|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Academic Year: | 2022/2023 | Term: | Spring 2023 |  |
| Course Code: | ELC 4028 | Course Title: | Artificial Neural Networks and its Applications |

Cairo University

Faculty of Engineering

Electronics and Communications Engineering Department – 4th Year

Neural Networks Applications

- Text-to-Image synthesis using Stable Diffusion -

*Submitted to: Dr. Mona Riad*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | BN | Sec | ID | رقم الجلوس |
| احمد محمود حسيني عطية | 22 | 1 | 9180178 | 34022 |
| علي ماهر عبدالسلام نبيه | 5 | 3 | 9190067 | 34117 |
| محمد احمد طه السيد | 30 | 3 | 9191043 | 34142 |
| محمد حسام عثمان يسن | 35 | 3 | 9191083 | 34147 |
| محمد عاطف ربيع | 43 | 3 | 9190924 | 34155 |

Table of Contents

[Abstract 1](#_Toc132153691)

[1. Introduction 2](#_Toc132153692)

[1.1. Motivation for Stable Diffusion 3](#_Toc132153693)

[2. Overview of Stable Diffusion Main Components 3](#_Toc132153694)

[2.1. Text Encoder 3](#_Toc132153695)

[2.2. Image Information Creator 4](#_Toc132153696)

[2.2.1. Diffusion 4](#_Toc132153697)

[2.3. Image Decoder 4](#_Toc132153698)

[References 6](#_Toc132153699)

**List of Figures**

Figure 1: text to image examples by Lexica 2

Figure 2: image to image block diagram example 3

Figure 3: Stable Diffusion Main Components 4

Figure 4: diffusion example from a noise pattern then slowly morphs into a desired output 4

Figure 5: full cycle of SD 5

**List of Abbreviations**

# Abstract

Text to image synthesis is a process of generating an image from a given text description. It is a form of natural language processing that involves the use of deep learning algorithms to generate an image from a text description. The goal of text to image synthesis is to create an image that accurately reflects the content of the text. In this report, we will discuss a paper called “High-Resolution Image Synthesis with Latent Diffusion Models” [1] which uses Stable Diffusion (SD) also called Latent Diffusion Model (LDM) to achieve new state-of-the-art scores for image inpainting and class-conditional image synthesis and highly competitive performance on various tasks, including text-to-image synthesis, unconditional image generation and super-resolution, while signiﬁcantly reducing computational requirements compared to the previous pixel-based Diffusion Models (DM).

# 1. Introduction

Artificial Intelligence (AI) art has become extremely popular in recent years as technology advances and more people became interested in exploring the possibilities of using AI to create artwork. The Public release of SD in August 22, 2022 [2] has massively impacted this field as it’s not only a high-performance model competitive with present AI image generation models such as DALL-E by OpenAI or Imagen by Google but also its model weights and source code are fully open to anyone which allows anyone to download the model and tinker with it and adjust the internal parameters in a way that they can’t do with the closed solutions as DALL-E and Imagen. [3]. In addition, as we will discuss in this report, the latent space approach allowed relatively low resources and memory requirements [1].

the most common way of generation AI art from text is by typing a prompt to the model and it will generate the image, The best platform for finding examples and the prompts used to generate images is Lexica [4], which archives over 10 million sample artworks. Each artwork includes its full prompt and the seed number, some examples of images and their prompt are shown below

A realistic cute adorable baby owl made of crystal ball with low poly eye's surrounded by glowing aura highly detailed intricated concept art with vivid beautiful colors trending artstation 8k

A cute wizard elven, blue long hair, golden eyes, in a forest, simple dress, intricate, elegant, highly detailed, digital painting, hyper realistic, fantasy, dungeons and dragons, art by artgerm and greg rutkowski



streets of 1980s pripyat with neon lights in a foggy rainy night



Shabby Chic Teal and Pink Flowers with Fluffy Realistic Rabbit

Figure 1: text to image examples by Lexica

In addition, painting images from text isn’t the only feature of SD but also image to image and object removal is also possible. Also, various SD examples can be found in twitter social platform with the hashtag #StableDiffusion.

## 1.1. Motivation for Stable Diffusion

Before we dive into the model, we have to know the disadvantages of previous models.

DMs belong to the class of likelihood-based models, whose mode-covering behavior makes them prone to spend excessive amounts of capacity (and thus compute resources) on modeling imperceptible details of the data. Leading to 2 consequences for the research community and users in general which are: [1]

1. Training such a model requires massive computational resources only available to a small fraction of the ﬁeld. [1]
2. Evaluating an already trained model is also expensive in time and memory, since the same model architecture must run sequentially for a large number of steps. [1]

Hence, for these reasons a method is needed to increase the accessibility of this powerful model class and reduce its significant resource consumption without impairing their performance. [1]

They decided to work completely in the image information space (or latent space) which makes it faster than previous DMs that worked in pixel space. [1]

# 2. Overview of Stable Diffusion Main Components

The operation of SD is either text-to-image or can take both as an input to help it figure out the output image [5]



Figure 2: image to image block diagram example

This SD block consists of several components and models which we will look into.

