المجلة المصرية لعلوم المعلومات

Building Text to Image Generative AI Model (T2I) for Images Retrieval from Text Queries

A Practical Study

DR. Moamen Sayed Othman El-Nasharty

Lecturer at the Department of
Libraries, Documents, and Information, Faculty of Arts, Cairo University

DOI	تاريخ القبول	تاريخ الإرسال
10.21608/jesi.2024.312821.1137	2024/8/28	2024/8/15

Abstract:

This study aims at investigating the role of generative models in revolutionizing image retrieval methods. By leveraging generative models' ability to understand and generate visual content, and to explores the abilities of Generative Adversarial Networks (GANs), that have demonstrated the capability to capture intricate patterns and semantic information within images, opening new avenues for content-based image retrieval. At the same time This research presents the development and implementation of a sophisticated Text to Image Generative AI Model (T2I) tailored for image retrieval tasks based on natural language descriptions. The most prominent results have been demonstrated by the extensive evaluation was the model's effectiveness and efficiency in generating accurate and contextually relevant images corresponding to given textual queries.

Keywords: Image Retrieval System – TBIR – CBIR -SBIR- Generative Models – GANs – T2I.

المستخلص

تهدف هذه الدراسة إلى التحقيق في دور النماذج التوليدية في إحداث ثورة في طرق استرجاع الصور. من خلال الاستفادة من قدرة النماذج التوليدية على فهم وتوليد المحتوى المرئي، تستكشف الدراسة قدرات الشبكات التوليدية (GANs)، التي أظهرت القدرة على التقاط الأنماط المعقدة والمعلومات الدلالية داخل الصور، مما يفتح آفاقًا جديدة لاسترجاع الصور القائم على المحتوى. في الوقت نفسه، يقدم هذا البحث تطوير وتنفيذ نموذج ذكاء اصطناعي توليدي متطور لتحويل النصوص إلى صور (T2I) مصمم خصيصًا لمهام استرجاع الصور بناءً على الأوصاف اللغوية الطبيعية. وقد تجلت أبرز النتائج من خلال التقييم الشامل لفعالية وكفاءة النموذج في توليد صور دقيقة وذات صلة سياقية تتوافق مع الاستعلامات النصية المقدمة.

الكلمات المفتاحية: نظام استرجاع الصور – استرجاع الصور المستند إلى النص – (TBIR) استرجاع الصور القائم على المحتوى – (CBIR) استرجاع الصور القائم على الشكل – (SBIR) النماذج التوليدية – شبكات GANs – تحويل النص إلى صورة.(T2I)

Introduction:

Nowadays, the world has a ton of images, and finding the right one quickly is a big challenge with exponentially growing of multimedia contents, especially the images, the search or retrieve a relevant image from an open platform like Facebook, Twitter, and Instagram is a challenging research problem. In the last few years, the Images Retrieval has seen significant advancements through the incorporation of generative models.

According to (Priyatharshini, 2013) An image retrieval system is a computer system for browsing, searching, and retrieving images from a large database of digital images.

The nature of Image retrieval is the task of finding visually similar images in a database given a query by text or Images, where traditional image retrieval techniques often rely on metadata or simplistic features, that leads to limiting their ability to handle the complex and diverse nature of Images and visual data (Zhang, 2012, T. Khalil, 2018).

Traditional methods often rely on predefined rules and patterns, which may fall short when dealing with the complexity and diversity of real-images data, adding to that, the continuous advancements in digital image processing, and data storage has reached an optimal level, makes image search and retrieval a difficult task (Dahake, 2028).

To solve the image retrieval problem, many techniques have been devised addressing the requirement of different applications.

This challenge has fueled the exploration of advanced techniques in the realm of artificial intelligence, leading to the emergence of Generative Models as powerful tools in information retrieval. The image retrieval systems and engines has seen significant advancements through the incorporation of Generative Models, which are smart computer programs, can help us search for images in a much better way.

The generative models have revolutionized the way we perceive and interact with visual data, where these models are offering the ability not just to replicate existing data but to generate entirely new and meaningful content.

Problem Statement:

In the era of rapid digital content creation and consumption, the demand for efficient and intuitive methods of retrieving visual information (images) from textual queries has surged, at the same time, finding the right one quickly is a big challenge.

According to (Qiu, 2022) more than 3 billion images and 700,000 hours of video are shared on social media daily. When dealing with such a flood of content, researchers and practitioners are confronted with the challenge of how to efficiently index the image and video data and develop friendly tools to enable users to quickly find what they are looking for. Indexing and retrieval of images and videos are very challenging due to the primitive nature of their raw data representations which lack readily available structural and semantic information. Performing image and video retrieval requires to first process the data to extract discriminative, meaningful, and interpretable features, and then to gain high level understanding at the object, scene, and semantic levels, and finally to develop systems and tools to efficiently index the data and to help users to find what they are looking for intuitively, easily, and accurately.

In the domain of image retrieval, there are many challenges at the three main approaches in retrieving images:

1- In the images retrieval based on text which is done manually by a human. This method necessitates the use of text to represent an image and thus it needs a significant amount of effort and time. The traditional methods often fall short in capturing the complexity and diversity of visual data because based on textual annotations and metadata that leading to fall short when dealing with the sheer volume and complexity of image data available today, leaving a significant gap in bridging the semantic understanding between human-generated textual descriptions and the corresponding visual representations.

2- To get around this limitation, a new approach named as CBIR (Content-Based Image Retrieval) has been developed which describes images based on their features like texture, color, and shape (Agrawal, 2022). Content-Based Image Retrieval (CBIR) describes images based on their features like texture, color, and shape (S. Rubini, 2018), it is a technology that allows users to search for images based on their visual content, rather than relying on text-based descriptions or metadata (Thamotharan, 2013). While CBIR offers several advantages such as: going beyond textual descriptions and understands the visual content of the image, (Silva, 2013). But CBIR, has some limitations and challenges, such as Semantic Gap It refers to the disconnect between low-level visual features extracted from images (e.g., color, texture, shape) and high-level human perception and interpretation of those features. Bridging this gap to accurately understand the meaning and context of images is a challenging task (Abioui, 2019).

3- Semantic-based Image Retrieval (SBIR) system has been introduced by (Jing Li, 2006), (Yiqing Guo, 2018), (Hossein, 2008), (Xiao-Feng Wang, 2008), (Kunshan, 2016), (Ajay, 2022), The general architecture of semantic-based image retrieval system is divided into two categories, namely feature extraction and query processing. The feature extraction section concerns with the extraction of semantic features of the database images. in the query processing phase the user specifies the query using the keywords or textual words. When the user submits the query using the keywords or texts, the system will go through two main steps that are semantic feature translator and semantic image mapping (Jagtap, 2021). However, SBIR has set of limitations as: the training and running complex SBIR models can be computationally expensive, especially for large datasets. This can limit its accessibility and scalability (Barz, 2021).

According to (Abioui, 2019) the best solution that combines text-based image retrieval and the content-based techniques and semantic based image retrieval by using Machine Learning and AI techniques. That is what this study is trying to do.

The Study Objectives:

The primary objectives of this research are as follows:

1. Handling shortcomes of both Content-Based Image Retrieval (CBIR) and Semantic-based Image Retrieval (SBIR) by Examine limitations and challenges and how generative models may be used for tasks of (CBIR) & (SBIR), enabling users to generating images based their demand.
2. Building a Text to Image AI Generative Model (T2I): by Exploring Various Generative Models and their abilities, with focusing to Generative Adversarial Networks (GANs), and their variants and understand these underlying mechanisms, strengths, and limitations in the context of image retrieval, to generate a new image from text queries, not replicate or retrieving images have been collected or capturing before and how these generative models can make image searching easier and more accurate.

Methodology:

This study has been conducted based on two methods: the first method was the analytical approach by conducting an extensive literature review to understand the generative models used in image generating, retrieval and analyze.

The second approach was applying the GAN main algorithms to build Generative Model has ability to create and generating (not retrieving) a new image based on text queries.

Literature Review:

The nature of Image retrieval is the task of finding visually similar images in a database given a query by text or Images, where traditional image retrieval techniques often rely on metadata or simplistic features, that leads to limiting their ability to handle the complex and diverse nature of Images and visual data.

According to (Farooque, 2003), the origin of “Images” term backs to the Latin word "Imago", which means a pictorial representation of a person, scene, or object. An image is a realistic or semi-realistic representation of a variety of subjects produced by several methods and in several different styles. The term "picture" is also frequently used in the literature; therefore, the terms "picture" and "image" are both used in appropriate contexts.

In other words, the image can be described here is any object that could be considered graphical in nature, and image is defined as a combination of physical attributes set, referring to the image content, and metadata set referring to its context (Enser, 2000). That’s include, but is not limited to, photographs, slides, digital images, and any object that is not textual in nature (Farooque, 2003).

The basic diagram of the automatic image retrieval system has been presented by (Pradhan, 2021) In Fig. 1, according to that, the user gives the input query, and the searching algorithm finds the most relevant images from the annotated image dataset.

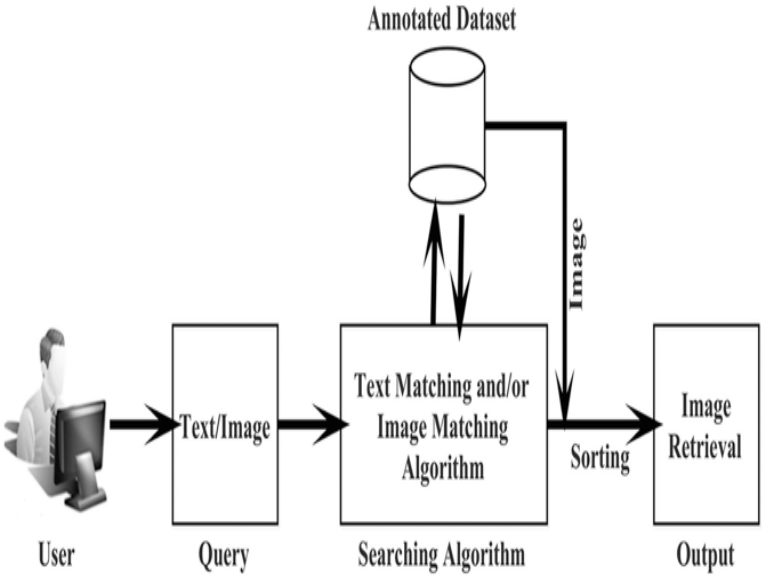


Fig (1) Explain the automatic image retrieval system (Pradhan, 2021).

