

ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

**Xây dựng ứng dụng thay đổi thuộc tính khuôn mặt
dựa trên mô hình StarGAN**

Lý Văn Hiếu

hieu.lv240108e@sis.hust.edu.vn

**Chương trình đào tạo: Kỹ sư chuyên sâu Trí tuệ nhân tạo tạo
sinh**

Giảng viên hướng dẫn: TS. Đỗ Bá Lâm

Khoa: Khoa học máy tính

Trường: Công nghệ thông tin và Truyền thông

HÀ NỘI, 01/2026

ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Xây dựng ứng dụng thay đổi thuộc tính khuôn mặt
dựa trên mô hình StarGAN

Lý Văn Hiếu

hieu.lv240108e@sis.hust.edu.vn

Chương trình đào tạo: Kỹ sư chuyên sâu Trí tuệ nhân tạo tạo
sinh

Giảng viên hướng dẫn: TS. Đỗ Bá Lâm

Chữ ký GVHD

Khoa: Khoa học máy tính

Trường: Công nghệ Thông tin và Truyền thông

HÀ NỘI, 01/2026

LỜI CẢM ƠN

Trước hết, em xin bày tỏ lòng biết ơn sâu sắc tới TS. Đỗ Bá Lâm, người đã trực tiếp hướng dẫn, định hướng nghiên cứu và luôn tận tình hỗ trợ, tạo điều kiện thuận lợi để em có thể hoàn thành đồ án tốt nghiệp này.

Em xin trân trọng cảm ơn các thầy cô Trường Công nghệ Thông tin và Truyền thông, cũng như các thầy cô tại Đại học Bách khoa Hà Nội, những người đã truyền đạt cho em những kiến thức quý báu trong suốt quá trình học tập và rèn luyện tại trường, là nền tảng quan trọng giúp em hoàn thành tốt đồ án này.

Cuối cùng, em xin gửi lời cảm ơn chân thành tới gia đình và bạn bè, những người luôn bên cạnh động viên, khích lệ và tạo động lực cho em trong suốt quá trình học tập cũng như thực hiện đồ án tốt nghiệp.

Em xin chân thành cảm ơn!

TÓM TẮT NỘI DUNG ĐỒ ÁN

Trong những năm gần đây, bài toán chỉnh sửa và thay đổi thuộc tính khuôn mặt trên ảnh chân dung đã thu hút nhiều sự quan tâm trong lĩnh vực thị giác máy tính và học máy, học sâu, đặc biệt với sự phát triển mạnh mẽ của các mô hình sinh đối kháng (Generative Adversarial Networks – GANs). Các ứng dụng của bài toán này rất đa dạng, từ chỉnh sửa ảnh, giải trí, đến hỗ trợ nghiên cứu và các hệ thống tương tác người – máy. Nhiều hướng tiếp cận đã được đề xuất, tiêu biểu như các mô hình GAN truyền thống cho từng thuộc tính riêng lẻ, hoặc các mô hình học có giám sát mạnh. Tuy nhiên, các phương pháp này thường gặp hạn chế về khả năng mở rộng, chi phí huấn luyện cao và khó đảm bảo tính nhất quán của hình ảnh khi thay đổi nhiều thuộc tính khác nhau.

Trước những hạn chế đó, đồ án này lựa chọn hướng tiếp cận dựa trên mô hình StarGAN, một kiến trúc GAN đa miền cho phép thực hiện thay đổi nhiều thuộc tính khuôn mặt chỉ với một mô hình duy nhất. Hướng tiếp cận này được lựa chọn nhờ khả năng linh hoạt, hiệu quả trong huấn luyện và đã được chứng minh tính hiệu quả trong các nghiên cứu trước đó.

Trên cơ sở hướng tiếp cận đã chọn, đồ án tiến hành xây dựng quy trình huấn luyện mô hình StarGAN trên tập dữ liệu CelebA, bao gồm các bước tiền xử lý dữ liệu, thiết kế chiến lược huấn luyện, lựa chọn siêu tham số và tối ưu mô hình. Bên cạnh đó, đồ án đề xuất và triển khai quy trình đánh giá chất lượng mô hình thông qua các chỉ số định lượng như FID và SSIM, đồng thời phân tích kết quả theo từng thuộc tính khuôn mặt thay đổi trên ảnh. Cuối cùng là triển khai một số ứng dụng sử dụng mô hình.

Đóng góp chính của đồ án là xây dựng thành công một mô hình StarGAN có khả năng thay đổi nhiều thuộc tính khuôn mặt trên ảnh với chất lượng ổn định, đồng thời phát triển ứng dụng minh họa cho phép người dùng tương tác và chỉnh sửa ảnh khuôn mặt một cách trực quan. Kết quả thực nghiệm cho thấy mô hình đạt được chất lượng sinh ảnh khá tốt, giữ được đặc trưng khuôn mặt gốc và đáp ứng mục tiêu đề ra của đồ án. Tuy nhiên vẫn có thể cải thiện thêm ở một số hướng, sẽ được trình bày ở phần kết luận.

