Augmented Reality
Lab 9

Gareth Young, Binh-Son Hua
Trinity College Dublin

In this lab, you will learn about image (2D), multi-view, and 3D synthesis using diffusion models. We will base our exploration on Stable Diffusion, a pre-trained diffusion model commonly used for image generation, and MVDream, a pre-trained diffusion model for multi-view image generation. We will condition the data generation using text prompts.

1. Image Synthesis

Overview

We will perform image generation using StableDiffusion, a popular diffusion model for text-to-image generation. There are a few variants of StableDiffusion circulated on the Internet. Here is a quick list.

- The vanilla LatentDiffusion model (LDM) version resulted from the CVPR 2022 paper: https://github.com/CompVis/latent-diffusion. The pretrained model is a 1.45B model trained on the LAION-400M dataset.
- The v1 StableDiffusion model improved upon LDM, outputting 512x512 image, trained on a subset of LAION-5B dataset. The model was pretrained on 256x256 and then finetuned to generate 512x512 images. More information at: https://github.com/CompVis/stable-diffusion/blob/main/Stable_Diffusion_v1_Model_Card.md
- The v2 StableDiffusion model https://github.com/Stability-AI/stablediffusion. Output 768x768 and 1024x1024 images.
- SDXL is an upgraded version of StableDiffusion in terms of network scale.
- SDXL Turbo is a variant of SDXL with knowledge distillation to allow single-step image generation, reduce the step count from 50 to 1. It can be used for real-time image editing, as on ClipDrop platform: https://clipdrop.co/stable-diffusion-turbo
- StableCascade https://github.com/Stability-AI/StableCascade is StableDiffusion with smaller latent space for faster inference and training.
- SDXL Lightning is another variant of SDXL that can generate images in a few steps. https://huggingface.co/ByteDance/SDXL-Lightning. Real-time demo here: https://huggingface.co/spaces/radames/Real-Time-Text-to-Image-SDXL-Lightning.

Checkout Hugging Face Spaces for other models from the community.

Stable Diffusion

The easiest method to try StableDiffusion is via online inference. Hugging Face is a machine learning community platform for running inferences on pretrained models. The platform provides the popular version 2.1 of Stable Diffusion model: https://huggingface.co/spaces/stabilityai/stable-diffusion

Open Stable Diffusion 2.1, and generate a few images by inputting a few text prompts to describe the content in the images. You can try the following prompts (input to the positive prompt textbox):

- A high-quality photo of an icecream sundae.

Example result:

- A beautiful dress made out of fruit, on a mannequin. Studio lighting, high quality, high resolution.

Example result:

Let's now try generating some images using your own prompts. It is recommended to store your prompts somewhere for reproducibility.

State-of-the-Art Generators

Try DreamStudio, a separate online platform provided by Stability AI, which gives access to the latest version of Stable Diffusion model.
https://beta.dreamstudio.ai/generate

Try commercial platforms for text-to-image generation, e.g., ChatGPT's DALL-E 3, Adobe Firefly. Compare the image quality obtained from these approaches. Which model gives you the best quality images?

How can you engineer your prompts effectively to ensure that the generated images meet your expectations?

Front end

If you run StableDiffusion locally on your machine, the following front end might be of your interest.

The front end of DreamStudio is released as an open source named StableStudio hosted on Github: https://github.com/Stability-AI/StableStudio. It will use Stability API for image synthesis and editing.

Another nice UI if you want to run StableDiffusion offline is ComfyUI: https://github.com/comfyanonymous/ComfyUI.

Note that a front end is not always needed. You can simply run inference in a terminal.

2. Multi-View Image Synthesis

Between image and 3D, an intermediate pseudo-3D representation is multi-view representation. The basic idea is to represent a 3D object by a sparse set of images, each capturing a view of the object, e.g., front, left, right, back view of an object. The architecture of text-to-image diffusion model can be repurposed for multi-view image generation, utilizing the prior knowledge learned from single-view image generation.