The user can provide input query in form of text as well as an image. In this automatic image retrieval system, the searching algorithm access the meta-data of the annotated data-set and uses the text/image matching approach to compare the query information and data-set information (Pradhan, 2021).

Further, based on comparison score, it will perform sorting of top matched images. Finally, from the sorted images it will retrieve the most similar images as a final output. The images are processed as based on two categories:

1. First category is internal metadata extracted from the image itself, is processing images as an array of pixels that addressing an image feature related to the content-based image retrieval (CBIR) method which are: the image colors (Sakhare, 2011), the shape (Zhang, 2004), and the texture features (Westerveld, 2000).
2. The second category is external metadata collected from the text surrounding the image, is presenting attributes of meaning or metadata which define the image context, and qualified (Tamura, 1987).

Researchers around the globe have introduced different kinds of automatic image retrieval systems which take text, meta-data, and/or image as a query input and retrieve the most similar images from the image dataset. Here, Fig. 2 shows the taxonomy of the image retrieval system (Pradhan, 2021).

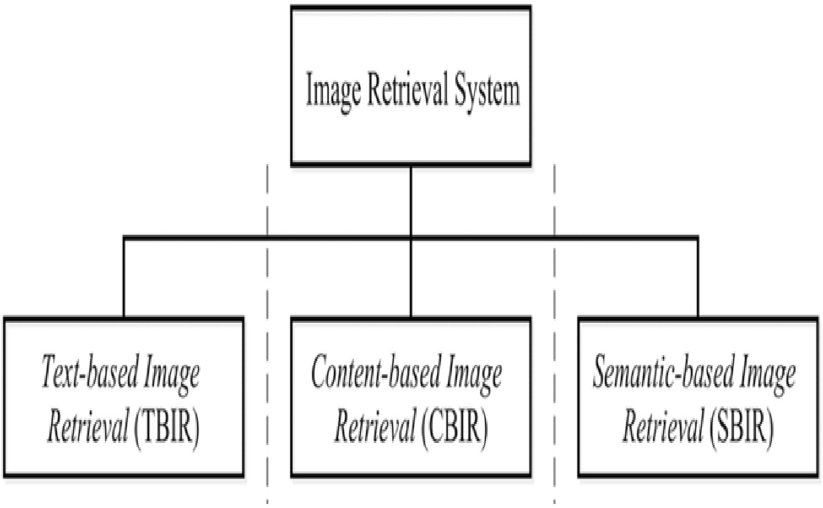


Image Retrieval System
Text-based Image Retrieval (TBIR)
Content-based Image Retrieval (CBIR)
Semantic-based Image Retrieval (SBIR)

Fig 2: Taxonomy of the image retrieval systems (Pradhan, 2021).

Traditional image retrieval methods primarily relied on metadata, manually annotated keywords, or textual descriptions associated with images.

(Pradhan, 2019) indicated that in last decade contemporary researchers have suggested many images retrieval schemes which use text-based image information, content-based image information or semantic information of an image.

Text-based Image Retrieval (TBIR) methods were using in automatic image retrieval systems, which relied on texts that contained one or more keywords related to the picture (Sean, 2002).

Traditional image retrieval schemes (Burger, 2009) use the text information associated with the images like, name of the image, image number, note title, file name, indexing icons and keywords for retrieval purposes. (Gong, 1994) proposed an image retrieval method which uses numerical index generated from primitive image features by using some set of rules and the traditional descriptive keywords.

This approach had several limitations, including the need for human annotation, inconsistencies in keyword assignment, and difficulty in retrieving images based on their content, which often resulted in suboptimal user experiences.

Two serious problems with TBIR systems exist. First, human intervention is required in the tag assignment procedure for allocating tags to each unique database image. the second is the primary shortcoming of tag-based image retrieval systems is that a single term is insufficient to represent all the information contained in a picture (Pradhan, 2021).

CBIR Content-based image retrieval (CBIR) (Gudivada, 1995) (Huang, 2003) (Rui Y, 1999) or Query by image content (QBIC), is considers a type of image retrieval process, it considers the salient visual contents from images to accelerate the retrieval process since the straightforward image to image searching is not realistic to handle such kind of large volume of data. That is depending on the visual features like colors, shapes, and spatial relations, which are incorporated in the retrieval of most relevant images from the image database.

CBIR (Gudivada, 1995) The Content-Based Image Retrieval technique involves searching for images based on their content. CBIR is a straightforward method that consists of three phases: feature extraction, similarity assessment, and picture retrieval (Ghaleb, 2022).

CBIR can help with fingerprint recognition, digital libraries, biodiversity data management, web image analysis and education. Many efforts used this technique to improve images retrieval accuracy (Rani, 2020), (Ramya, 2018), (Tadasare, 2018), (Wang, 2013).

Usually, a CBIR system works in four steps which are as follows (Pradhan, 2021):

1- Feature Extraction: In this step, the researchers use color, texture, shape, structural, co-relational feature extraction schemes, or combination of these schemes to extract visual features of the query image.

2- Feature Space Generation: In this step, is being constructed feature vector for all database images.

3- Similarity Matching: In this step, different distance measurements schemes have been used to find the similarity between query image feature vector and the feature vectors of the feature space.

4- Image Retrieval: This is the final step in which sorting of images have been performed based on their similarity score.

CBIR system usually works better when image database contains relatively a smaller number of images. As the size of database increases it adversely affects the retrieval efficiency and speed of the CBIR system (Pourghassem, 2008).

Content-Based Image Retrieval (CBIR) faces limitations (A.W.M, 2006), (Y. Rui, 2003), (M. Datar, 2007), such as:

- Semantic gap: CBIR systems rely on low-level visual features like color, texture, and shape. This can lead to the "semantic gap" problem, where visually similar images might not be semantically relevant to the query.

- Subjectivity: Image interpretation is highly subjective, varying from one person to another. What one user considers a relevant image may not match another user's perception. CBIR systems may not always capture these subjective nuances effectively.
- Limited by Feature Extraction: CBIR systems rely on feature extraction techniques to represent images. The choice of features and their quality significantly impacts retrieval results. Different feature sets may be more suitable for specific types of images, making it challenging to develop a one-size-fits-all CBIR system.
- Scalability: Scalability can be an issue when dealing with large image databases. As the number of images increases, the time required for feature extraction and retrieval can become a bottleneck, leading to decreased system efficiency.
- Data Annotation: Building a large-scale CBIR system often requires extensive manual or automated annotation of images. This process can be time-consuming and costly, and errors in annotations can lead to retrieval inaccuracies.
- Relevance Ranking: CBIR systems may not always produce a ranking of images that matches user expectations. They often require post-processing or additional context to refine results.
- Noise and Irrelevant Features: CBIR systems can be sensitive to noise in images and irrelevant features. Noisy or cluttered images may produce suboptimal retrieval results.

Semantic-Based Image Retrieval (SBIR) is a technique that aims to retrieve images based on their semantic meaning, rather than just their visual features. so that, it builds bridges on the gap between the low-level features and the high-level concepts that humans use to understand images which used in traditional Content-Based Image Retrieval (CBIR) systems (Alzu’bi, 2015).

The Key Components of SBIR are:

- Semantic Feature Extraction: which includes 3 elements:
  - ○ Textual Annotation: Automatically or manually adding keywords, tags, or descriptions to images.
  - ○ Concept Learning: Using semantic techniques to extract high-level concepts (e.g., "dog," "beach," "wedding") from image content and associated text.

- Knowledge Graphs: Leveraging structured knowledge bases to link images with semantic relationships and entities.

- Semantic Similarity Measures: which includes.
  - Employing techniques that go beyond visual similarity to measure the semantic relatedness of images.
  - Using semantic similarity metrics based on knowledge graphs, word embeddings, or other semantic representations.

SBIR has many advantages such as: more Relevant Results: Retrieves images based on their meaning, improving relevance to user queries. Another point of SBIR is reduced Semantic Gap: Addresses the discrepancy between low-level visual features and high-level human understanding of images. SBIR, Enhanced User Experience: Allows for more intuitive and natural image search based on concepts and ideas. SBIR provides Cross-Modal Retrieval: Enables retrieval of images based on text queries and vice versa.

Despite that, SBIR faces many challenges such as

- Complex Semantic Representation: While SBIR systems leverage textual descriptions or semantic concepts for retrieval, accurately representing complex and nuanced semantics in text can be difficult. Some descriptions might be abstract or ambiguous, leading to challenges in interpretation.
- Ambiguity in Language: Natural language is inherently ambiguous. The same word or phrase can have multiple meanings based on the context. Resolving these ambiguities in textual queries is crucial to generating images that align with the user's intent.
- Dependency on Text Quality: The quality and accuracy of the textual annotations or semantic concepts provided heavily influence the retrieval results. Inaccurate or incomplete annotations can lead to incorrect or irrelevant image matches.
- User Intent Variability: Users may have varied intents when describing a particular scene or object. Different users might use different words or phrases to describe the same visual content, making it challenging to create a one-size-fits-all retrieval system.

- Limited Vocabulary: The effectiveness of SBIR systems is constrained by the vocabulary used for annotation. If the provided concepts do not align with user queries, relevant images might be missed (Zarchi, 2014).