Sinh viên thực hiện
(Ký và ghi rõ họ tên)

MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI.....	1
1.1 Đặt vấn đề.....	1
1.2 Mục tiêu và phạm vi đề tài.....	1
1.3 Định hướng giải pháp.....	2
1.4 Bố cục đồ án	3
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT	5
2.1 Tổng quan về xử lý ảnh.....	5
2.2 Mạng sinh đôi kháng - GAN.....	6
2.3 Bài toán thay đổi thuộc tính khuôn mặt.....	8
2.4 Mô hình StarGAN và các nghiên cứu liên quan	10
2.4.1 Tổng quan	10
2.4.2 Ý tưởng của StarGAN	10
2.4.3 Các hàm măt măt trong StarGAN	11
2.4.4 Ưu điểm của StarGAN trong bài toán thay đổi thuộc tính khuôn mặt	12
2.4.5 Các biến thể và nghiên cứu mở rộng từ StarGAN	12
2.4.6 Đánh giá tổng quan	12
2.5 Các chỉ số đánh giá chất lượng ảnh sinh.....	13
2.5.1 Chỉ số tương đồng cấu trúc SSIM	13
2.5.2 Chỉ số Fréchet Inception Distance (FID)	14
CHƯƠNG 3. PHÂN TÍCH THIẾT KẾ HỆ THỐNG.....	16
3.1 Kiến trúc tổng thể của hệ thống.....	16
3.1.1 Mô tả tổng quan.....	16
3.1.2 Khối chuẩn bị dữ liệu và tiền xử lý.....	16

3.1.3 Khối huấn luyện mô hình StarGAN	17
3.1.4 Khối ứng dụng và triển khai	17
3.2 Thiết kế pipeline huấn luyện mô hình	18
3.2.1 Tổng quan pipeline huấn luyện	18
3.2.2 Tổ chức dữ liệu huấn luyện.....	19
3.2.3 Huấn luyện mô hình	19
3.2.4 Lưu checkpoint và theo dõi quá trình huấn luyện.....	20
3.2.5 Tính mở rộng và khả năng tái sử dụng của pipeline	20
3.3 Thiết kế pipeline sinh ảnh và đánh giá	20
3.3.1 Pipeline sinh ảnh từ mô hình đã huấn luyện	21
3.3.2 Đánh giá định lượng	22
3.3.3 Đánh giá định tính.....	22
3.3.4 Phân tích và so sánh giữa các checkpoint huấn luyện	23
CHƯƠNG 4. HUẤN LUYỆN MÔ HÌNH	24
4.1 Chuẩn bị dữ liệu và thiết lập mô hình	24
4.1.1 Giới thiệu bộ dữ liệu CelebA.....	24
4.1.2 Lựa chọn và sử dụng tập thuộc tính.....	25
4.1.3 Tiền xử lý dữ liệu	25
4.1.4 Thiết lập mô hình StarGAN	26
4.2 Triển khai huấn luyện mô hình.....	26
4.2.1 Quy trình huấn luyện tổng thể	27
4.2.2 Hàm mất mát và chiến lược tối ưu.....	27
4.2.3 Thiết lập siêu tham số và cấu hình huấn luyện	28
4.2.4 Môi trường triển khai	29
4.2.5 Lưu checkpoint và chiến lược đánh giá trung gian	29

4.3 Các vấn đề gặp phải và cách khắc phục.....	30
4.3.1 Hạn chế về tài nguyên tính toán	30
4.3.2 Hạn chế về thời gian huấn luyện	31
4.3.3 Hạn chế trong đánh giá định lượng	31
CHƯƠNG 5. THỰC NGHIỆM VÀ ĐÁNH GIÁ	32
5.1 Thiết lập môi trường thực nghiệm	32
5.2 Phương pháp đánh giá mô hình	33
5.3 Kết quả đánh giá tổng thể	35
5.4 Phân tích kết quả theo từng thuộc tính thay đổi trên ảnh khuôn mặt	36
5.5 Nhận xét và so sánh các checkpoint huấn luyện	40
CHƯƠNG 6. XÂY DỰNG ỨNG DỤNG THỦ NGHIỆM	43
6.1 Chức năng của ứng dụng	43
6.2 Cốt lõi của ứng dụng	44
6.3 Xây dựng chatbot sử dụng mô hình được huấn luyện.....	46
6.4 Xây dựng web site sử dụng mô hình được huấn luyện.....	50
CHƯƠNG 7. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	53
7.1 Kết luận	53
7.2 Hạn chế của đồ án hiện tại	53
7.3 Hướng phát triển.....	54
TÀI LIỆU THAM KHẢO.....	56

DANH MỤC HÌNH VẼ

Hình 2.1	Hình minh họa StarGAN thể hiện cấu trúc hình sao kết nối đa miền [7]	11
Hình 3.1	Tổng quan các khôi	16
Hình 3.2	Các giai đoạn trong quá trình huấn luyện	18
Hình 3.3	Các giai đoạn của việc sinh ảnh và đánh giá	21
Hình 4.1	Giới thiệu về bộ dữ liệu CelebA	24
Hình 5.1	Ví dụ về ảnh sinh ra	33
Hình 5.2	Biểu đồ chỉ số SSIM qua các iteration	35
Hình 5.3	Biểu đồ chỉ số FID qua các iteration	35
Hình 5.4	Biểu đồ FID của nhóm ảnh liên quan đến tóc	37
Hình 5.5	Biểu đồ FID của nhóm ảnh liên quan đến râu	38
Hình 5.6	Biểu đồ FID của nhóm ảnh liên quan đến hình thái khuôn mặt, màu da mặt	39
Hình 5.7	Biểu đồ FID của nhóm ảnh liên quan đến biểu cảm khuôn mặt, hình miệng	39
Hình 5.8	Biểu đồ FID của nhóm ảnh liên quan đến trang điểm, phụ kiện	40
Hình 6.1	Minh họa chức năng ứng dụng	43
Hình 6.2	Các bước xử lý của hệ thống sinh ảnh	44
Hình 6.3	Minh họa sử dụng chatbot: tìm kiếm bot trên Telegram.	48
Hình 6.4	Minh họa sử dụng chatbot: chọn ảnh đầu vào.	48
Hình 6.5	Minh họa sử dụng chatbot: chọn tiêu chí.	49
Hình 6.6	Minh họa sử dụng chatbot: nhận ảnh kết quả.	49
Hình 6.7	Minh họa sử dụng web: chọn ảnh đầu vào.	51
Hình 6.8	Minh họa sử dụng web: chọn tiêu chí và sinh ảnh.	51
Hình 6.9	Minh họa sử dụng web: nhận thông báo và ảnh sinh ra.	52

