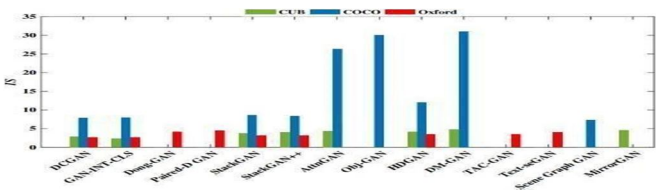



Fig.5. Performance comparison of different GANs with respect to Inception Score (IS)

## VI. CONCLUSION

We are looking for new methods to consume video content from subtitles or text content as the use of videos on the web grows tremendously. Which will enhance the learning process through video caption accessibility for children with learning disabilities. So, it is possible to improve performance by creating neural networks with greater strength. From our study, we proposed that the General Adversarial network is a trending research area for the generation of images from the corresponding caption or text. Consequently, this not only assures the compatibility of the semantics of text and image pairing but also givesa visual representation that will increase the comprehension of captions.

This study can be extended into the accessibility of video captions in different languages other than English. Furthermore, plenty of research is being carried out in the area of GAN so, there will be various fields in the future where text-to-image synthesis will be utilized rather than multimedia accessibility. Medical research and advanced treatment will certainly get a boost with this technique. Medical images dataset can be added up with synthetic images by using improved text-to-image synthesis, thus assisting medical practitioners in better diagnosis.

## REFERENCES

[1] R. Ismail and A. Jaafar, “Interactive screen-based design for dyslexic children,” in 2011 International Conference on User Science and Engineering (i-USEr ), 2011, pp. 168–171

[2] L. Rello and R. Baeza-Yates, “How to present more readable text for people with dyslexia,” Univers Access Inf Soc, vol. 16, no. 1, pp. 29– 49, Mar. 2017, doi: 10.1007/s10209-015-0438-8.

[3] N. I. R. Ab Llah, K. H. A. Firdaus, and J. Ibrahim, “Modeling the needfor decision support systems for dyslexic children using BMC,” in Proceedings - International Conference on Information and Communication Technology for the Muslim World 2018, ICT4M 2018,Dec. 2018, pp. 114–119. doi: 10.1109/ICT4M.2018.00030.

[4] J. Rodriguez, Ed. D and M. V. Diaz, M.S.D, “Media with Captions and Description to Support Learning among Children with Sensory Disabilities,” Universal Journal of Educational Research, vol. 5, no. 11, pp. 2016–2025, Nov. 2017, doi: 10.13189/ujer.2017.051118.

[5] B. Bhavya et al., “Exploring collaborative caption editing to augment video-based learning,” Educational Technology Research and Development, Oct.2022, doi: 10.1007/s11423-022-10137-5.

[6] H. Yang and C. Meinel, “Content based lecture video retrieval using speech and video text information,” IEEE Transactions on Learning Technologies,vol. 7, no. 2, pp. 142–154, 2014, doi: 10.1109/TLT.2014.2307305.

[7] H. Gowhary, Z. Pourhalashi, A. Jamalinesari, and A. Azizifar, “Investigating the Effect of Video Captioning on Iranian EFL Learners’ Listening Comprehension,” Procedia Soc Behav Sci, vol. 192, pp. 205– 212, Jun. 2015, doi: 10.1016/j.sbspro.2015.06.029.

[8] M. Li, “A High-Efficiency Knowledge Distillation Image Caption Technology,” in Lecture Notes in Electrical Engineering, 2022, vol. 942 LNEE, pp. 912–917. doi: 10.1007/978-981-19-2456-9_92.

[9] M. Pantula and K. S. Kuppusamy, “A metric to assess the readability of video closed captions for the persons with low literacy skills,” ComputerJournal, vol. 63, no. 7, pp. 1063–1075, 2020, doi: 10.1093/COMJNL/BXZ074.

[10] M. Pantula and K. S. Kuppusamy, “Fast captions: Towards making thevideo content consumption quicker,” Computers in Entertainment, vol.16, no. 4, Nov. 2018, doi: 10.1145/3276323.

[11] M. Pantula and K. S. Kuppusamy, “A model to measure readability of captions with temporal dimension,” in Smart Innovation, Systems and Technologies, 2018, vol. 79, pp. 225–234. doi: 10.1007/978-981-10- 5828-8_22.

[12] M. Pantula and K. S. Kuppusamy, “AuDIVA: A tool for embedding Audio Descriptions to enhance Video Accessibility for Persons with Visual Impairments,” Multimed Tools Appl, vol. 78, no. 14, pp. 20005–20018, Jul. 2019, doi: 10.1007/s11042-019-7363-4.

[13] J. Liu, M. Yang, C. Li, and R. Xu, “Improving Cross-Modal Image- Text Retrieval with Teacher-Student Learning,” IEEE Transactions onCircuits andSystems for Video Technology, vol. 31, no. 8, pp. 3242– 3253, Aug. 2021, doi: 10.1109/TCSVT.2020.3037661.