Addressing these challenges involves advancements in machine learning techniques, natural language processing, and multimodal learning to create more robust and accurate image retrieval systems, and the last one is Generative models.

Generative AI Models:

Artificial Intelligence (AI) Is the Simulation of Human Intelligence Processes by Machines. Using AI in the field of images has been gone beyond retrieval tasks, it could picture categorization, feature extraction, and objects detection at images (Ghaleb et al., 2021).

Machine learning (Krizhevsky et al., 2012; Wang 2015; Wang et al., 2016) and deep learning (Hinton et al., 2012; Shafaey et al., 2018) are two types of AI that have attained the highest levels in images retrieval accuracy. The Deep learning has many techniques that achieved good accuracies in image retrieving and classification such as CNN, LSTM, and GRU would affect the efficiency of CBIR performance (Ghaleb et al, 2022).

CNN (Convolutional Neural Networks) is a deep learning technology that can be used to extract features from images and classify them (LeCun et al., 2015; Jiang 2009). CNN is based on collecting features from the data itself in many layers. (Lecun, Y. 1985), is the one to introduce convolutional neural networks in the 1980s. CNN stands for Convolutional Neural Network, which is a type of neural network designed to handle data in the form of a 2D matrix, such as pictures. CNNs are commonly employed in the detection and categorization of images (Ghaleb, 2022).

(Ghaleb et al, 2021) introduced a CNN model that measures the retrieval accuracy of 10 object images and 10-digit images with 92.9 and 99.8% average accuracy, respectively. (Tan et al, 2020) utilized three distinct Convolutional Neural Network (CNN) models, namely pre-trained AlexNet, fine-tuned AlexNet, and D-Leaf, to extract features. Most machine learning algorithms were used to classify these features: Support Vector Machine (SVM), Artificial Neural Network (ANN), k-Nearest-Neighbour (k-NN), Nave-Bayes (NB), and CNN. (Galab, 2022).

Despite the promised benefits of CNN but it has many issues related to in handling variable-length sequences and context awareness, adding to that, training effective CNNs often requires large amounts of labeled data, which can be expensive and time-consuming to collect (Cao, 2018), that’s leading to Generative Models.

Generative AI refers to "a branch of artificial intelligence that focuses on enabling machines to generate new and original content". Unlike traditional AI systems that follow predefined rules and patterns, generative AI leverages advanced algorithms and neural networks to autonomously produce outputs that mimic human creativity and decision-making (Takyar, 2023).

Generative AI models are designed to learn from large datasets and capture the under lying patterns and structures within the data. These models can then generate new content, such as images, text, music, or even videos, that closely resemble the examples they were trained on. By analyzing the data and understanding its inherent characteristics, generative AI algorithms can generate outputs that exhibit similar patterns, styles, and semantic coherence.

The power of generative AI lies in its ability to go beyond simple replication and mimicry. It can create novel and unique content that hasn’t been explicitly programmed into the system. This opens exciting possibilities for various applications, including art, design, storytelling, virtual reality, and more.

A generative AI model, on the other hand, refers to a specific implementation or architecture designed to perform generative tasks. It is a type of artificial intelligence model that learns from existing data and generates new output that is like the training data it was exposed to.

Generative models have a long history in AI, dating back to the 1950s with the development of Hidden Markov Models (HMMs) (Knill, 1997), and Gaussian Mixture Models (GMMs) (Reynolds, 2009). These models generated sequential data such as speech and time series. However, it wasn’t until the advent of deep learning that generative models saw significant improvements in performance.

According to (Takyar, 2023) Generative AI models are used in various fields, including image generation, text generation, music composition, and more.

Generative AI models skipping retrieving process and their issues to generate new content, simulate human-like behavior, and create stunning visual art by leveraging vast amounts of data and the power of machine learning, by analyzing the data and understanding its inherent characteristics, generative AI algorithms can generate outputs that exhibit similar patterns, styles, and semantic coherence (Takyar, 2023).

There are many key issues in image retrieval that necessitate the use of generative models, Visual mental imaging. Now generative model is crucial for that, and for a variety of cognitive functions, including reasoning, memory, and spatial navigation, where humans visualize stories in their minds by creating mental images when they hear or read them now generative models can drawing these mental images. on other hand, generative models will solve many issues related to complex link between language and the visual world to make human can see what that thinking, or as known as "seeing with the mind's eye," (Kosslyn, 2001).

Building a system that comprehends the connection between language and vision and can produce visuals that accurately represent text descriptions is a significant step towards developing intelligence like to that of humans. This system is inspired by the way people see scenes (Frolov, 2021).

Add to that, The Images in real-world datasets are diverse in terms of objects, scenes, lighting conditions, and viewpoints. Generating images that cover this diverse visual space based on textual descriptions requires generative models capable of capturing this variability.

More that, in applications such as virtual reality or chatbots, there is a need for real-time generation of images from text. Balancing the quality and speed of image generation is a significant challenge in these contexts. That's what Generative AI does, by building images based on textual descriptions, to being a new method in image retrieval systems.

In recent years, Artificial Intelligence Generated Content (AIGC) has acquired widespread attention in many communities outside of the computer science, with the public being interested in the numerous content creation tools developed by prominent tech businesses such as Bard, ChatGpt, and DALL-E- 2 (L. Yunjiu, 2022).

AIGC refers to content that is generated using advanced Generative AI (GAI) techniques, as opposed to being created by human authors, which can create of large amounts of content in a short amount of time. The generation process usually consists of two steps: extracting intent information from human instructions and generating content according to the extracted intentions (Cao, 2023).

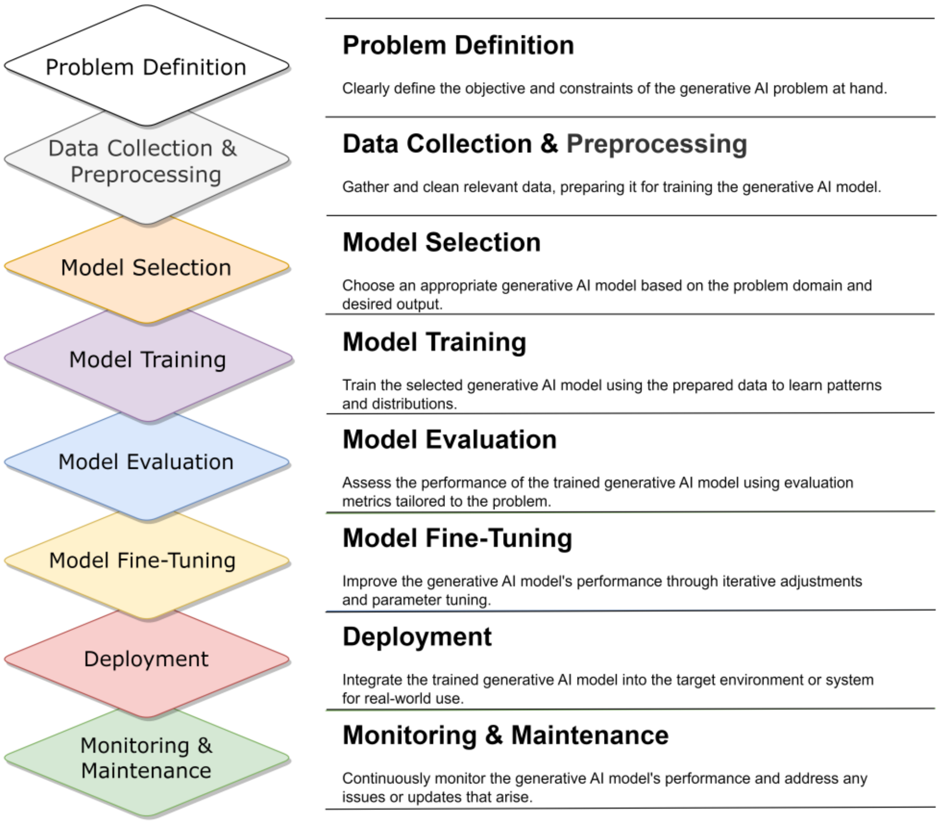
For example, OpenAI created the AIGC model called DALL-E-3, it can produce original, excellent images from textual descriptions in a matter of minutes, many researchers believe it will be the new era of AI and make significant impacts on the image retrieval systems (DALL-E-3, 2023).

When it comes to the components that constitute generative AI models, it's important to note that not all models share the same set of components (Takyar, 2023).

According to (Bandi, 2023) the process of building generative AI models requires various stages that must be addressed in a systematic way to obtain the required results. While the exact terminologies and steps may vary depending on the specific approach and context, as illustrated in Figure 3, the common phases involved in generative AI are:

- Problem Definition: The initial phase revolves around clearly defining the problem that the generative AI model aims to address or generate. This encompasses identifying the desired outcomes, data requirements, and any constraints.
- Data Collection & Preprocessing: The second phase focuses on data collection, which entails gathering a large and representative dataset that encapsulates the patterns and characteristics that the generative model intends to learn using appropriate devices for capturing data, such as web scraping tools, cameras, or sensors.
- Model Selection: Following data collection, the model selection phase begins, where the most suitable generative model architecture is chosen. This phase involves considering popular options such as VAEs, GANs, transformers, or diffusion models.

- Model Training: Once the generative model architecture is determined, the model training phase commences. This stage involves training the selected model using the collected or available dataset. Through this process, the model learns the underlying patterns and statistical relationships within the data. Training generative models frequently necessitates substantial computing resources, particularly for large-scale datasets and sophisticated models. The specific training algorithm varies depending on the chosen model, with GANs, for instance, training a generator network to produce realistic samples while concurrently training a discriminator network to differentiate between real and generated samples.
- Model Evaluating: Once the model training is completed, the subsequent phase involves evaluating and validating its performance. Evaluation metrics are tailored to the specific task or domain. In the case of image generation, metrics such as inception score, Frechet inception distance (FID) or visual inspection can be utilized to assess the quality and diversity of the generated samples.