DANH MỤC TỪ VIẾT TẮT

Viết tắt	Tên tiếng Anh	Tên tiếng Việt
GAN	Generative Adversarial Network	Mạng sinh đối kháng
StarGAN	Star Generative Adversarial Network	Mạng sinh đối kháng đa miền Star-GAN
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
AI	Artificial Intelligence	Trí tuệ nhân tạo
DL	Deep Learning	Học sâu
CV	Computer Vision	Thị giác máy tính
FID	Fréchet Inception Distance	Chỉ số đánh giá độ khác biệt Fréchet Inception
SSIM	Structural Similarity Index Measure	Chỉ số đo độ tương đồng cấu trúc
GPU	Graphics Processing Unit	Bộ xử lý đồ họa
CPU	Central Processing Unit	Bộ xử lý trung tâm
ĐATN		Đô án tốt nghiệp
SV		Sinh viên

DANH MỤC THUẬT NGỮ

Tên tiếng Anh	Ý nghĩa
Pipeline	Chuỗi xử lý (quy trình gồm nhiều bước liên tiếp để xử lý dữ liệu hoặc huấn luyện mô hình)
Dataset	Tập dữ liệu dùng cho huấn luyện, kiểm tra hoặc đánh giá mô hình
Training	Quá trình huấn luyện mô hình học máy
Checkpoint	Trạng thái lưu lại của mô hình tại một thời điểm huấn luyện nhất định
Evaluation	Quá trình đánh giá hiệu năng của mô hình
Iteration	Một vòng lặp huấn luyện, đại diện cho một lần cập nhật tham số của mô hình

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

1.1 Đặt vấn đề

Trong những năm gần đây, cùng với sự phát triển mạnh mẽ của công nghệ thông tin và trí tuệ nhân tạo, các ứng dụng xử lý và phân tích hình ảnh ngày càng trở nên phổ biến và đóng vai trò quan trọng trong nhiều lĩnh vực của đời sống. Đặc biệt, bài toán xử lý và chỉnh sửa ảnh khuôn mặt đang nhận được sự quan tâm lớn từ cả giới nghiên cứu lẫn cộng đồng người dùng, do khuôn mặt là một trong những đối tượng mang tính nhận diện cao và gắn liền với nhiều ứng dụng thực tiễn như mạng xã hội, giải trí, thương mại điện tử, an ninh – giám sát và truyền thông đa phương tiện.

Trong thực tế, nhu cầu thay đổi hoặc điều chỉnh một số thuộc tính trên khuôn mặt, chẳng hạn như kiểu tóc, giới tính, biểu cảm hay phong cách trang điểm, xuất hiện ngày càng nhiều. Các công cụ chỉnh sửa ảnh truyền thống thường đòi hỏi người dùng có kỹ năng chuyên môn, thao tác thủ công phức tạp và khó đảm bảo tính tự nhiên cũng như nhất quán của khuôn mặt sau khi chỉnh sửa. Điều này gây ra không ít hạn chế khi áp dụng trên quy mô lớn hoặc trong các hệ thống tự động.

Bên cạnh đó, với sự bùng nổ của dữ liệu hình ảnh và sự phát triển của các nền tảng trực tuyến, việc xây dựng các hệ thống có khả năng tự động xử lý, biến đổi và tạo ra hình ảnh khuôn mặt một cách linh hoạt, chân thực và hiệu quả đang trở thành một yêu cầu cấp thiết. Nếu bài toán này được giải quyết tốt, nó không chỉ giúp nâng cao trải nghiệm người dùng trong các ứng dụng chỉnh sửa ảnh, mà còn mở ra nhiều cơ hội ứng dụng trong các lĩnh vực khác như hỗ trợ sản xuất nội dung số, mô phỏng nhân vật, nghiên cứu hành vi thị giác, hay tạo dữ liệu huấn luyện cho các bài toán thị giác máy tính.

Từ những yêu cầu và thách thức nêu trên, bài toán thay đổi thuộc tính khuôn mặt một cách tự động, đảm bảo giữ nguyên đặc trưng nhận dạng của khuôn mặt gốc đồng thời tạo ra các biến đổi hợp lý, tự nhiên, đang đặt ra nhiều vấn đề cần được nghiên cứu và đánh giá một cách nghiêm túc trong bối cảnh hiện nay.

1.2 Mục tiêu và phạm vi đề tài

Từ những vấn đề đã nêu ở Mục 1.1, có thể thấy rằng nhu cầu thay đổi thuộc tính khuôn mặt một cách linh hoạt, tự nhiên và có khả năng kiểm soát đang ngày càng trở nên phổ biến trong nhiều lĩnh vực khác nhau.