[14] M. Pantula and K. S. Kuppusamy, “CORDIF: A machine learning- based approach to identify complex words using intra-word feature set,” in Lecture Notes in Electrical Engineering, vol. 478, Springer Verlag, 2019, pp. 285–296. doi: 10.1007/978-981-13-1642-5_26.

[15] H. Parikh, H. Sawant, B. Parmar, R. Shah, S. Chapaneri, and D. Jayaswal, “Encoder-decoder architecture for image caption generation,” in 2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA), 2020, pp. 174–179.

[16] M. Singh and A. Kaur, “An efficient hybrid scheme for key frame extraction and text localization in video,” in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015, pp. 1250–1254.

[17] S. Komanduri, Y. M. Roopa, and M. Madhu Bala, “Novel approach forimage text recognition and translation,” in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 596–599.

[18] X. Wang, T. Qiao, J. Zhu, A. Hanjalic, and O. Scharenborg, “Generating Images from Spoken Descriptions,” IEEE/ACM Trans Audio Speech Lang Process, vol. 29, pp. 850–865, 2021, doi: 10.1109/TASLP.2021.3053391.

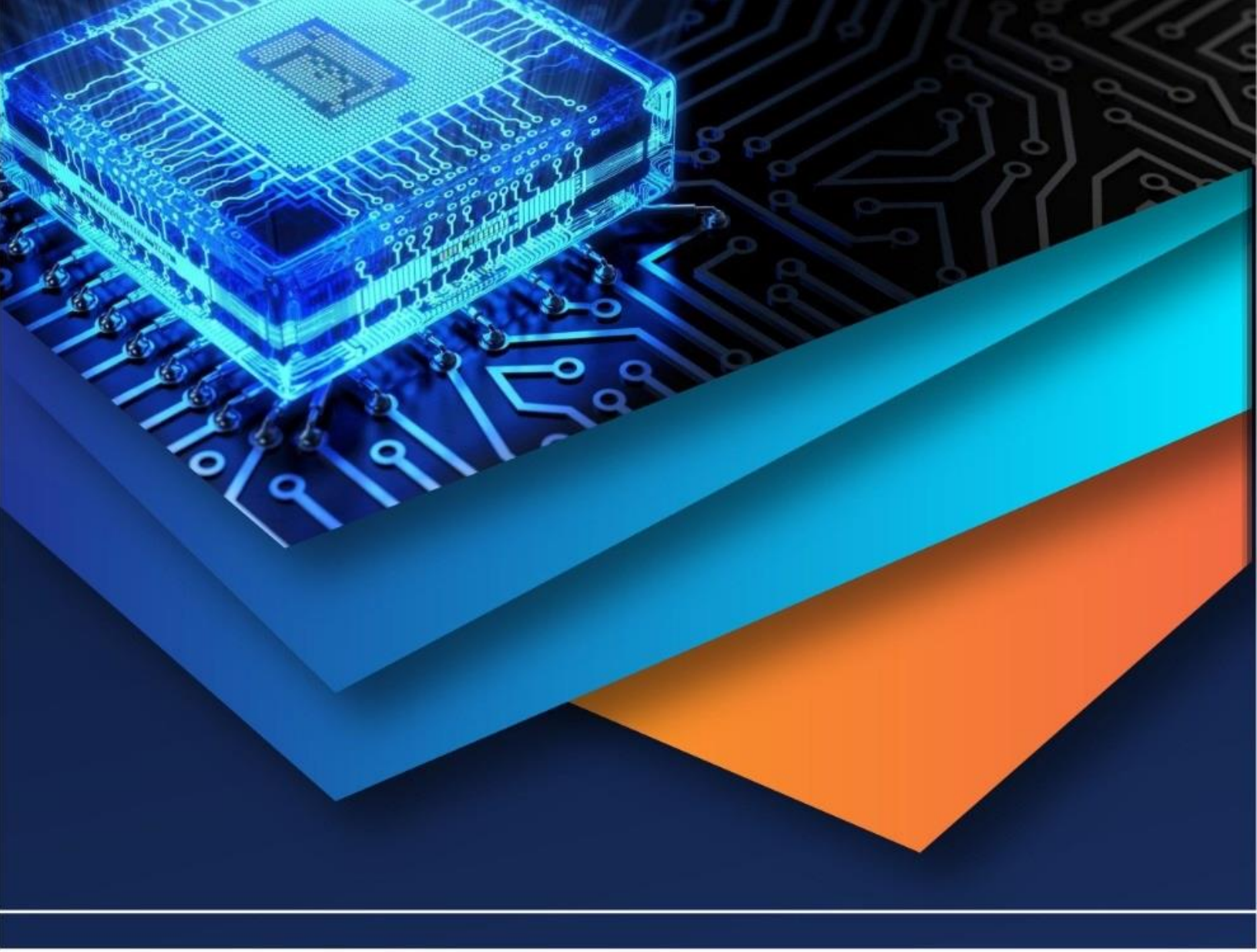
[19] M. Liu, H. Hu, L. Li, Y. Yu, and W. Guan, “Chinese Image Caption Generation via Visual Attention and Topic Modeling,” IEEE Trans Cybern, vol. 52, no. 2, pp. 1247–1257, Feb. 2022, doi: 10.1109/TCYB.2020.2997034.