Problem Definition

Clearly define the objective and constraints of the generative AI problem at hand.

Data Collection & Preprocessing

Gather and clean relevant data, preparing it for training the generative AI model.

Model Selection

Choose an appropriate generative AI model based on the problem domain and desired output.

Model Training

Train the selected generative AI model using the prepared data to learn patterns and distributions.

Model Evaluation

Assess the performance of the trained generative AI model using evaluation metrics tailored to the problem.

Model Fine-Tuning

Improve the generative AI model's performance through iterative adjustments and parameter tuning.

Deployment

Integrate the trained generative AI model into the target environment or system for real-world use.

Monitoring & Maintenance

Continuously monitor the generative AI model's performance and address any issues or updates that arise.

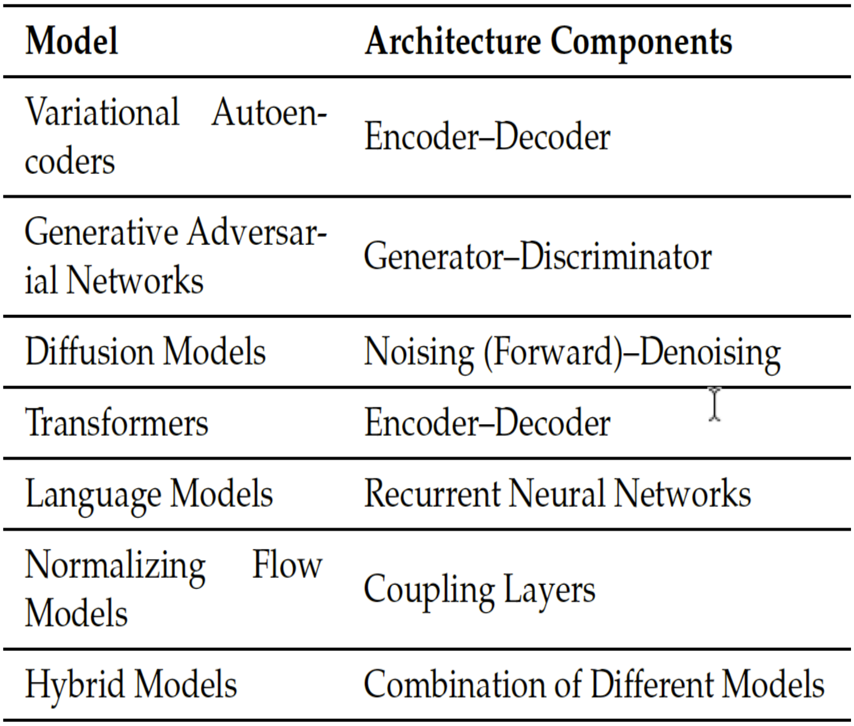
Figure 3: Implementation phases of generative AI (Bandi, 2023).

- Model Fine Tuning: Hyperparameter tuning constitutes a significant aspect of the model training phase. Various hyperparameters, including learning rate, batch size, network architecture, and regularization techniques, influence the behavior and performance of the generative model.
- Refinement: it depends on human specialists, when they polish or improve the generated content. This may entail choosing the best results from the generative AI model or making modest tweaks to ensure the content meets certain criteria or requirements (Takyar, 2023)
- Deployment: Upon successful training and validation, the generative model is ready for deployment to generate new samples.

When it comes to the components that constitute generative AI models, Generative AI Model Architecture plays the main role by includes how its layers or neural networks and components are arranged and organized. The model's architecture determines how it processes and generates information, which makes it a critical aspect of its functionality and suitable for specific tasks. It's important to note that not all models share the same set of components. The specific components of a generative AI model can vary depending on the architecture and purpose of the model. Table 1 describes the architecture components and training methods that are used in the generative AI models (Bandi, 2023).

By understanding these Architectures, researchers and practitioners can choose the most suitable generative model for their specific task or explore hybrid approaches that combine different models to leverage their respective strengths.

The specific components of a generative AI model can vary depending on the architecture and purpose of the model as (Table 1) (Bandi, 2023).



Model	Architecture Components
Variational Autoencoders	Encoder–Decoder
Generative Adversarial Networks	Generator–Discriminator
Diffusion Models	Noising (Forward)–Denoising
Transformers	Encoder–Decoder
Language Models	Recurrent Neural Networks
Normalizing Flow Models	Coupling Layers
Hybrid Models	Combination of Different Models

Table 1. Architecture components used in generative AI models (Bandi, 2023).

Different types of generative AI models may employ various components or variations of them.

It's important to note that the types and design of components in a generative AI model depend on the specific requirements of the generative AI task and the desired output. Different models may prioritize different aspects, such as image generation, text generation, or music composition, leading to variations in the components they employ.

the essential requirements for generative AI can be categorized as (Bandi, 2023):

- Hardware: the collection of data for generative AI tasks involves leveraging cameras, sensors, and existing datasets curated by researchers for images generating. For the training, fine-tuning, and hyperparameter optimization stages, powerful hardware configurations like Tesla V100 16 GB, RTX 2080Ti, NVIDIA RTX 3090 with 24 GB, and TPUs are commonly employed. However, for smaller-scale models, a GTX 1060 6 GB of DDR5 can suffice. Sample generation, which is an integral part of the generative AI process, can be achieved using more basic configurations like a CPU with an i7 3.4 GHz clock speed and a GPU such as the GTX970.

- On the software side, various tools and frameworks play a crucial role in different phases of generative AI. Data collection and preprocessing rely on frameworks like web scraping frameworks, Pandas, NumPy, scikit-image, and RDKit. Additionally, specialized tools for data acquisition, motion capture are employed. To train generative models effectively, deep learning frameworks, such as TensorFlow, PyTorch, scikit-learn, and SciPy, provide comprehensive support for various model architectures and optimization algorithms. These frameworks are also instrumental in evaluating and validating the models. Furthermore, post-processing and model refinement can be facilitate using libraries like OpenCV-Python and NLTK.
- User experience requirements for generative AI models are critical in ensuring user satisfaction and successful outcomes. High-quality and realistic outputs are expected, along with customization and control options to align the generated content with user preferences.

There are many generative AI models, each one has a unique structures, components, and applications. Some common generative AI models are:

- **Variational Autoencoders (VAEs)**: consist of an encoder network, a decoder network, and a latent space. The encoder maps the input data to a latent spacer presentation, while the decoder generates new outputs from the latent space.
- **Generative Adversarial Networks (GANs):** GANs comprise two main components: a generator and a discriminator. The generator generates new samples, such as images, while the discriminator evaluates the generated samples and distinguishes them from real ones. The Study will focus on these types of Models.
- **Transformers Models:** Transformers are widely used in natural language processing tasks. They consist of encoder and decoder layers that enable the model to generate sequences of text or translate between different languages.
- **Autoencoders Models:** Autoencoders consist of an encoder and a decoder. The encoder compresses the input data into a latent representation, and the decoder reconstructs the original data from the latent space.
-

Generative Adversarial Network (GAN):

Generative Adversarial Networks (GAN): it is advanced neural network; it consists of a generator network that creates new instances and a discriminator network that tries to distinguish between the generated instances and real ones. Through an iterative training process, the generator learns to produce increasingly realistic outputs that can deceive the discriminator (Takyar, 2023).

GANs (Goodfellow, 2014) was suggested in 2014, it was a key milestone in the processing and retrieval of images. Based on GAN Model, many proposed models have also been developed for more fine-grained control over the image generation process and the ability to generate high-quality images (Song, Y, 2019).

GAN is a deep learning framework in which two models, a generative model G and a discriminative model D, are trained simultaneously. The objective of G is to capture the distribution of some target data (e.g., distributions of pixel intensity in images). D aids the training of G by examining the data generated by G in reference to "real" data, and thereby helping G learn the distribution that underpins the real data (Wang, S, 2017).

(Mansimov et al. 2015) was the initial work and foundation for modern machine learning techniques for generate images from text synthesis. They demonstrated that the generative model developed by DRAW Gregor et al. (2015) could produce new visual sceneries when it was modified to condition on picture captions (Ramesh, 2021).

GANs are commonly used in image and video generation tasks, where they have shown impressive results in generating realistic images, creating animations, and even generating synthetic human faces. They are also being used in other areas, such as natural language processing, music generation, and fashion design (Takyar, 2023).

GANs have gained immense popularity in image retrieval due to their ability to generate highly realistic images (Radford, A., Metz, L., & Chintala, S, 2015). Researchers have harnessed GANs to generate images that are visually like query images, enabling improved content-based image retrieval.

Generative Adversarial Networks (GANs) address CBIR & SBIR challenges by learning the underlying patterns and distributions in the textual and visual data, these models effectively overcome the semantic gap, handle ambiguity, capture diverse visual content, and incorporate fine-grained details by generating images based on textual descriptions. As a result, they play a crucial role in addressing the complexities of image retrieval based on textual queries, providing a bridge between the language of users and the visual world of images.

GAN training process is depending on updating the generator network to improve its ability to generate realistic samples, while the discriminator network is updated to improve its ability to distinguish between real and generated samples. The training is done in an iterative process, where the generator and discriminator networks are updated alternately to reach a Nash equilibrium (Takyar, 2023).

In another meaning, GAN is based on the minimax two-person zero-sum game, in which one player profits only when the other suffers an equal loss. The two players in GAN are the generator and the discriminator. The generator's purpose is to trick the discriminator, while the discriminator's goal is to identify whether a sample is from a true distribution. The discriminator's output is a probability that the input sample is a true sample. A higher probability suggests that the sample is drawn from real-world data. In contrast, the closer the probability is to zero, the more probable the sample is a fake. When the probability approaches one-half infinity, the optimal answer is reached because the discriminator finds it difficult to check fake samples (Pan, Z, 2023).