Các công trình nghiên cứu gần đây cho thấy các mô hình GAN đa miền có tiềm năng lớn trong việc giải quyết bài toán thay đổi nhiều thuộc tính khuôn mặt trên

cùng một mô hình, thay vì phải huấn luyện riêng lẻ cho từng thuộc tính. Tuy nhiên, trong nhiều trường hợp, việc đánh giá chất lượng mô hình mới chỉ dừng lại ở quan sát trực quan, chưa đi kèm với các phân tích định lượng của ảnh sinh ra. Bên cạnh đó, không phải nghiên cứu nào cũng hướng tới việc xây dựng một ứng dụng hoàn chỉnh miễn phí hoàn toàn.

Trên cơ sở đó, đồ án này hướng tới mục tiêu nghiên cứu và xây dựng một hệ thống thay đổi thuộc tính khuôn mặt dựa trên mô hình StarGAN, trong đó tập trung vào ba khía cạnh chính. Thứ nhất, tiến hành huấn luyện mô hình với nhiều thuộc tính khuôn mặt khác nhau nhằm khai thác khả năng biến đổi đa miền của StarGAN. Thứ hai, đánh giá chất lượng mô hình một cách định lượng thông qua các chỉ số phù hợp. Thứ ba, xây dựng một ứng dụng thử nghiệm cho phép người dùng tương tác với mô hình đã huấn luyện, qua đó minh họa khả năng ứng dụng của mô hình trong thực tế.

Phạm vi của đề tài được giới hạn trong bài toán thay đổi thuộc tính khuôn mặt trên tập dữ liệu ảnh khuôn mặt phổ biến. Đồ án tập trung vào việc nghiên cứu, triển khai, đánh giá và ứng dụng mô hình ở mức độ một đồ án học tập và nghiên cứu, làm cơ sở cho các hướng phát triển sâu hơn trong tương lai, chưa đạt đến giá trị mà có thể đưa vào thị trường ứng dụng hoặc kinh doanh.

1.3 Định hướng giải pháp

Xuất phát từ các mục tiêu và phạm vi đã xác định ở Mục 1.2, đồ án lựa chọn tiếp cận bài toán thay đổi thuộc tính khuôn mặt theo hướng sử dụng mô hình học sâu dựa trên GAN, cụ thể là mô hình StarGAN – một kiến trúc GAN đa miền cho phép xử lý nhiều thuộc tính trên cùng một mô hình thống nhất. Định hướng này được lựa chọn do StarGAN có khả năng thay đổi linh hoạt nhiều thuộc tính khuôn mặt khác nhau mà vẫn giữ được các đặc trưng nhận dạng chính, đồng thời giảm đáng kể chi phí huấn luyện so với việc xây dựng nhiều mô hình riêng lẻ cho từng thuộc tính.

Trên cơ sở định hướng đó, giải pháp của đồ án tập trung vào việc huấn luyện mô hình StarGAN trên tập dữ liệu ảnh khuôn mặt với nhiều thuộc tính khác nhau (ở đồ án hiện tại là 17 thuộc tính khác nhau), từ đó tạo ra các ảnh khuôn mặt mới với thuộc tính được thay đổi theo yêu cầu. Bên cạnh việc huấn luyện, chất lượng ảnh sinh ra bởi mô hình được đánh giá một cách hệ thống thông qua cả quan sát trực quan và các chỉ số định lượng phù hợp. Cuối cùng, mô hình sau huấn luyện được tích hợp vào một ứng dụng thử nghiệm, cho phép người dùng trực tiếp trải nghiệm quá trình thay đổi thuộc tính khuôn mặt.

Đồ án này đã xây dựng và triển khai thành công một quy trình hoàn chỉnh từ huấn luyện, đánh giá đến ứng dụng mô hình StarGAN cho bài toán thay đổi thuộc

tính khuôn mặt. Kết quả đạt được cho thấy mô hình có khả năng biến đổi nhiều thuộc tính khác nhau với chất lượng ổn định. Bên cạnh đó, ứng dụng được xây dựng giúp minh họa tính khả thi của việc đưa mô hình vào sử dụng thực tế, tạo tiền đề cho các hướng phát triển và mở rộng trong tương lai.

1.4 Bố cục đồ án

Phần còn lại của quyển đồ án tốt nghiệp này được tổ chức như sau.

Chương 2 trình bày các cơ sở lý thuyết liên quan trực tiếp đến bài toán thay đổi thuộc tính khuôn mặt. Chương này sẽ giới thiệu tổng quan về xử lý ảnh và các khái niệm nền tảng thường được sử dụng trong lĩnh vực xử lý ảnh số. Tiếp đó, chương đi sâu vào mạng sinh đối kháng GAN, bao gồm nguyên lý hoạt động, cấu trúc cơ bản và các biến thể tiêu biểu. Trên cơ sở đó, bài toán thay đổi thuộc tính khuôn mặt được phân tích rõ hơn, làm tiền đề để giới thiệu mô hình StarGAN cùng các nghiên cứu liên quan. Cuối cùng, chương trình bày các chỉ số đánh giá chất lượng ảnh sinh, đóng vai trò quan trọng trong việc đánh giá kết quả huấn luyện và thực nghiệm ở các chương sau.

Chương 3 tập trung vào phân tích và thiết kế hệ thống tổng thể của đồ án. Chương này mô tả kiến trúc chung của hệ thống. Tiếp theo, chương sẽ trình bày các bước trong quá trình huấn luyện mô hình và quá trình đánh giá, là tiền đề cho hai chương phân tích chi tiết phía sau.