## 2.1. Text Encoder

It is a text understanding component that translates the text prompt into a numeric representation so that the image generator can process. It is a special Transformer language model; It takes the input text and outputs a list of numbers representing each word/token in the text. [5]

## 2.2. Image Information Creator

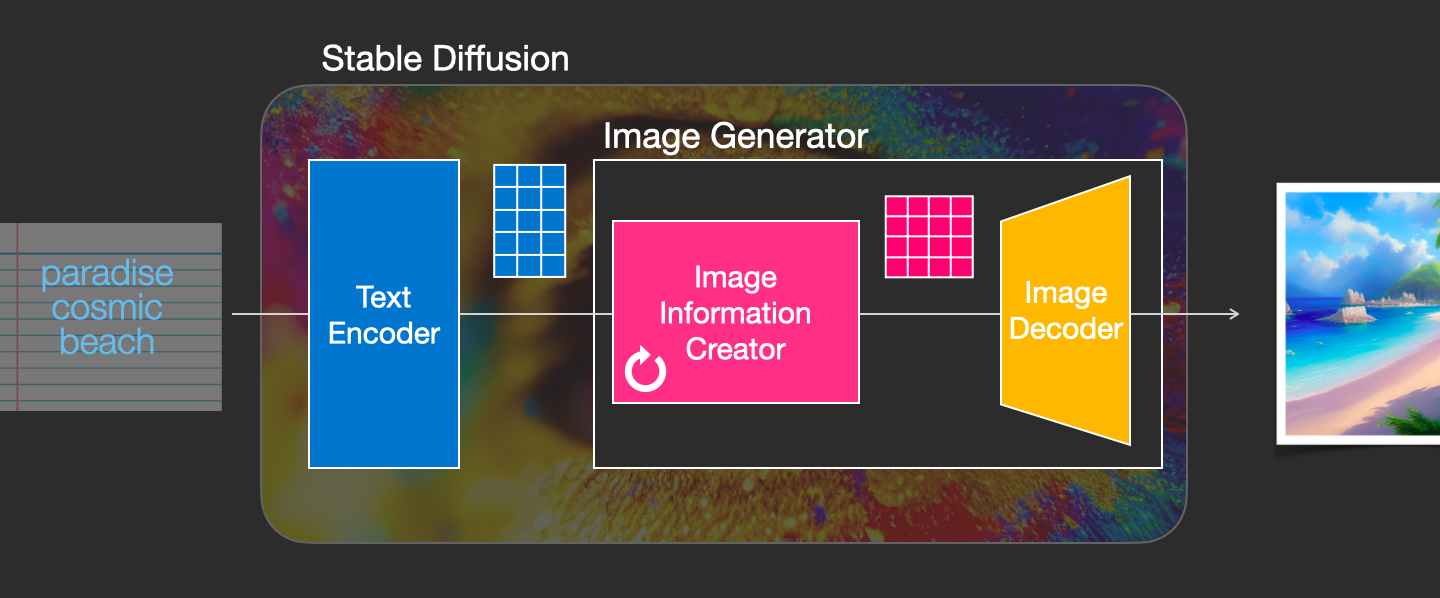


Figure 3: Stable Diffusion Main Components

Basically, what this component does from its name is generation image information from the data received from text encoder. It’s where a lot of the performance gain over previous models is achieved. This component runs for multiple steps to generate image information. This is the steps parameter in SD interfaces and libraries which often defaults to 50 or 100. This component is made up of a UNet-NN and a scheduling algorithm. [5]

### 2.2.1. Diffusion

The term diffusion means the step-by-step processing of information, from just a noise shaped pattern to a high-quality image with a meaning. [3]

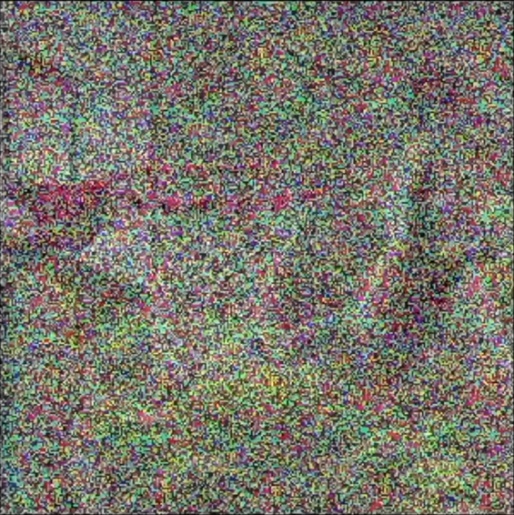


Figure 4: diffusion example from a noise pattern then slowly morphs into a desired output

## 2.3. Image Decoder

At the end of the process, it runs only once to produce the final pixel image. [5]