Typically, generator (G) and discriminator (D) are implemented using deep neural networks, working as latent function representations. The architecture of the GAN, illustrated in Figure 4 (Bandi, 2023).

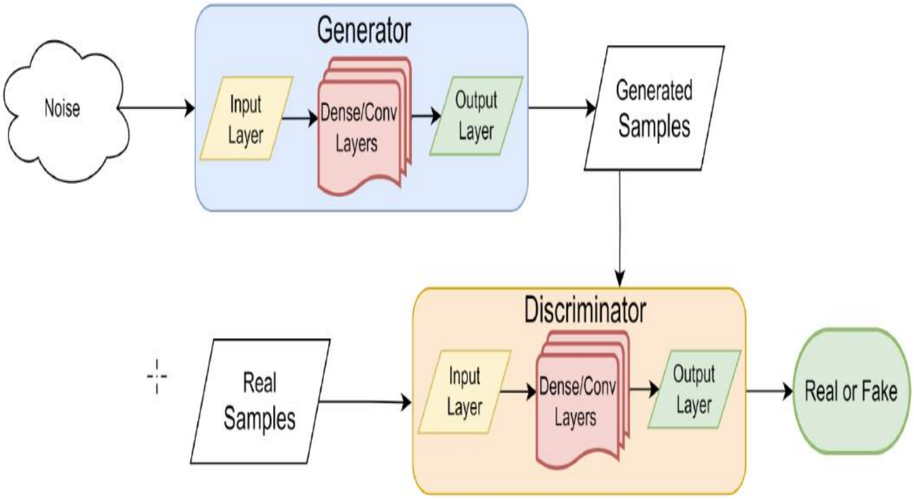


Figure 4: Explains the GAN Architecture (Bandi, 2023).

Involves the G learning the data distribution from real samples and mapping it to a new space (generated samples) using dense/convolutional layers accompanied by its corresponding probability distribution. The primary objective of the GAN is to ensure that this probability distribution closely resembles the distribution of the training samples.

The D receives input data, which can be either real data (x) from the training set or generated data produced by the generator. The discriminator then outputs a probability using dense/convolutional layers or scalar value that indicates whether the input is likely to come from the real data distribution (Bandi, 2023).

Generally, GAN models can be categorized into two types: Unimodal models and Multimodal models. Unimodal models receive instructions from the same modality as the generated content modality, whereas multimodal models accept cross-modal instructions and produce results of different modalities, as shown in Figure 4 (Cao, 2023).

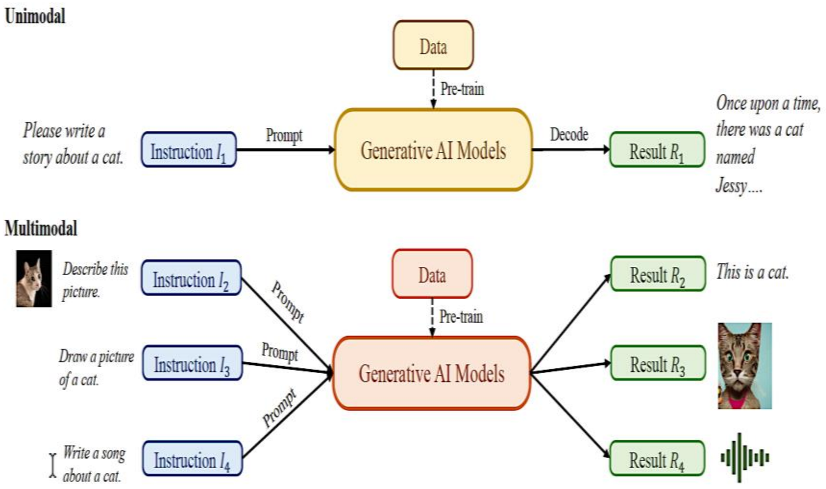


Figure 5: explains the two types of GAN: Unimodal & Multimodal (Cao, 2023).

The Unimodels are designed to accept a specific raw data modality as input, such as text or images, and then generate predictions in the same modality as the input. One of the shortcomings of the Unimodal generation, it based on one type of data, just has ability to generate text from text, it can't allow to geneate images or videos or soud, based on a text query, such as ChatGPT (Cao, 2023).

Multimodal learning, aims to build models that make predictions based on such multimodal information, Multimodal learning is especially important for robots that need to operate properly in the real world, because they need to make sense of the world based on the various types of information they receive through their onboard sensors (Suzuki. M, 2022).

Nowadays, multimodal generation plays a crucial role in images retrieval. Learning the multimodal connection and interaction from data to create a model that creates raw modalities is the aim of multimodal generation (Pan, Z, 2023).

The general structure of multimodal generative vision language can be separate the generation process into encoder part and decoder part, where encoder models will encode the inputs into a latent representation and then the decoder will de-code this representation into a generated output.

The Multimodal (based on the encoder-decoder architecture) is a widely used framework for solving unimodal generation problems in computer vision and natural language processing. In multimodal generation, particularly in vision-language generation, this method is often used as a foundation architecture. The encoder is responsible for learning a contextualized representation of the input data, while the decoder is used to generate raw modalities that reflect cross-modal interactions, structure, and coherence in the representation.

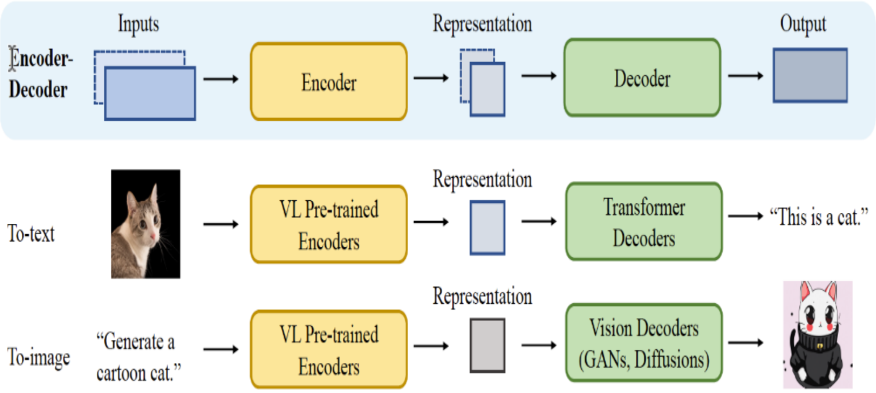


Figure 6:

The general generative multimodal which separates into encoder part and decoder part (Bandi, 2023).

Generative AI in Image Retrieval form Text (Hybrid Models):

GANs is one of the most popular Hybrid Model to image retrieval from text, which involves combining text-based information retrieval techniques with image processing methods. This model is essential in scenarios where users want to search for images using textual queries.

Generative AI hybrid models refer to artificial intelligence systems that combine different generative modeling techniques to produce more sophisticated and versatile outcomes.

GANs can learn generate images that is not just a mere copy of existing images but rather a novel synthesis.

This capability is particularly valuable in situations where traditional images retrieval methods fall short, such as when users seek images that matches their preferences, but the exact keywords are unknown or when searching for images based on visual similarity.

GANS are being harnessed to address several critical challenges in images retrieval. GANs can play a significant role in enhancing various aspects of image retrieval, from data augmentation to generating realistic images for training and improving the overall performance of image retrieval models. GANs can support image retrieval as follow:

- Image Generation: based on the main components of GANs a generator and a discriminator: The generator creates new images, and the discriminator evaluates how realistic those generated images are compared to real ones. GANs has a capability for generating synthetic data for training image retrieval models, especially when the available dataset is limited (solving the CBIR issues).
- Style Transfer and Image Retrieval: GANs, particularly those designed for style transfer, can be used to modify the visual appearance of images while preserving their content. This capability can be useful in image retrieval tasks where the style or appearance of an image is crucial.
- Semantic Understanding: GANs trained for tasks like image-to-image translation or image synthesis can contribute to a better understanding of the semantics of images. This enhanced understanding can be beneficial for improving the performance of image retrieval models.
- Conditional GANs: GANs allow for the generation of images conditioned on specific attributes or classes. This can be applied to create synthetic images for specific retrieval scenarios, enabling the training of models tailored to those scenarios.
- Domain Adaptation: GANs can be used for domain adaptation in image retrieval, helping to bridge the gap between different datasets with varying characteristics. This is especially useful when deploying image retrieval models in real-world scenarios where the target domain might differ from the source domain used for training.

By incorporating text embeddings and conditioning into the GAN architecture, systems can achieve the synthesis of images that are semantically relevant to the input textual descriptions. This text-to-image synthesis capability has applications in various domains, including creative image generation, content creation, and enhancing image retrieval systems.

Generating images from text using GANs involves a process where a conditional GAN (GAN) is trained to understand the relationship between textual descriptions and corresponding images. Here's a step-by-step breakdown of the process:

1- Data Preparation: Gather a dataset containing pairs of textual descriptions and corresponding images. Ensure that the dataset is diverse and representative of the types of images you want to generate.

2- Text Embedding: Convert the textual descriptions into numerical vectors known as text embeddings. Techniques like Word Embeddings (Word2Vec, GloVe) or pre-trained models like BERT can be used to convert words or phrases into numerical representations.

3- Conditional GAN Architecture: by setting a conditional GAN architecture, consisting of a generator and a discriminator, and make the generator takes random noise and the text embedding as input and generates images. The text embedding conditions the generator to produce images based on the provided textual descriptions. The discriminator is also conditioned on the text embedding and evaluates the realism of the generated images in the context of the input text.

4- Training: by Define appropriate loss functions for both the generator and discriminator. During training, the generator aims to produce images that are realistic and align with the input text, while the discriminator aims to correctly classify real and generated images. The training process involves iteratively updating the parameters of the generator and discriminator through backpropagation and optimization techniques.

5- Adversarial Training: The adversarial training process involves a constant interplay between the generator and discriminator. The generator tries to generate images that are indistinguishable from real images, and the discriminator aims to improve its ability to distinguish between real and generated images.

