Exploring Face Synthesis and Detection on a Pre-Trained Network with StyleGAN2

Alyssa Riter
Advisor: Patrick Flynn
Date: 04/04/2025

Introduction - Problem Statement

This project investigates StyleGAN2's ability to generate faces using a pre-trained network that has been fine-tuned on a unique set of images, which we will refer to as the FRGC dataset, and are distinct from those used to train the original model. We will evaluate how effectively InsightFace can recognize faces generated by this new network. Additionally, we will explore the latent space to identify regions where "monsters"—unrecognizable or unrealistic faces—are located, and determine how far we must traverse in latent space to reach them.

Background - Generative Models & The Latent Space

For the face synthesis portion of this project, StyleGAN2 with adaptive discriminator augmentation (ADA) was utilized [1]. The main purpose of StyleGAN2 is to reduce the appearance of artifacts (e.g., water droplets) generated in the original StyleGAN image. In general, the purpose of StyleGAN is to synthesize realistic photo images. It is a very robust general adversarial network (GAN) architecture as it generates extremely realistic images at a high resolution. GANs consist of two neural networks, a generator and a discriminator that are trained together. The generator creates images from random noise while the discriminator tries to distinguish between real images and the fake ones produced by the generator. The primary component is the use of adaptive instance normalization (AdaIN) which is a normalization method that aligns the mean and variance of the content features with those of the style features. In the case of my programs, it provides a mapping network from a latent vector into W. We see progressive changes from low-resolution images to high-resolution images and as we walk along the latent space, some features such as the eyes are fixed into place which can lead to undesired artifacts in the generated images. It was suggested that the droplet artifacts found in the original StyleGAN design are a result of the generator intentionally creating a strong spike to scale the signal as it likes elsewhere. These undesired results led to the development of StyleGAN2.

The developers of StyleGAN2 remove the AdaIN operator and replace it with weight modulation and demodulation in each step. Compared to instance normalization, the demodulation technique is less powerful as it relies on statistical assumptions about the signal rather than the actual contents of the feature maps leading to fewer artifacts found in the generated images. Along with this development, there was the implementation of StyleGAN2-ADA which is a method of training GAN with a limited set of data and reduces the amount of discriminator overfitting that can be found in GANs, especially when the amount of available data is low. Data augmentation involves randomly applying transformations that maintain the integrity of the data to the input data to generate various realistic versions of it, thus increasing the amount of training data available and reducing the likelihood of overfitting.

The benchmark network typically used for facial generation is StyleGAN2 trained on the FFHQ dataset at 1024 x 1024 resolution. This dataset consists of 70,000 PNG images with considerable diversity in terms of age and ethnicity. [2]

After using StyleGAN2-ADA for the face synthesis, InsightFace's SCRFD [3] was used for facial recognition and embeddings. InsightFace generates face embeddings using deep convolutional neural networks, after performing feature extraction the network can take them and convert them into a compact, numerical vector - the face embedding. The vector embedding represents the face's identity in a multi-dimensional space. The key principles behind SCRFD are feature extraction, sample and computation redistribution, bounding box regression, and non-maximum suppression. Feature extraction involves the network analyzing the input image, and extracting features like edges, textures, and shapes in various scales. Sample and computation redistribution strategically allocates more computational resources to areas of the image more likely to contain faces, overall improving efficiency. Bounding box regression allows the network to predict the location and size of potential faces by drawing a bounding box around the face. Lastly, non-maximum suppression ensures that the most accurate bounding box for each face is kept, removing overlapping or redundant boxes.

Methodology

To conduct the experiments and obtain the percent yield and the confidence score for each set of generations we ran the face synthesis algorithm (StyleGAN2-ada) on a random walk of 500 images. We simulated this random walk at least 10 times per network using a different random seed each time to get a variety of image sets. Starting at the origin, we traversed the latent space using a step size of 0.01. After generating these images we ran them through InsightFace to get the confidence score, bounding region, and embeddings for each image. The extraction of vector embeddings will allow us to visualize the vector space and detect any clusters forming in the region. To better visualize the results we graphed the confidence scores for each batch of 500 images, with undetectable faces yielding a confidence score of 0.00. We repeated the experiment twice, once for the original FFHQ dataset that StyleGAN is trained on, and a second time for the newly fine-tuned network using the FRGC images.

