Advancing Explainable AI Models for Prognosing Lung Disease via Feature Disentanglement

Zhiling Yue,
Yingying Fang,
Guang Yang

Abstract

This project investigates semantic disentanglement techniques to analyse the intermediate latent spaces in generative AI models, such as StyleGAN and Diffusion models. By isolating and utilising these disentangled features, we aim to develop prognostic models for lung diseases that are more interpretable for clinicians and researchers. This approach will facilitate the identification of novel biomarkers for lung disease prognosis

Introduction

Background

High-Resolution Computed Tomography (HRCT) is essential for diagnosing various lung diseases, but interpreting tissue patterns is challenging due to diverse manifestations and similarities across different diseases. Deep learning has demonstrated exceptional capabilities in medical classification, offering cost-effective and accurate prognostic predictions from medical images.

Limitation

- The 'black-box' nature of AI models makes their results opaque and hard to integrate clinically.
- Existing studies focus on natural images, which are easier to interpret than medical images.

Objective

- Investigate the disentanglement of latent space abilities across different AI models.
- Develop interpretable AI tools for lung disease diagnosis and prognosis.

Methods

Disentanglement of latent spaces

- Develop a generative-based autoencoder and integrate it with StyleGAN model [1] .
- Investigate the disentanglement abilities of StyleGAN's latent spaces, including the random latent space Z and intermediate spaces W and S, which better reflect the disentangled nature of the learned distribution [2].

Semantic Channel Identification and Manipulation

- Develop methods to automatically identify semantic channels that control specific target attributes.

Train AI model and evaluate semantic feature for lung disease

- Train and fine-tune StyleGAN model on HRCT dataset.
- Evaluate feature disentanglement ability for fibrosis feature generation.
- Prognose fibrosis lung disease from images with difficult-to-observe biomarker tissue patterns.

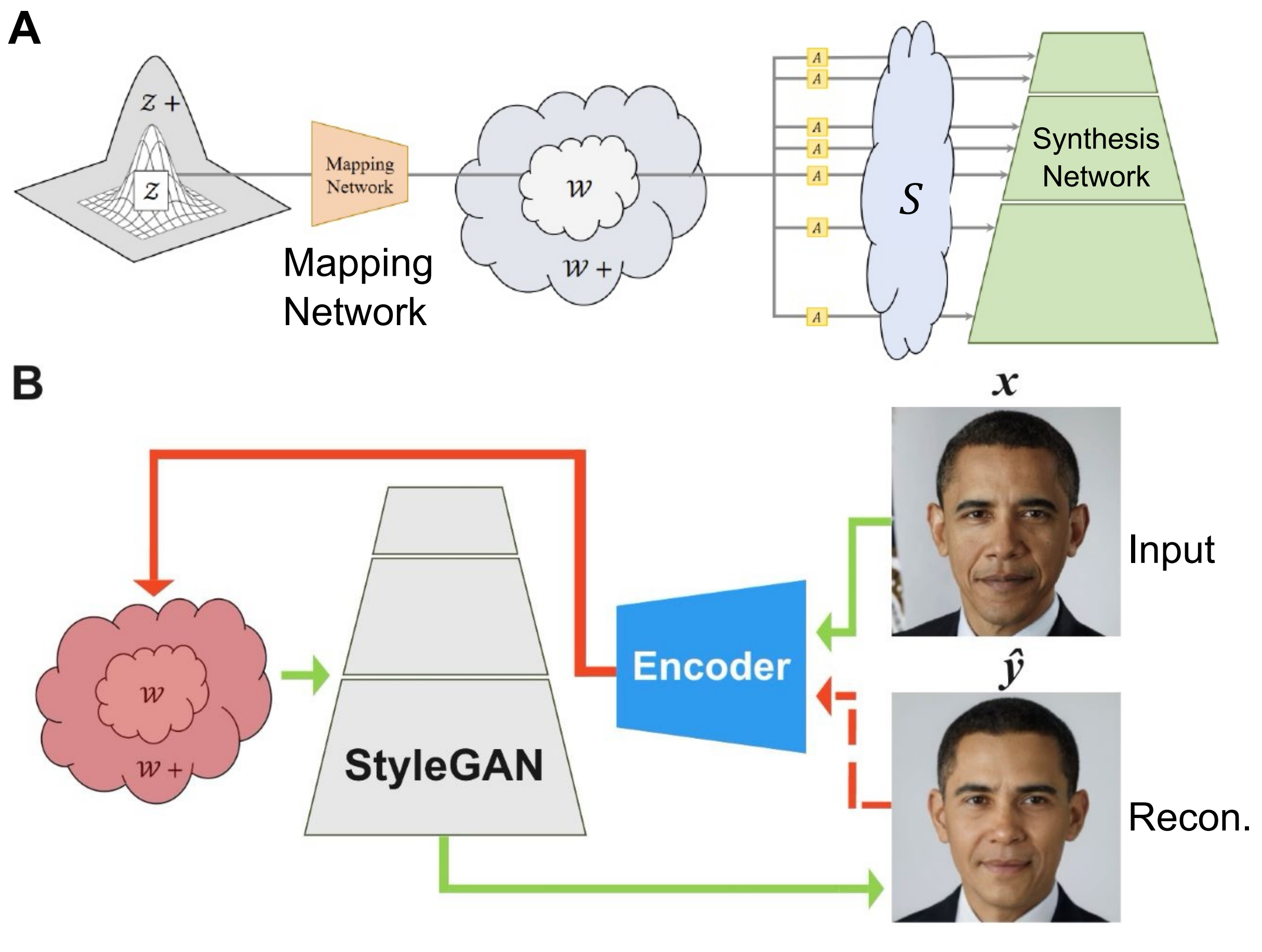

Fig. 1. Structure of StyleGAN and Encoder Integrated Model [3].
A: *The Latent Spaces in StyleGAN .* ***B:*** *Encoder-based mapping process of translating raw images directly into W Latent Space for manipulating. The model output will be a reconstruction of original input image.*