6- Evaluation: Periodically evaluate the performance of the generator using validation data. This can involve assessing the quality of generated images and the alignment between textual descriptions and generated content.

7- Text-to-Image Synthesis: Once the GAN is trained, to generate an image from a new text description, provide the corresponding text embedding to the generator. The generator generates an image based on the given text, leveraging the learned relationships between text embeddings and image generation during training.

8- Fine-tuning (Optional): Depending on the performance and application, fine-tune the GAN or adjust hyperparameters to improve the quality and diversity of generated images.

9- Generate Images: After training, use the generator to generate images from textual descriptions. Provide new text inputs and observe the corresponding generated images.

10- Evaluate and Iterate: Evaluate the quality of generated images using metrics like Inception Score, or user feedback. Iterate on the model and training process to improve results.

11- Deploy the Model: Once satisfied with the performance, deploy the model for generating images based on text in real-world applications.

12- Application: Use the trained GAN for various applications such as creative content generation, image synthesis for missing data, or enhancing image retrieval systems (Reed, S., et al. 2016), (Zhang, H., et al, 2017), (Xu, T., et al. 2018).

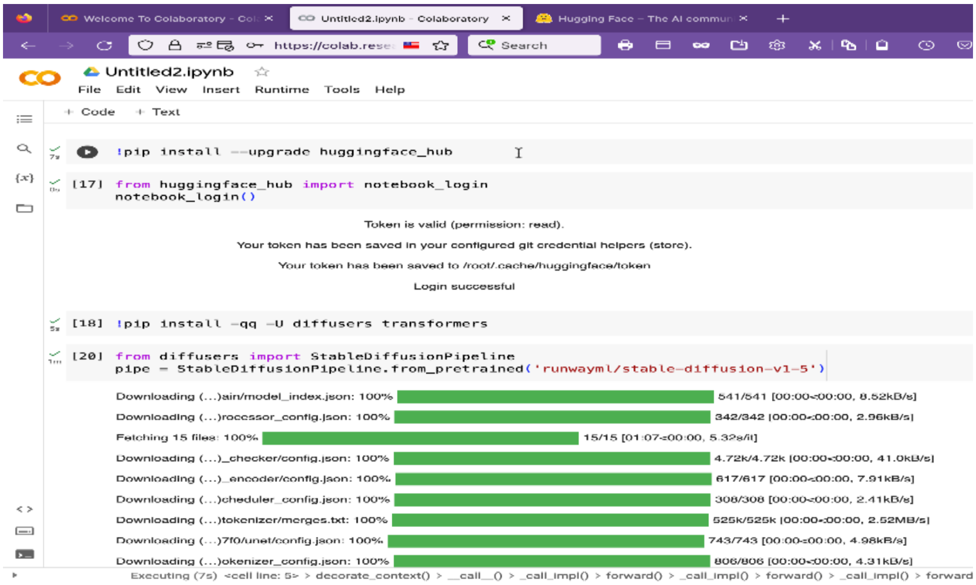
The Model:

Before building Text to Image T2I generative model, the first question should be known ask is whether pre-train model by yourself should or use an existing one. According to (Li, Rebecca, 2023), There are three basic approaches:

- Option 1: Use the API of a commercial Generative model, e.g. dall-e-2 (OpenAI, 2020).
- Option 2: Use an existing open-sourced T2I, e.g. (AttnGAN: Fine-Grained Text to Image Generation with Attentional GANs), or (StackGAN: Text to Photo-realistic Image Synthesis with Stacked GANs).
- Option 3: Pre-train GAN Generative model by yourself or with consultants. There are a lot of details to consider when making the choice (Li, Rebecca, 2023).

The proposed T2I model leverages the power of deep learning techniques, combining natural language processing (NLP) and computer vision (CV) methodologies to bridge the semantic gap between textual queries and visual representations. the research employs state-of-the-art neural architectures, integrating advanced recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to map textual descriptions to corresponding image features then, training GAN model on large-scale, diverse datasets, enabling it to learn relationships between textual cues and visual elements.

The main prerequisites of this model are installing Hugging Face python library and authenticate using an API token as Figure 7.


Welcome To Colaboratory - Col
Untitled2.ipynb - Colaboratory
Hugging Face – The AI commun
https://colab.rese
Search
Untitled2.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
+ Code
+ Text

```
!pip install --upgrade huggingface_hub
```

```
[17] from huggingface_hub import notebook_login
     notebook_login()
```

```
Token is valid (permission: read).
Your token has been saved in your configured git credential helpers (store).
Your token has been saved to /root/.cache/huggingface/token
Login successful
```

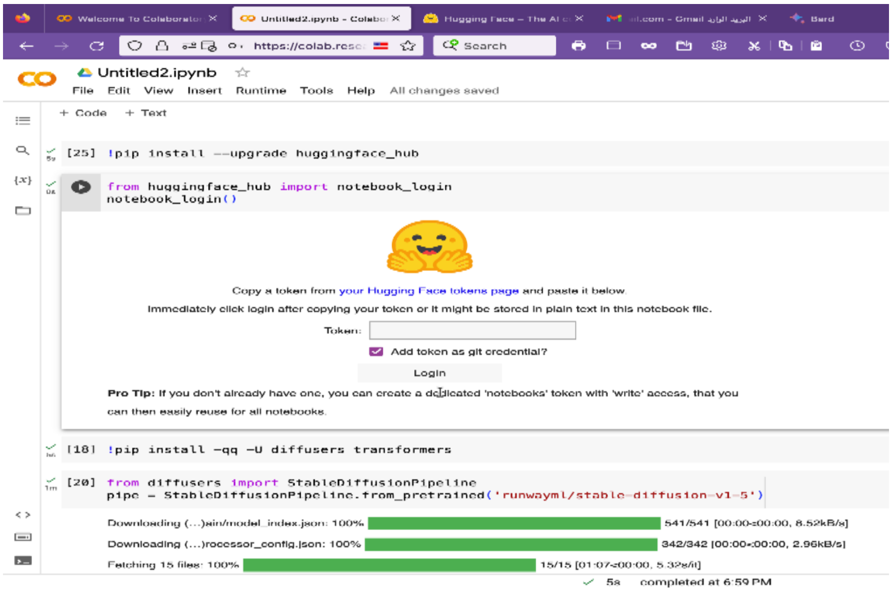
```
[18] !pip install -qq -U diffusers transformers
```

```
[20] from diffusers import StableDiffusionPipeline
     pipe = StableDiffusionPipeline.from_pretrained('runwayml/stable-diffusion-v1-5')
```

```
Downloading (…)ain/model_index.json: 100%      541/541 [00:00<00:00, 8.52kB/s]
Downloading (…)rocessor_config.json: 100%      342/342 [00:00<00:00, 2.96kB/s]
Fetching 15 files: 100%      15/15 [01:07<00:00, 5.32s/it]
Downloading (…)_checker/config.json: 100%      4.72k/4.72k [00:00<00:00, 41.0kB/s]
Downloading (…)_encoder/config.json: 100%      617/617 [00:00<00:00, 7.91kB/s]
Downloading (…)cheduler_config.json: 100%      308/308 [00:00<00:00, 2.41kB/s]
Downloading (…)tokenizer/merges.txt: 100%      525k/525k [00:00<00:00, 2.52MB/s]
Downloading (…)7f0/unet/config.json: 100%      743/743 [00:00<00:00, 4.98kB/s]
Downloading (…)okenizer_config.json: 100%      806/806 [00:00<00:00, 4.31kB/s]
Executing (7s) <cell line: 5> > decorate_context() > __call__() > _call_impl() > forward() > _call_impl() > forward() > _call_impl() > forward
```

Figure 7: Explain importing python libraries to install it.

After successfully importing Hugging Face Library, the study going to download the diffusers and transformers python libraries, as Figure 8.


Welcome To Colaborator
Untitled2.ipynb - Colabor
Hugging Face – The AI c
il.com - Gmail البريد الوارد
Bard
https://colab.rese
Search
Untitled2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text

```
[25] !pip install --upgrade huggingface_hub
```

```
from huggingface_hub import notebook_login
notebook_login()
```

Copy a token from your Hugging Face tokens page and paste it below.

Immediately click login after copying your token or it might be stored in plain text in this notebook file.

Token:

☑ Add token as git credential?

Login

Pro Tip: If you don't already have one, you can create a dedicated 'notebooks' token with 'write' access, that you can then easily reuse for all notebooks.

```
[18] !pip install -qq -U diffusers transformers
```

```
[20] from diffusers import StableDiffusionPipeline
pipe = StableDiffusionPipeline.from_pretrained('runwayml/stable-diffusion-v1-5')
```

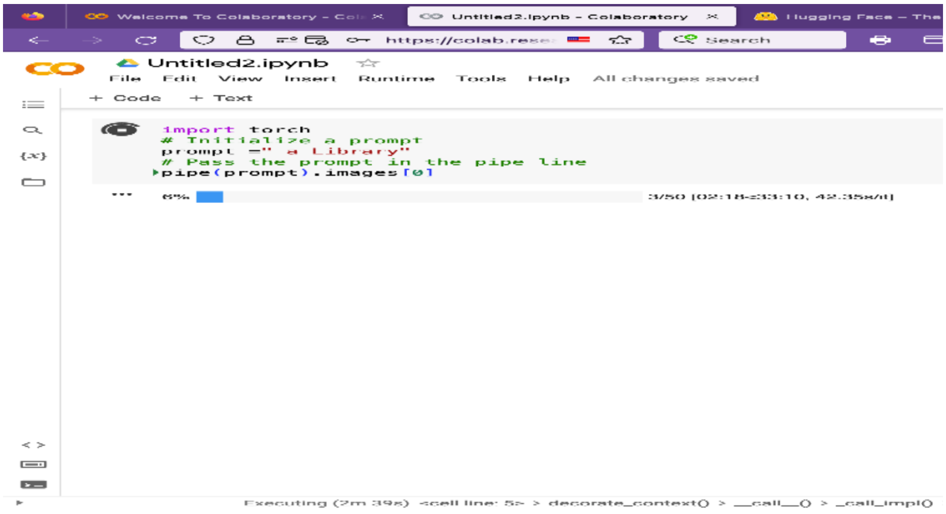
```
Downloading (…)ain/model_index.json: 100%      541/541 [00:00<00:00, 8.52kB/s]
Downloading (…)rocessor_config.json: 100%      342/342 [00:00<00:00, 2.96kB/s]
Fetching 15 files: 100%      15/15 [01:07<00:00, 5.32s/it]
```

✓ 5s completed at 6:59 PM

Figure 8: Explain download the diffusers and transformers python libraries.