Now the three main components of SD are mentioned

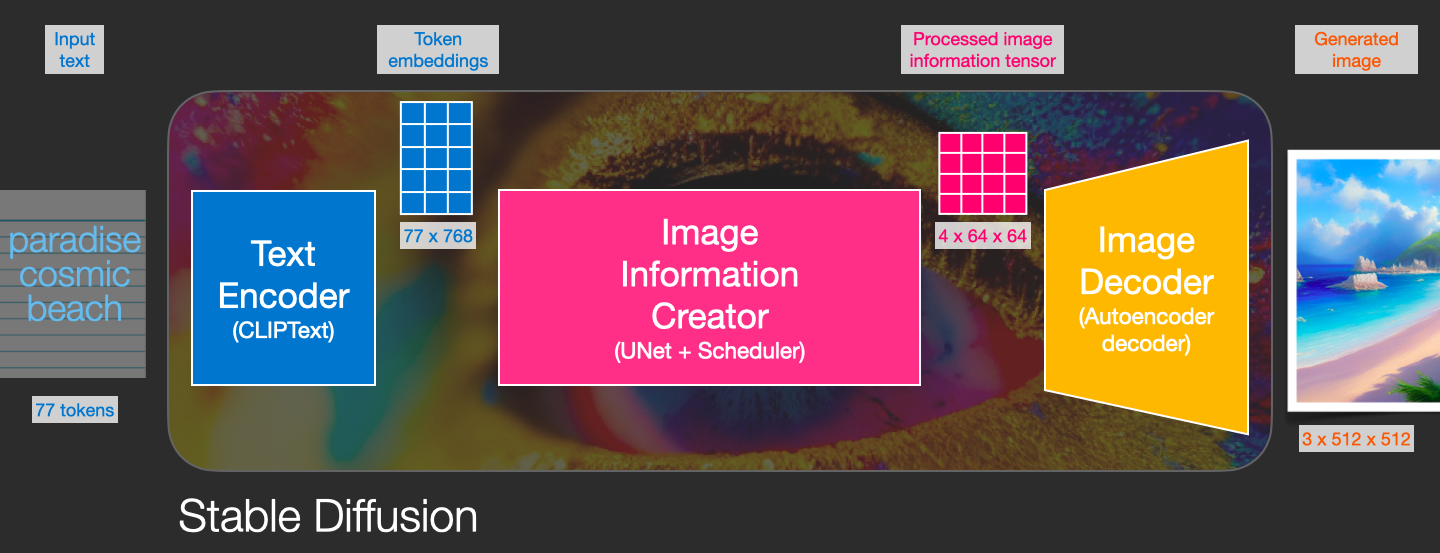


Figure 5: full cycle of SD

Each component consists of its own neural network model

Table 1: summary of Stable Diffusion main components

|  |  |  |  |
| --- | --- | --- | --- |
|  | ClipText | UNet + Scheduler | Autoencoder Decoder |
| Description | For text encoding | Gradually process the info in the latent space. | paints the final image using the processed info array. |
| Input | Text prompt | Text array made up of noise patterns | The processed info |
| Output | 77 token embeddings vectors, each in 768 dimensions. [5] | A processed info array | The resulting image (dimensions: (3, 512, 512) which are (RGB, width, height)) [5] |

# References

|  |  |
| --- | --- |
| [1] | R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models | Papers With Code," 2022. [Online]. Available: https://paperswithcode.com/paper/high-resolution-image-synthesis-with-latent. [Accessed 2023]. |
| [2] | Stability AI, "Stable Diffusion Public Release — Stability AI," [Online]. Available: https://stability.ai/blog/stable-diffusion-public-release. [Accessed 2023]. |
| [3] | Two Minute Papers, "Stable Diffusion: DALL-E 2 For Free, For Everyone! - YouTube," 6 Septemper 2022. [Online]. Available: https://youtu.be/nVhmFski3vg. [Accessed 2023]. |
| [4] | "Lexica," [Online]. Available: https://lexica.art/. [Accessed 2023]. |
| [5] | J. Alammar, "The Illustrated Stable Diffusion," 2022. [Online]. Available: https://jalammar.github.io/illustrated-stable-diffusion/. [Accessed 2023]. |
| [6] | CompVis, "High-Resolution Image Synthesis with Latent Diffusion Models - Computer Vision & Learning Group," [Online]. Available: https://ommer-lab.com/research/latent-diffusion-models/. [Accessed 2023]. |
| [7] | Stability AI, "Stable Diffusion v2.1 and DreamStudio Updates 7-Dec 22 — Stability AI," [Online]. Available: https://stability.ai/blog/stablediffusion2-1-release7-dec-2022. [Accessed 2023]. |
| [8] | Stability AI, "Stable Diffusion 2-1 - a Hugging Face Space by stabilityai," [Online]. Available: https://huggingface.co/spaces/stabilityai/stable-diffusion. [Accessed 2023]. |
| [9] | CompVis, "CompVis/latent-diffusion: High-Resolution Image Synthesis with Latent Diffusion Models," [Online]. Available: https://github.com/CompVis/latent-diffusion.git. [Accessed 2023]. |