Preliminary Results

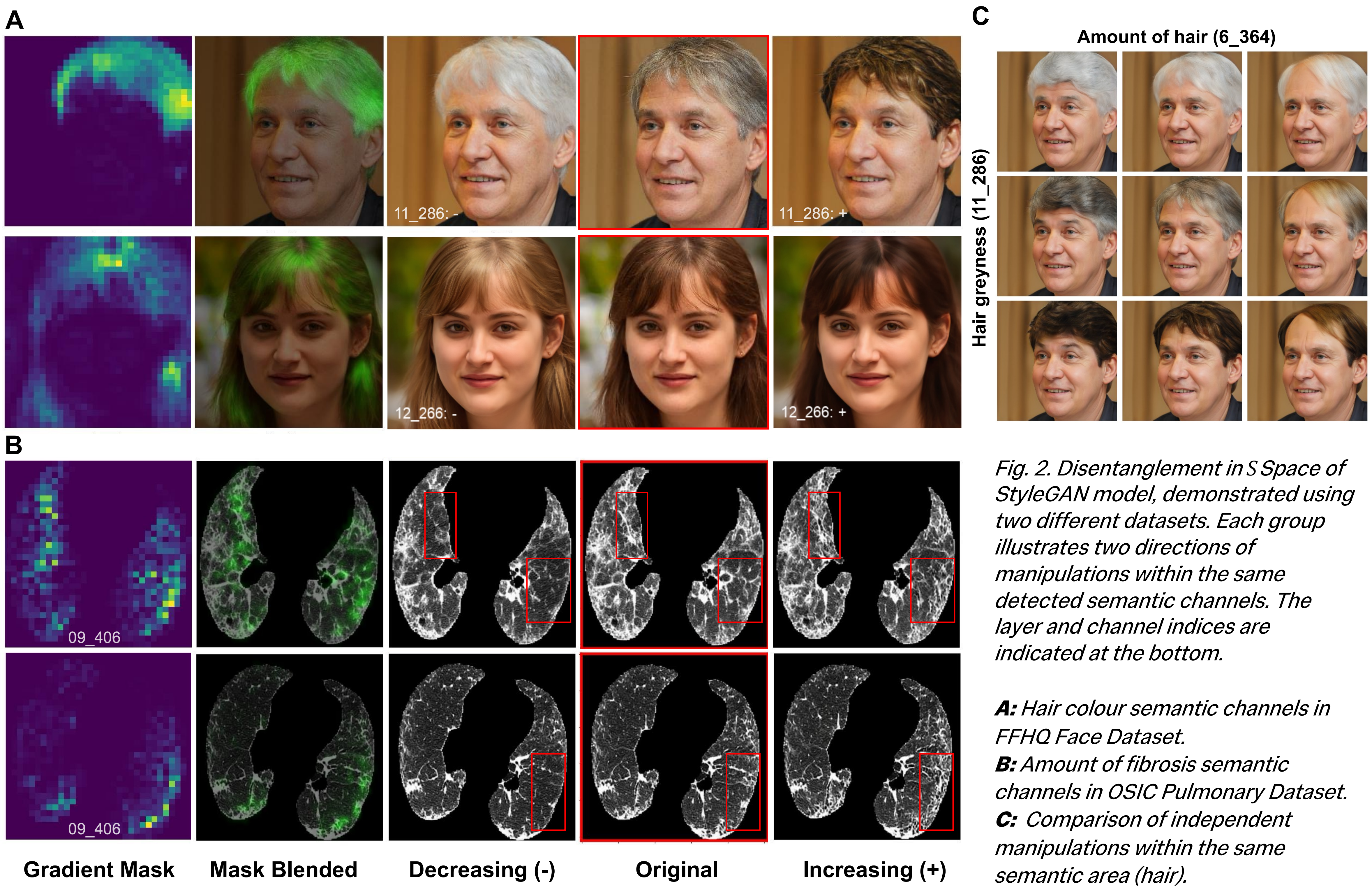

Fig. 2. Disentanglement in S Space of StyleGAN model, demonstrated using two different datasets. Each group illustrates two directions of manipulations within the same detected semantic channels. The layer and channel indices are indicated at the bottom.

A: *Hair colour semantic channels in FFHQ Face Dataset.*
B: *Amount of fibrosis semantic channels in OSIC Pulmonary Dataset.*
C: *Comparison of independent manipulations within the same semantic area (hair).*

Future Plan

Enhance Prognostic Models

Continue developing and refining image prognostic models utilising synthesised fibrosis feature.

Quantify Disentangled Features

Develop methods to identify and quantify the disentangled features that are closely associated with the model's decision.

Pattern Analysis

Analyse identified patterns in relation to disease outcomes through statistical analysis and expert consultations.

Reference

[1] Karras, T., et al. (2020). Analyzing and Improving the Image Quality of StyleGAN. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[2] Wu, Z., et al. (2021). StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE.
[3] Melnik, A., et al. (2024). "Face generation and editing with stylegan: A survey." IEEE Transactions on Pattern Analysis and Machine Intelligence.

Hamlyn Centre
for Robotic Surgery