Chương 4 trình bày quá trình huấn luyện mô hình StarGAN trong đồ án. Nội dung chương bắt đầu với việc chuẩn bị dữ liệu và thiết lập các cấu hình cần thiết cho mô hình. Tiếp theo, quá trình triển khai huấn luyện được mô tả chi tiết.

Chương 5 tập trung vào thực nghiệm và đánh giá mô hình sau huấn luyện. Chương này đều tiên sẽ trình bày môi trường thực nghiệm. Sau đó sẽ mô tả phương pháp đánh giá mô hình, kết hợp giữa đánh giá định tính và định lượng. Cuối chương sẽ so sánh các checkpoint từ quá trình huấn luyện nhằm làm rõ ảnh hưởng của quá trình huấn luyện đến chất lượng ảnh sinh ra.

Chương 6 trình bày việc xây dựng ứng dụng thử nghiệm sử dụng mô hình đã được huấn luyện. Chương này mô tả các chức năng chính của ứng dụng, cũng như phần cốt lõi chung của việc tích hợp mô hình StarGAN vào hệ thống. Sau đó, sẽ trình bày việc xây dựng chatbot và website sử dụng mô hình được huấn luyện để nhằm minh họa khả năng ứng dụng thực tế của kết quả đồ án.

Cuối cùng, chương 7 tổng kết các kết quả đạt được của đồ án. Chương này đưa ra các kết luận chính, chỉ ra những hạn chế còn tồn tại trong quá trình thực hiện, đồng thời đề xuất các hướng phát triển và mở rộng trong tương lai nhằm nâng cao

hiệu quả và phạm vi ứng dụng của mô hình.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

Chương này trình bày các cơ sở lý thuyết nền tảng phục vụ cho việc nghiên cứu và triển khai bài toán thay đổi thuộc tính khuôn mặt trong đồ án. Trước hết, chương giới thiệu những khái niệm cơ bản của xử lý ảnh số và vai trò của xử lý ảnh trong các hệ thống thị giác máy tính hiện đại. Tiếp theo, các nguyên lý của mạng sinh đối kháng (GAN) được trình bày nhằm làm rõ cơ sở hình thành các mô hình sinh ảnh. Trên nền tảng đó, bài toán thay đổi thuộc tính khuôn mặt và mô hình StarGAN cùng các nghiên cứu liên quan được phân tích. Cuối cùng, chương giới thiệu các chỉ số đánh giá chất lượng ảnh sinh, làm tiền đề cho quá trình thực nghiệm và đánh giá mô hình ở các chương sau.

2.1 Tổng quan về xử lý ảnh

Xử lý ảnh số (Digital Image Processing) là một lĩnh vực quan trọng của khoa học máy tính và thị giác máy tính, nghiên cứu các phương pháp và thuật toán nhằm phân tích, biến đổi và trích xuất thông tin từ ảnh số. Với sự phát triển mạnh mẽ của các thiết bị thu nhận ảnh như máy ảnh số, camera giám sát, thiết bị di động và hệ thống cảm biến, xử lý ảnh ngày càng đóng vai trò then chốt trong nhiều lĩnh vực ứng dụng như nhận dạng khuôn mặt, y sinh, giám sát an ninh, xe tự hành và giải trí số [1].

Về mặt toán học, một ảnh số có thể được biểu diễn như một hàm hai chiều:

$$f(x, y)$$

trong đó x và y là các tọa độ không gian rời rạc, còn $f(x, y)$ biểu diễn cường độ sáng (intensity) hoặc mức xám của điểm ảnh (pixel) tại vị trí tương ứng. Đối với ảnh màu, mỗi điểm ảnh thường được biểu diễn bởi một vector nhiều chiều, phổ biến nhất là trong không gian màu RGB với ba thành phần R , G và B [1].

Ảnh số thực chất là kết quả của quá trình lấy mẫu (sampling) và lượng tử hóa (quantization) ảnh liên tục trong thế giới thực. Độ phân giải không gian của ảnh phụ thuộc vào mật độ lấy mẫu, trong khi độ phân giải cường độ phụ thuộc vào số mức xám được sử dụng để biểu diễn giá trị $f(x, y)$. Hai yếu tố này ảnh hưởng trực tiếp đến chất lượng ảnh cũng như khả năng xử lý và phân tích về sau.

Các bài toán trong xử lý ảnh số có thể được phân thành nhiều nhóm khác nhau. Nhóm đầu tiên là các bài toán tiền xử lý ảnh, bao gồm lọc nhiễu, tăng cường ảnh, cân bằng histogram và chuẩn hóa ảnh. Mục tiêu của các kỹ thuật này là cải thiện chất lượng ảnh đầu vào nhằm phục vụ tốt hơn cho các bước phân tích tiếp theo.

Nhóm thứ hai là các bài toán phân tích ảnh, bao gồm phân đoạn ảnh, phát hiện biên, trích xuất đặc trưng và mô tả đối tượng. Các bài toán này nhằm chuyển đổi ảnh từ dạng dữ liệu mức thấp (pixel) sang các biểu diễn mức cao hơn, phản ánh cấu trúc và nội dung ngữ nghĩa của ảnh.