The study needs to create a stable diffusion model pipeline so we can basically pass the model some text and have it generate an image based on that prompt. notice that one of the parameters we're passing is a path to a Stable Diffusion model hosted on Hugging Face. The examples in this model were tested using (v2.1) at the time of study writing.

Before the study going to build images based on this model GANs with Diffusion transformers, it was necessary build a prompt box to receive a user text queries, then writing the any text query, when model receive the text query, it will be starting to analysis the text to generate images that matches query, as Figure 9.



Welcome To Colaboratory - Col
Untitled2.ipynb - Colaboratory
Hugging Face – The
https://colab.rese
Search
Untitled2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text

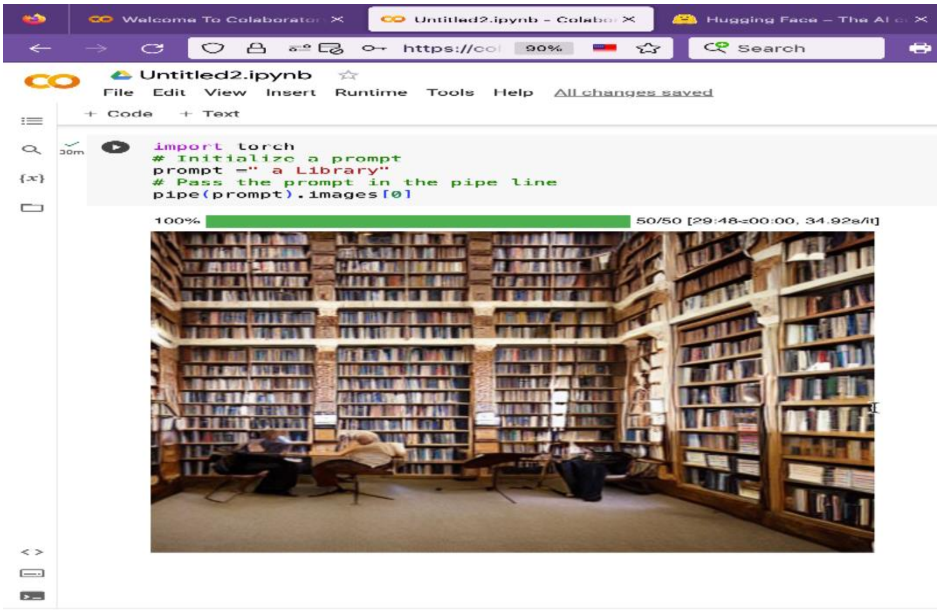
```
import torch
# Initialize a prompt
prompt =" a Library"
# Pass the prompt in the pipe line
pipe(prompt).images[0]
```

```
6%  3/50 [02:18<33:10, 42.35s/it]
```

```
Executing (2m 39s) <cell line: 5> > decorate_context() > __call__() > _call_impl()
```

Figure 9: Explain Model receiving a text query and start to generate image for "Library".

After model complete generate image e.g., Library image, it will be displayed on notebook as Figure 10.



```
import torch
# Initialize a prompt
prompt =" a Library"
# Pass the prompt in the pipe line
pipe(prompt).images[0]
```

```
100%  50/50 [29:48<00:00, 34.92s/it]
```

Figure 10: Explain the image "Library" after generating GAN Model to it.

Conclusion:

In conclusion, the integration of generative models into image retrieval has opened new avenues for image search, exploration, and retrieval. This field is continuing to push the boundaries of generative model capabilities, seeking ways to enhance the accuracy, scalability, and ethical considerations of image retrieval systems. As the field evolves, the fusion of generative models with other cutting-edge technologies like deep learning and natural language processing is likely to redefine the future of image retrieval.

This study introduced GANs as an innovative mechanism within the T2I model, enabling it to focus on specific parts of the input text and image during the generation process. This mechanism enhances the model's interpretability and generates more precise results, aligning with the user's intent. the most prominent results have been demonstrated by the extensive evaluation was the model's effectiveness and efficiency in generating accurate and contextually relevant images corresponding to given textual queries.

The research's findings have significant influences for several applications, including image retrieval based on text queries, and creative content generation, and virtual/augmented reality experiences. As the digital landscape continues to evolve, the developed T2I model stands at the forefront, offering an intelligent and scalable solution for bridging the gap between textual queries and image retrieval, thereby revolutionizing the way users interact with visual data.

It's important to note that training GANs can be challenging, and hyperparameter tuning, careful loss function design, and proper dataset preprocessing are crucial for achieving desirable results.

Currently, no standardized framework exists to merge between generative models, often leading to fragmented efforts and suboptimal results.

As generative models continue to evolve and improve, they hold the potential to revolutionize how we access and interact with information.

References:

1. A.W.M. Smeets, A. Hamdi, R.C. Veltkamp, M. Worring. (2006). "Content-based image retrieval at low bit-rates," in IEEE Transactions on Multimedia, vol. 8, no. 5, pp. 791-803.
2. Abioui, Hasna & Idarrou, Ali & Ali, Bouzit & Mammass, D. (2019). Review: Automatic Image Annotation for Semantic Image Retrieval. Available at: https://www.researchgate.net/publication/337533787_Review_Automatic_Image_Annotation_for_Semantic_Image_Retrieval
3. Agrawal, D., Agarwal, A., & Sharma, D. K. (2022). Content-Based Image Retrieval (CBIR): A Review. Recent Innovations in Computing: Proceedings of ICRIC 2021, Volume 2, 439-452.
4. Ajay, K.D.K., Malleswara Rao, V. (2022). ‘Recent Techniques in Image Retrieval: A Comprehensive Survey”. In “Soft Computing and Signal Processing. ICSCSP 2021. Advances in Intelligent Systems and Computing, vol 1413. Springer, Singapore. https://doi.org/10.1007/978-981-16-7088-6_41
5. Alzu’bi, A., Amira, A., & Ramzan, N. (2015). Semantic content-based image retrieval: A comprehensive study. Journal of Visual Communication and Image Representation, 32, 20-54.
6. Bandi, A.; Adapa, P.V.S.R.; Kuchi, Y.E.V.P.K. (2023). The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges. Future Internet. https://doi.org/10.3390/fi15080260
7. Barz, B., & Denzler, J. (2021). Content-based image retrieval and the semantic gap in the deep learning era. In Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II (pp. 245-260). Springer International Publishing.
8. Burger W, Burge MJ (2009). “Principals of digital image processing: core algorithms”. Springer.
9. Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). “A comprehensive survey of ai-generated content (AIGC): A history of generative ai from GAN to ChatGPT”. arXiv preprint: arXiv:2303.04226.
10. Dahake, P. A., & Thakare, S. S. (2018). “Content based image retrieval: A review. Int. Res. J. Eng. Technol, 5, 1059-1061.
11. Elad, M., & Milanfar, P. (2019). “A Guided Tour of Image Super-Resolution”. IEEE Transactions on Image Processing, 30, 2341-2358
12. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... & Olah, C. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread, 1.
13. Enser, P. (2000). Visual image retrieval: seeking the alliance of concept-based and content-based paradigms. J. Inf. Sci. 26(4), 199–210.
14. Farooque, M. (2003). Image indexing and retrieval. In DRTC Workshop on Digital Libraries: Theory and Practice March DRTC.
15. Frolov, S., Hinz, T., Raue, F., Hees, J., & Dengel, A. (2021). Adversarial text-to-image synthesis: A review. Neural Networks, 144, 187-209.

16.Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., & Mikolov, T. (2013). “DeViSE: A Deep Visual-Semantic Embedding Model”. In Advances in neural information processing systems (NeurIPS), 2013.

17.Ghaleb, M. S., Ebied, H. M., Shedeed, H. A., & Tolba, M. F. (2021). “Content based image retrieval based on convolutional neural network”. 10th International Conference on Intelligent Computing and Information science (ICICS), 149-153, Cairo, Egypt.

18.Ghaleb, M. S., Ebied, H. M., Shedeed, H. A., & Tolba, M. F. (2022). “Image Retrieval based on deep learning”. J. Syst. Manag. Sci, 12(2), 477-496.

19.Ghazvininejad, M., Shi, Y., Creswell, A., Saatchi, Y., Goldman, J., & Lowe, D. G. (2019). “Hugging Face's Transformers: State-of-the-art Natural Language Processing”. arXiv preprint arXiv:1910.03771.

20.Gong Y, Zhang H, Chuan HC, Sakauchi M (1994). “An image database system with content capturing and fast image indexing abilities”. In: 1994 Proceedings of IEEE international conference on multimedia computing and systems, pp 121–130Return to ref 11 in article

21.Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). “Generative adversarial nets”. Advances in neural information processing systems, 27.

22.Gregor, K., Danihelka, I., Graves, A., Rezende, D., & Wierstra, D. (2015, June). “Draw: A recurrent neural network for image generation”. In International conference on machine learning (pp. 1462-1471). PMLR.

23.Gregor, K., et. (2015). “A recurrent neural network for image generation”. In International Conference on Machine Learning, pp. 1462–1471. PMLR, 2015.