The graphs utilized for visualization included a histogram of the confidence scores to understand the range of confidence InsightFace in detecting the generated faces. Generating a histogram for each batch of 500 images from each network allowed for comparison in how quantities of monsters compared in the output of StyleGAN across the two networks.

Experimental Process

Defining Experimental Parameters: Images were generated, projected, and detected using a pretrained network on FRGC data. The resulting serialized file was then used as the network input for the StyleGAN2-ada model.

Data Description: The pretrained FRGC network was developed using a series of facial images captured in diverse locations and under varying lighting conditions. To evaluate the impact of environmental variation, the same subjects were photographed in each setting, enabling analysis of how different contexts influence model performance. The dataset is composed of primarily Caucasian and Asian volunteers, causing for there to be scarce diversity in StyleGAN's output.

Computing the Average Latent Vector: To compute the average latent vector we will use as our starting point the `random_walk.py` script which is implemented upon StyleGAN2's `projector.py` the script generates 10,000 random latent vectors (z) from a normal distribution. The script then maps those vectors into the intermediate latent space (W) and then maps the latent space (W) to an image. We only use the first W vector per sample to compute the average with the 10,000 samples. Using the average latent vector allows for smoothness in the generations since it is close to the center of the distribution of realistic faces.

Random Walk in Latent Space: The random_walk.py script creates a sequence of slightly different latent vectors starting from the average latent vector. Each new vector is a very small step from the previous one. Using large steps would make us stray into the regions of monsters very quickly since we are reaching far beyond the origin. This small step size keeps us localized while giving us better insight into how far we must stray from the origin to get unrecognizable faces. A small step size also creates a smooth transition through the space of images the GAN can generate.

Image Synthesis: Images were synthesized using StyleGAN2-ada's `projector.py` and `generate.py` scripts. To explore the latent space more broadly, I developed a custom script called `random_walk.py`, which incrementally perturbs latent vectors in random directions. This allowed for the identification of regions in latent space where faces become unrecognizable or distorted.

Face Detection: For facial detection, we utilized InsightFace, which allowed us to evaluate each synthesized image by checking: whether a face was detected, the structure of the latent vector, and the recognizability of the face. Using a script called `face_detector_embeddings.py`, we parallelized a face embedding extractor to take a set of image files, the 500 generated in each batch, and outputs face detection scores and embeddings (512-D vectors) to a CSV file.

Graphing (Minimum Spanning Tree):

Results & Discussion

Initial experiments with incorrectly cropped input images resulted in highly distorted reconstructions—often generating surreal, cyclopean "monsters." These distortions were due to misalignments that violated the input constraints expected by the StyleGAN2 model. Once images were properly aligned and cropped to match the model's expected dimensions, the quality of latent vector projections improved significantly, producing more recognizable and human-like facial reconstructions.

Notably, the network trained on the FRGC dataset consistently produced less recognizable faces compared to the standard StyleGAN2-ADA model trained on the FFHQ dataset. This discrepancy highlights the importance of both dataset quality and alignment in achieving high-fidelity projections.