Ở mức cao hơn, xử lý ảnh thường gắn liền với các bài toán nhận dạng và hiểu ảnh, chẳng hạn như nhận dạng khuôn mặt, nhận dạng đối tượng, ước lượng tư thế hay phân tích biểu cảm. Đây là những bài toán phức tạp, đòi hỏi sự kết hợp giữa xử lý ảnh, học máy và trí tuệ nhân tạo.

Trong giai đoạn đầu, các phương pháp xử lý ảnh chủ yếu dựa trên các kỹ thuật thủ công (hand-crafted features), sử dụng các toán tử tuyến tính, phi tuyến và các mô hình toán học cổ điển. Mặc dù hiệu quả trong một số bài toán cụ thể, các phương pháp này thường gặp khó khăn khi đối mặt với sự đa dạng lớn của dữ liệu ảnh trong thực tế.

Sự phát triển của học sâu (Deep Learning), đặc biệt là các mạng nơ-ron tích chập (Convolutional Neural Networks – CNN), đã tạo ra bước đột phá lớn trong lĩnh vực xử lý ảnh. Thay vì thiết kế đặc trưng thủ công, các mô hình học sâu có khả năng tự động học đặc trưng từ dữ liệu, cho phép đạt được hiệu suất vượt trội trong nhiều bài toán phức tạp.

Các mô hình sinh ảnh dựa trên học sâu, tiêu biểu là Mạng sinh đối kháng (Generative Adversarial Networks – GAN), đã mở ra một hướng tiếp cận mới trong xử lý ảnh, không chỉ phân tích mà còn có khả năng tổng hợp và biến đổi ảnh với mức độ chân thực cao. Đây chính là nền tảng quan trọng cho các bài toán thay đổi thuộc tính khuôn mặt được nghiên cứu trong đồ án này.

Phần này đã trình bày tổng quan về xử lý ảnh số, từ cách biểu diễn ảnh, các bài toán cơ bản cho đến sự chuyển dịch từ các phương pháp truyền thống sang các kỹ thuật dựa trên học sâu. Những kiến thức nền tảng này đóng vai trò quan trọng trong việc hiểu rõ các mô hình sinh ảnh hiện đại. Trên cơ sở đó, chương tiếp theo sẽ tập trung trình bày chi tiết về Mạng sinh đối kháng (GAN) – nền tảng cốt lõi của mô hình StarGAN được sử dụng trong đồ án.

2.2 Mạng sinh đối kháng - GAN

Mạng sinh đối kháng (Generative Adversarial Network – GAN) là một trong những mô hình học sâu tiêu biểu cho bài toán sinh dữ liệu, được Ian Goodfellow và cộng sự đề xuất lần đầu vào năm 2014 [2]. Kể từ đó, GAN đã trở thành nền tảng cho nhiều hướng nghiên cứu quan trọng trong lĩnh vực xử lý ảnh và thị giác máy tính, đặc biệt là các bài toán sinh ảnh, chuyển đổi ảnh và chỉnh sửa nội dung ảnh.

Về mặt khái niệm, GAN bao gồm hai mô hình học sâu được huấn luyện đồng thời trong một quá trình mang tính đối kháng: mô hình sinh (Generator) và mô hình phân biệt (Discriminator). Mục tiêu của Generator là học cách sinh ra các mẫu dữ liệu mới sao cho chúng có phân bố gần nhất với dữ liệu thật, trong khi Discriminator có nhiệm vụ phân biệt giữa dữ liệu thật và dữ liệu do Generator tạo ra. Hai mô hình này được huấn luyện đồng thời và cạnh tranh lẫn nhau, từ đó thúc đẩy Generator ngày càng sinh ra các mẫu dữ liệu có chất lượng cao hơn.

Về mặt toán học, quá trình huấn luyện GAN có thể được mô tả như một bài toán tối ưu hai người chơi dạng min–max. Cụ thể, hàm mục tiêu của GAN được định nghĩa như sau [2]:

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log (1 - D(G(\mathbf{z})))] ,$$

trong đó G là Generator, D là Discriminator, \mathbf{x} là mẫu dữ liệu thật, và \mathbf{z} là biến nhiễu ngẫu nhiên được lấy từ một phân bố đơn giản (thường là phân bố chuẩn hoặc phân bố đều). Mục tiêu của Discriminator là tối đa hóa khả năng phân biệt đúng, trong khi Generator cố gắng tối thiểu hóa khả năng bị Discriminator phát hiện [2].

Một ưu điểm quan trọng của GAN là khả năng học phân bố dữ liệu một cách linh hoạt mà không cần giả định rõ ràng về dạng phân bố xác suất. Nhờ đó, GAN có thể sinh ra các ảnh có độ sắc nét cao, chi tiết phong phú và mang tính chân thực vượt trội so với nhiều mô hình sinh truyền thống trước đây. Đặc biệt trong lĩnh vực xử lý ảnh, GAN đã được ứng dụng rộng rãi trong các bài toán như sinh ảnh khuôn mặt, siêu phân giải ảnh, tô màu ảnh, khôi phục ảnh và chuyển đổi phong cách ảnh.

Tuy nhiên, việc huấn luyện GAN cũng đặt ra nhiều thách thức đáng kể. Một trong những vấn đề phổ biến nhất là hiện tượng mất cân bằng trong quá trình huấn luyện, khi Discriminator trở nên quá mạnh khiến Generator không thể học hiệu quả. Bên cạnh đó, GAN còn gặp các vấn đề như không hội tụ, dao động trong quá trình huấn luyện và hiện tượng *mode collapse*, trong đó Generator chỉ sinh ra một số ít mẫu lặp lại thay vì bao phủ đầy đủ phân bố dữ liệu.