24.Gudivada VN, Raghavan VV. (1995). “Design and evaluation of algorithms for image retrieval by spatial similarity”. ACM Trans Inf Syst 13(2):115–144.

25.Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). “Deep neural networks for acoustic modeling in speech recognition”. The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.

26.Hossein Pourghassem and Hassan Ghassemian. (2008) “Content-based medical image classification using a new hierarchical merging scheme. Computerized Medical Imaging and Graphics, 32(8):651–661.

27.Huang PW, Dai SK. (2003). “Image retrieval by texture similarity. Pattern Recognition” 36(3):665–679Return to ref 16 in article

28.Jagtap, J., & Bhosle, N. (2021). “A comprehensive survey on the reduction of the semantic gap in content-based image retrieval”. International Journal of Applied Pattern Recognition, 6(3), 254-271.

29.Jing Li, Nigel Allinson, Dacheng Tao, and Xuelong Li. 2006. “Multi-training support vector machine for image retrieval”. IEEE Transactions on Image Processing, 15(11):3597–3601.

30.Kingma, D. P., & Welling, M. (2013). “Auto-Encoding Variational Bayes”. arXiv preprint arXiv:1312.6114.

31.Knill, K., & Young, S. (1997). “Hidden Markov models in speech and language processing. In Corpus-based methods in language and speech processing” (pp. 27-68). Dordrecht: Springer Netherlands.

32.Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). “Neural foundations of imagery’. Nature reviews neuroscience, 2(9), 635-642.

33.Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ‘Image-net classification with deep convolutional neural networks”. In Advances in neural information processing systems, 1097-1105.

34.Huang, K., Li, S., Kang, X., & Fang, L. (2016). Spectral–spatial hyperspectral image classification based on KNN. Sensing and Imaging, 17, 1-13.

35.L. Yunjiu, W. Wei, and Y. Zheng. (2022). “Artificial intelligence-generated and human expert-designed vocabulary tests: A comparative study,” SAGE Open, vol. 12, no. 1, p. 21582440221082130.

36.Lecun, Y. (1985). “Une Procedure dśapprentissage pour reseau a seuil assymetrique cog'nitiva 85: A la Frontiere de lqIntelligence Artificielle des Sciences de la Connais' sance des Neurosciences. Paris, France.

37.Li, Rebecca. (2023). “Current Best Practices for Training LLMs from Scratch” available at: www.wandb.ai

38.M. Datar, P. M. Hayes, P. Singla, P. N. Yianilos, (2007). "A scalable end-to-end system for efficient image retrieval," in Proceedings of the ACM SIGIR conference on Research and development in information retrieval, pp. 341-350.

39.Mansimov, E., Parisotto, E., Ba, J. (2015). ‘Generating images from captions with attention”. https://arxiv.org/abs/1511.02793

40.Merriam-Webster. (2023). Image definition & meaning. Merriam-Webster. https://www.merriamwebster.com/dictionary/image

41.OpenAI. (2023). ‘DALL-E 3”. https://openai.com/dall-e-3

42.Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). ‘Training language models to follow instructions with human feedback”. Advances in Neural Information Processing Systems, 35, 27730-27744.

43.Pan, Z., Yu, W., Yi, X., Khan, A., Yuan, F., & Zheng, Y. (2019). “Recent progress on generative adversarial networks (GANs): A survey”. IEEE access, 7, 36322-36333.

44.Pourghassem, H. H. Ghassemian. (2008). Content-based medical image classification using a new hierarchical merging scheme. Computerized Medical Imaging and Graphics, 32(8):651–661.

45.Pradhan, J., Pal, A. K., & Banka, H. (2019). Principal texture direction-based block level image reordering and use of color edge features for application of object-based image retrieval. Multimedia Tools and Applications, 78, 1685-1717.

46.Pradhan, J., Pal, A.K., Banka, H. (2021). Medical Image Retrieval System Using Deep Learning Techniques. In: Elloumi, M. (eds) “Deep Learning for Biomedical Data Analysis”. Springer, Cham. https://doi.org/10.1007/978-3-030-71676-9_5

47.Priyatharshini, R., Chitrakala, S. (2013). “Association Based Image Retrieval: A Survey”. In: Das, V.V., Chaba, Y. (eds) Mobile Communication and Power Engineering. AIM 2012.

Communications in Computer and Information Science, vol 296. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35864-7_3

48. Radford, A., Metz, L., & Chintala, S. (2015). “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks’. arXiv preprint arXiv:1511.06434.

49. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A. & Sutskever, I. (2021). “Zero-shot text-to-image generation”. In International Conference on Machine Learning (pp. 8821-8831). PMLR.

50. Ramya, V. (2018). ‘Content based image retrieval system using clustering with combined patterns”. International Journal of Scientific Research in Computer Science, Engineering, and Information Technology, IJSRCSEIT, ISSN, 2456-3307.

51. Rani, R. U. (2020). “Image qualification learning technique through content-based image retrieval”. Int. J. Future Revolut. Comput. Sci. Commun. Eng, 4(1).

52. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016, June). Generative adversarial text to image synthesis. In International conference on machine learning (pp. 1060-1069). PMLR.

53. Reynolds, D. A. (2009). “Gaussian mixture models”. Encyclopedia of biometrics, 741(659-663).

54. Rui Y, Huang TS, Chang SF (1999). “Image retrieval: current techniques, promising directions, and open issues”. J Vis Commun Image Represent 10(1):39–62Return to ref 46 in article

55. S. Rubini, R. Divya, G. Divyalakshmi, T.M.S. (2018). “Ganesan, Content based image retrieval (CBIR)”. Int. Res. J. Eng. Technol. (IRJET) 05(03).

56. Sakhare, S.V., Nasre, V.G. (2011). "Design of feature extraction in content based imageretrieval (CBIR) using color and texture". Int. J. Comput. Sci. Inform. 1(II) 8.

57. Sean, Xiang Zhou, Huang, T S. (2002). “Unifying keywords and visual contents in image retrieval”. IEEE Multimedia, 9(2):23–33.

58. Silva, Loshadi & Premaratne, Saminda. (2013). “Content Based Image Retrieval System”. IEEE - International Conference on Research and Development Prospectus on Engineering and Technology (ICRDPET 2013).

59. Song, Y., & Ermon, S. (2019). “Generative modeling by estimating gradients of the data distribution”. Advances in neural information processing systems, 32.

60. Suzuki, M., & Matsuo, Y. (2022). “A survey of multimodal deep generative models”. Advanced Robotics, 36(5-6), 261-278.

61. T. Khalil, M. U. Akram, H. Raja. (2018). "Detection of Glaucoma Using Cup to Disc Ratio from Spectral Domain Optical Coherence Tomography Images," in IEEE Access, vol. 6, pp. 4560-4576, doi: 10.1109/ACCESS.2018.2791427.

62. Tadasare, S. S., & Pawar, S. S. (2018). “Content based retinal image retrieval using lifting wavelet transform for classification of retinal fundus images”. Int. J. Elect. Electron. Comput. Sci. Eng, 5(1), 169-176.

63. Takyar, Akash. (2023). ‘Understanding generative AI models: A comprehensive overview”. Available at: https://www.leewayhertz.com/generative-ai-models/

64. Tamura, H., Mori, S., Yamawaki, T. (1978). “Textural features corresponding to visual perception. IEEE Trans. Syst. Man Cybern. 8(6), 460–473.

65. Thamotharan, Dharani & Aroquiaraj, Laurence. (2013). “A survey on content-based image retrieval”. Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, Prime 2013. 485-490. DOI: 10.1109/ICPRIME.2013.6496719.

66. Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). “Deep learning for identifying metastatic breast cancer”. ArXiv preprint arXiv: 1606.05718.

67. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, Y. (2014). “Learning Fine-grained Image Similarity with Deep Ranking”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

68. Wang, S. (2017). “Generative Adversarial Networks (GAN): A Gentle Introduction”. Tutorial on GAN in LIN395C: Research in Computational Linguistics, 1.

69. Wang, X. Y., Yang, H. Y., & Li, D. M. (2013). “A new content-based image retrieval technique using color and texture information”. Computers & Electrical Engineering, 39(3), 746-761.

70. Westerveld, T. (2000). "Image retrieval: content versus context". In: Content-Based Multi-media Information Access, vol. 1, pp. 276–284.

71. Xiao-Feng Wang, De-Shuang Huang, Ji-Xiang Du, Huan Xu, and Laurent Heutte. (2008). “Classification of plant leaf images with complicated background”. Applied mathematics and computation, 205(2):916–926.

72. Xu, H., Saenko, K., & Gould, S. (2015). “Jointly Modeling Deep Video and Compositional Text to Bridge Vision and Language in a Unified Framework”. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

73. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1316-1324).

74. Y. Rui, A. K. Jain, (2003). "Similarity retrieval from large image databases: Developments and trends," in The Handbook of Multimedia Information Retrieval, pp. 29-68.

75. Yiqing Guo, Xiuping Jia, David Paull. (2018). Effective sequential classifier training for svm-based multitemporal remote sensing image classification. IEEE Transactions on Image Processing, 27(6):3036–3048.

76. Zarchi, M. S., Monadjemi, A., & Jamshidi, K. (2014). A semantic model for general purpose content-based image retrieval systems. Computers & Electrical Engineering, 40(7), 2062-2071.

77. Zhang, D., Islam, M. M., & Lu, G. (2012). A review on automatic image annotation techniques. Pattern Recognition, 45(1), 346-362. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0031320311002391

78. Zhang, D., Lu, G. (2004). Review of shape representation and description techniques. Pattern Recognition. 37(1), 1–19.

79. Zhang, H. (2018). Realistic image synthesis with stacked generative adversarial networks. IEEE transactions on pattern analysis and machine intelligence, 41(8):1947–1962, x.

80. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 5907-5915).

81. Zhang, H.,et. (2017). Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision, pp. 5907–5915, 2017.