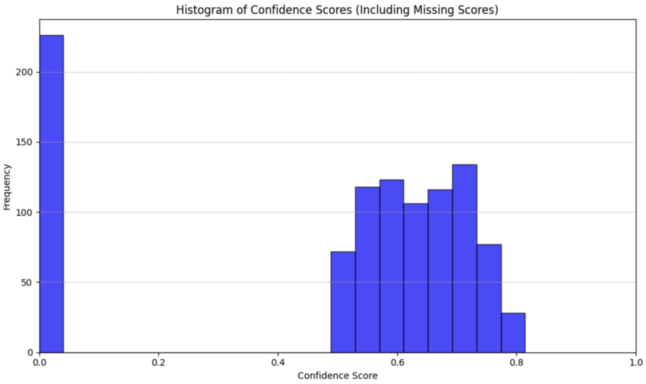

Figure 1. Histogram of Confidence Scores for FRGC Network

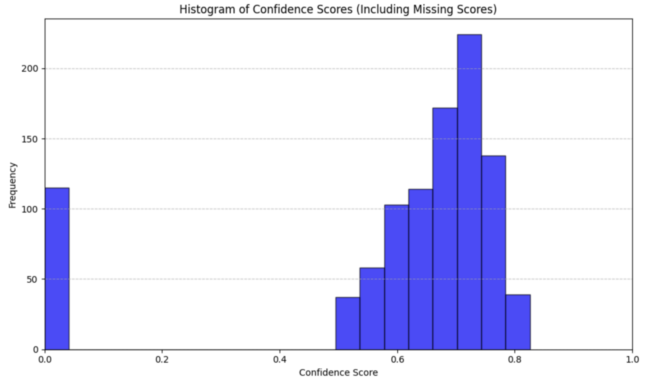

Figure 2. Histogram of Confidence Scores for FFHQ Network

As mentioned earlier, improperly formatted input images yielded latent projections that were highly distorted, often featuring only one visible eye. In contrast, properly cropped input images resulted in dramatically improved outputs. This effect is illustrated using the same subject: one projection based on a raw image (Figure 3) and another based on a well-cropped version of the same image (Figure 4). The stark contrast between these outputs underscores the importance of adhering strictly to StyleGAN2's input requirements in order to achieve meaningful and accurate reconstructions.

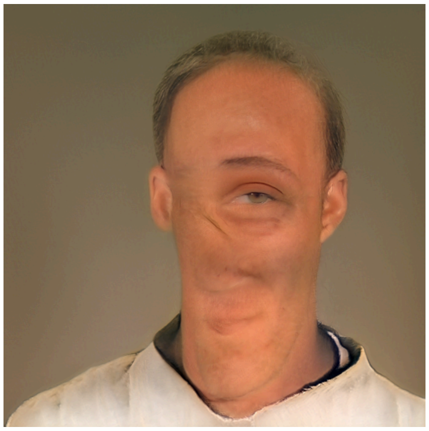

Figure 3. Image Projection using improperly cropped target

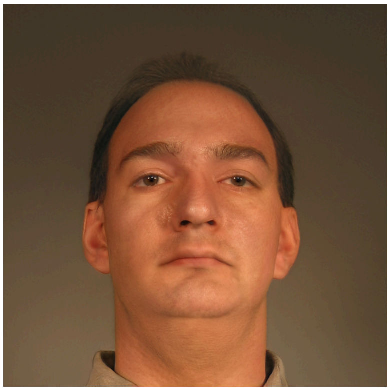

Figure 4. Image Projection using properly cropped target (same subject as figure 3)

An important point to note is that the newly fine-tuned network, trained on FRGC images, primarily consists of Caucasian and Asian faces and performs relatively well on those inputs. However, we have yet to evaluate its performance on other ethnicities, presenting an opportunity for a meaningful future experiment.

Future Work

Future experiments should scale to larger datasets to further evaluate how well the StyleGAN2 latent space captures and preserves identity. A key experiment would be Investigating how consistently StyleGAN can synthesize multiple distinct images of the same individual. Along with exploring where individual identities are clustered within the latent space. Additionally, tweaking StyleGAN's architecture or loss functions to balance identity preservation with generative diversity. Some guiding questions would be: How accurately does the StyleGAN latent space represent identity? How can we adjust StyleGAN to enhance synthesis while maintaining identity coherence? How are different individuals distributed across latent space?

References

[1] Training Generative Adversarial Networks with Limited Data
[2] FFHQ Dataset
[3] InsightFace Github