Để khắc phục các hạn chế trên, nhiều biến thể và cải tiến của GAN đã được đề xuất, bao gồm Deep Convolutional GAN (DCGAN), Wasserstein GAN (WGAN), Conditional GAN (cGAN) và các mô hình GAN đa miền [3]. Các hướng tiếp cận này đóng vai trò quan trọng trong việc cải thiện độ ổn định của quá trình huấn luyện cũng như nâng cao chất lượng dữ liệu sinh ra, đặc biệt là trong các bài toán liên quan đến chỉnh sửa và biến đổi ảnh khuôn mặt.

Nhìn chung, GAN đã tạo ra một bước ngoặt lớn trong lĩnh vực học sâu sinh dữ

liệu. Những đặc điểm và thách thức của GAN chính là nền tảng quan trọng để phát triển các mô hình nâng cao hơn, trong đó có các mô hình đa miền như StarGAN, sẽ được trình bày chi tiết trong các mục tiếp theo của đồ án.

2.3 Bài toán thay đổi thuộc tính khuôn mặt

Bài toán thay đổi thuộc tính khuôn mặt (facial attribute editing) là một bài toán quan trọng trong lĩnh vực thị giác máy tính và học sâu, với mục tiêu chỉnh sửa một hoặc nhiều thuộc tính ngữ nghĩa của khuôn mặt trong ảnh đầu vào trong khi vẫn bảo toàn các đặc trưng nhận dạng và cấu trúc tổng thể của khuôn mặt đó. Các thuộc tính thường được quan tâm bao gồm màu tóc, kiểu tóc, giới tính, độ tuổi, trạng thái biểu cảm, việc đeo kính, trang điểm, hoặc các đặc trưng mang tính phong cách khác. Bài toán này có mối liên hệ chặt chẽ với các bài toán dịch ảnh sang ảnh (image-to-image translation), tuy nhiên mang tính đặc thù cao do yêu cầu khắt khe về tính nhất quán danh tính và tính tự nhiên của ảnh sinh ra.

Về bản chất, thay đổi thuộc tính khuôn mặt không chỉ là việc biến đổi các pixel của ảnh mà còn là quá trình điều khiển các yếu tố ngữ nghĩa ở mức cao. Một mô hình giải quyết tốt bài toán này cần đảm bảo ba yêu cầu quan trọng: (i) thuộc tính mục tiêu phải được thể hiện rõ ràng và chính xác trong ảnh sinh, (ii) các thuộc tính không liên quan không bị thay đổi một cách không mong muốn, và (iii) ảnh kết quả phải có chất lượng thị giác cao, trông tự nhiên và không xuất hiện các hiện tượng méo mó hay nhiễu thị giác. Việc đồng thời thỏa mãn cả ba yêu cầu này tạo ra thách thức lớn đối với các mô hình học sâu, đặc biệt khi số lượng thuộc tính cần xử lý ngày càng tăng.

Trong các nghiên cứu ban đầu, bài toán thay đổi thuộc tính khuôn mặt thường được tiếp cận bằng các mô hình sinh ảnh có điều kiện (conditional generative models). Các phương pháp này sử dụng nhãn thuộc tính như một điều kiện đầu vào cho mô hình sinh, cho phép điều khiển kết quả sinh ra theo thuộc tính mong muốn. Tuy nhiên, phần lớn các phương pháp truyền thống chỉ tập trung vào việc sinh ảnh mới từ nhiều hoặc từ ảnh đầu vào đã được mã hóa, mà chưa giải quyết triệt để vấn đề bảo toàn thông tin gốc của khuôn mặt, đặc biệt là danh tính của đối tượng trong ảnh.

Sự ra đời của mạng sinh đối kháng (GAN) đã mở ra một hướng tiếp cận mới đầy triển vọng cho bài toán này. GAN cho phép học trực tiếp phân phối dữ liệu ảnh khuôn mặt và sinh ra các ảnh có chất lượng cao, tiệm cận ảnh thật. Dựa trên nền tảng này, nhiều mô hình thay đổi thuộc tính khuôn mặt đã được đề xuất, tiêu biểu là AttGAN – một trong những công trình có ảnh hưởng lớn trong lĩnh vực này. AttGAN được thiết kế với triết lý “chỉ thay đổi những gì cần thay đổi”, trong đó mô

hình được huấn luyện để chỉnh sửa thuộc tính mục tiêu trong khi duy trì tối đa các đặc trưng khác của ảnh gốc.

AttGAN sử dụng một kiến trúc encoder-decoder kết hợp với bộ phân loại thuộc tính và bộ phân biệt đối kháng. Trong quá trình huấn luyện, mô hình đồng thời tối ưu ba thành phần: ràng buộc phân loại thuộc tính nhằm đảm bảo ảnh sinh mang thuộc tính mong muốn, ràng buộc tái tạo nhằm giữ lại các thông tin không liên quan tới thuộc tính cần chỉnh sửa, và ràng buộc đối kháng nhằm nâng cao chất lượng thị giác của ảnh sinh. Cách tiếp cận này giúp AttGAN đạt được kết quả tốt trong nhiều kịch bản chỉnh sửa thuộc tính đơn lẻ hoặc kết hợp nhiều thuộc tính [4].