A notable open-source multi-view generator is MVDream, which shares a similar architecture to StableDiffusion, but can generate 4 views at a time. The setup instructions for this model can be found here: https://github.com/bytedance/MVDream

Task: try similar prompts from the previous section with MVDream.

Here are a few state-of-the-art view synthesis and multi-view generators, for your own reference.
https://github.com/cvlab-columbia/zero123
https://postech-cvlab.github.io/nvsadapter/
https://github.com/huanngzh/MV-Adapter

2. 3D Model Synthesis

In additional to multi-view generation, image-based diffusion model can also be repurposed to generate 3D data directly. Popular approaches include image-to-3D and text-to-3D generation.

First, let's try out a recent model for image-to-3D generation. It takes as input an image, and the model infers the 3D structure of the image, subject to background removal. Try the following model, also from StabilityAI:
https://huggingface.co/spaces/stabilityai/TripoSR

Second, a recent method for generating 3D models is to generate a neural radiance field based on the guidance from a text-to-image model via score distillation. The implementation of this technique can be quite involved. If you are interested in trying it (e.g., for your project), here is an open-source package to kickstart your exploration.
https://github.com/threestudio-project/threestudio

Challenge: setup threestudio by following the instructions on their Github, and run a text-to-3D generation method such as ProlificDreamer.

A Google Colab notebook is also available here but beware that the optimization will take a few hours, so it is good to have Colab Pro in this case:
https://colab.research.google.com/github/threestudio-project/threestudio/blob/main/threestudio.ipynb

Task: Generate 5-page digital booklet that tells a story about any topic of your choice. Each page should include a multi-view image or a 3D model generated by AI models and a short caption (human written or machine generated) to tell the story. Merge all pages into a single pdf or mp4 (if you would like to display a 3D object in multiple view angles). You can use any diffusion model that you have access to.

Submission:

Submit (1) the story pdf or mp4, and (2) a short report pdf (250 words max) to describe the steps to produce the multi-view images, or the 3D models. In the report, include all text prompts you used for data generation, and the URL or Github link to the pretrained model you used as well as any details necessary for reproducibility. Note that it is not practical to reproduce exact images due to the randomization in the inference process, but we will expect a reproduction of similar image content and quality. Package all the pdfs to lab9_result.zip and submit to Blackboard.

Deadline: Friday, Apr 11, 2025 at 11:59am (noon).

Marking:
You will get 1% from this lab upon a successful and high quality content generation.

Example. Below is an example of a story illustrated by Stable Diffusion. Note that this is single-view image generation. The prompt needs some finetuning to improve image quality as well as consistency across frames.

Once upon a time, in a colorful world of candy, there lived a tiny, friendly dragon named Sparkle. Sparkle loved to paint rainbows across the sky with her fiery breath. One day, she met a giggling unicorn who lost her way home. Together, they embarked on a magical adventure, finding new friends and hidden treasures. In the end, they found the unicorn's home, and Sparkle painted the brightest rainbow they had ever seen to celebrate.

We create an illustration image for each sentence. Some context are repeated for each prompt to improve image consistency.

Prompt 1: Once upon a time, in a colorful world of candy, there lived a tiny, friendly dragon named Sparkle

Prompt 2: In a colorful world of candy, there lived a tiny, friendly green dragon named Sparkle. Sparkle loved to paint rainbows across the sky with her fiery breath.


Prompt 3: In a colorful world of candy, there lived a tiny, friendly green dragon and a sad unicorn

Prompt 4: In a colorful world of candy, a tiny, friendly green dragon and a giggling unicorn embarked on a magical adventure, finding new friends and hidden treasures.


Prompt 5: In a colorful world of candy, a tiny, friendly green dragon and a pink unicorn found a home, and the dragon painted a bright rainbow onto the sky

Acknowledgment: Hubert Kompanowski for assisting to improve the manuscript of this lab.