[20] J. Zakraoui, S. Elloumi, J. M. Alja’am, and S. ben Yahia, “Improving Arabic Text to Image Mapping Using a Robust Machine Learning Technique,” IEEE Access, vol. 7, pp. 18772–18782, 2019, doi: 10.1109/ACCESS.2019.2896713.

[21] F. Quan, B. Lang, and Y. Liu, “ARRPNGAN: Text-to-image GAN withattention regularization and region proposal networks,” Signal ProcessImage Commun, vol. 106, Aug. 2022, doi: 10.1016/j.image.2022.116728.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[22] V. Talasila, N. M R, and M. M. V, "Optimized GAN for Text-to-ImageSynthesis: Hybrid Whale Optimization Algorithm and Dragonfly Algorithm," Advances in Engineering Software, vol. 173, Nov. 2022, doi: 10.1016/j.advengsoft.2022.103222.
[23] Z. Cheng et al., "FREE: A fast and robust end-to-end video text spotter," IEEE Transactions on Image Processing, vol. 30, pp. 822– 837, 2021, doi: 10.1109/TIP.2020.3038520.
[24] Z. Ji, W. Wang, B. Chen, and X. Han, "Text-to-image generation via semi-supervised training," in 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), 2020, pp. 265–268.
[25] A. Singh and S. Anekar, "Text to Image using Deep Learning; Text to Image using Deep Learning," www.ijert.org, vol. 10, 2021, [Online]. Available: www.ijert.org
[26] T. Hinz, S. Heinrich, and S. Wermter, "Semantic Object Accuracy for Generative Text-to-Image Synthesis," IEEE Trans Pattern Anal Mach Intell, vol.44, no. 3, pp. 1552–1565, Mar. 2022, doi: 10.1109/TPAMI.2020.3021209.
[27] Boyer, C. (2000) Libraries and section 508 of the rehabilitation act. Library Hi Tech News (Vol. 17)
[28] Improving speech playback using time-compression and speech recognition. Proc. SIGCHI Conf. Human Factors in Computing Systems, pp. 295–302. ACM. – 22.
[29] Caldwell, B., Cooper, M., Reid, L.G. and Vanderheiden, G. (2008) Time-based media: understanding guideline 1.2. WWW Consortium (W3C) (Vol. 11).
[30] A. Halko, "28 video stats for 2018," Insivia, 10-Jan-2018. [Online]. Available: https://www.insivia.com/28-video-stats-2018.
[31] Caldwell, B., Cooper, M., Reid, L.G. and Vanderheiden, G. (2008) Webcontent accessibility guidelines 2.0. W3C Recommendation (Vol. 11).
[32] Henry., S. L. (2016). Multimedia accessibility faq. https://www. w3.org/2008/06/video-notes/ (accessed October 19, 2017).
[33] R. Sawant, A. Shaikh, S. Sabat, and V. Bhole, "Text to image generation using GAN," SSRN Electron. J., 2021.
[34] Zhu, X., Goldberg, A. B., Eldawy, M., Dyer, C. R., & Strock, B. (2007).A text-to-picture synthesis system for augmenting communication. In Proceedings of the 22nd national conference on artificial intelligence (pp. 1590–1595). Vancouver, British Columbia, Canada: AAAI Press.
[35] Yan, X., Yang, J., Sohn, K., & Lee, H. (2016). Attribute2image: Conditional image generation from visual attributes. In Computer vision— ECCV 2016 (pp. 776–791). Amsterdam, The Netherlands: Springer International.
[36] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B.,& Bharath, A. (2018). Generative adversarial networks: An overview. In IEEE signal processing magazine (pp. 53–65). IEEE.
[37] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In NIPS, 2016.
[38] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local Nashequilibrium, in: Adv. Neural Inf. Process. Syst., 2017
[39] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. In Proceedings of the 33rd international conference on international conference on machine learning (pp. 1060–1069). New York, NY. Retrieved from JMLR.org
[40] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, Z., Huang, X., & Metaxas, D. (2017). Least squares generative adversarial networks. In 2017 IEEEinternational conference on computer vision (ICCV) (pp. 2813–2821). Venice, Italy: IEEE Computer Society.
[41] M. Zhu, P. Pan, W. Chen, Y. Yang, DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2019-June, 2019, pp. 5795–5803, http: //dx.doi.org/10.1109/CVPR.2019.00595.
[42] J. Yu et al., "Scaling autoregressive models for content-rich text-to- image generation," arXiv [cs.CV]



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)