Mặc dù đạt được nhiều kết quả khả quan, các phương pháp như AttGAN vẫn gặp phải những hạn chế nhất định. Một trong những vấn đề nổi bật là hiện tượng “rò rỉ thuộc tính”, trong đó việc chỉnh sửa một thuộc tính có thể vô tình làm thay đổi các thuộc tính khác có liên quan về mặt ngữ nghĩa. Ví dụ, việc chuyển đổi giới tính có thể kéo theo sự thay đổi không mong muốn về kiểu tóc hoặc trang điểm. Ngoài ra, khi chỉnh sửa nhiều thuộc tính cùng lúc, mô hình có thể gặp khó khăn trong việc cân bằng mức độ ảnh hưởng của từng thuộc tính, dẫn đến kết quả không ổn định.

Để khắc phục các hạn chế trên, một số nghiên cứu gần đây đã đề xuất sử dụng thông tin không gian và ngữ nghĩa để hướng dẫn quá trình chỉnh sửa. Tiêu biểu trong số đó là MagGAN, mô hình khai thác mặt nạ ngữ nghĩa (semantic mask) nhằm xác định chính xác vùng ảnh liên quan đến thuộc tính cần chỉnh sửa [5]. Bằng cách tập trung sự thay đổi vào các vùng quan trọng và hạn chế tác động lên các vùng khác, MagGAN cho phép sinh ra các ảnh có độ phân giải cao hơn và bảo toàn cấu trúc khuôn mặt tốt hơn so với các phương pháp trước đó.

Bên cạnh đó, một hướng tiếp cận quan trọng khác trong bài toán thay đổi thuộc tính khuôn mặt là việc xử lý đồng thời nhiều thuộc tính trong một mô hình duy nhất. Điều này đặc biệt quan trọng trong các ứng dụng thực tế, nơi người dùng thường có nhu cầu chỉnh sửa nhiều đặc trưng cùng lúc. Tuy nhiên, việc huấn luyện một mô hình có khả năng xử lý đa thuộc tính đặt ra yêu cầu cao về kiến trúc mạng, chiến lược huấn luyện cũng như dữ liệu huấn luyện đủ đa dạng. Các nghiên cứu chỉ ra rằng mỗi quan hệ phụ thuộc và tương tác giữa các thuộc tính là yếu tố then chốt ảnh hưởng đến chất lượng ảnh sinh trong các kịch bản đa thuộc tính.

Từ góc độ ứng dụng, bài toán thay đổi thuộc tính khuôn mặt có tiềm năng ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau. Trong ngành công nghiệp giải trí và truyền thông, các công nghệ này được sử dụng để chỉnh sửa ảnh chân dung, tạo hiệu ứng hình ảnh hoặc hỗ trợ sản xuất nội dung số. Trong lĩnh vực an ninh và pháp y, thay đổi thuộc tính khuôn mặt có thể hỗ trợ mô phỏng ngoại hình theo độ tuổi

hoặc đặc điểm mô tả. Ngoài ra, bài toán này cũng đóng vai trò quan trọng trong nghiên cứu về biểu diễn ngữ nghĩa của hình ảnh và khả năng kiểm soát quá trình sinh dữ liệu của các mô hình học sâu.

Tóm lại, bài toán thay đổi thuộc tính khuôn mặt là một bài toán phức tạp, đòi hỏi sự kết hợp giữa mô hình hóa ngữ nghĩa, bảo toàn thông tin gốc và đảm bảo chất lượng thị giác. Mặc dù đã có nhiều công trình nghiên cứu đạt được những kết quả đáng ghi nhận, vẫn còn tồn tại nhiều thách thức cần tiếp tục được giải quyết, đặc biệt trong bối cảnh chỉnh sửa đa thuộc tính và ứng dụng thực tế. Những vấn đề này cũng chính là động lực thúc đẩy sự ra đời của các mô hình tổng quát hơn như StarGAN, sẽ được trình bày chi tiết trong phần tiếp theo của chương này.

2.4 Mô hình StarGAN và các nghiên cứu liên quan

2.4.1 Tổng quan

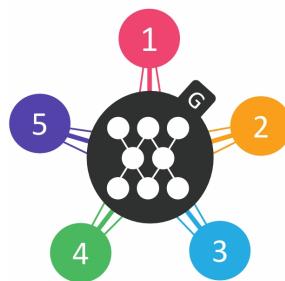
Như đã trình bày trong Mục 2.3, bài toán thay đổi thuộc tính khuôn mặt đặt ra yêu cầu cao về khả năng chỉnh sửa có kiểm soát các thuộc tính mong muốn, đồng thời bảo toàn các đặc trưng không liên quan trong ảnh gốc. Trong bối cảnh đó, các mô hình dựa trên GAN đã chứng minh hiệu quả vượt trội, đặc biệt là các mô hình dịch ảnh sang ảnh. Tuy nhiên, phần lớn các phương pháp ban đầu chỉ tập trung vào việc xử lý một cặp miền hoặc một thuộc tính cụ thể, dẫn đến hạn chế lớn về khả năng mở rộng. Mô hình StarGAN được đề xuất nhằm giải quyết trực tiếp vấn đề này bằng cách thống nhất việc học đa miền trong một kiến trúc duy nhất.

Trước khi StarGAN ra đời, nhiều nghiên cứu về dịch ảnh sang ảnh đa miền đã được đề xuất, tiêu biểu là CycleGAN và các biến thể của nó [6]. Mặc dù CycleGAN cho phép học ảnh xạ giữa hai miền ảnh không có cặp tương ứng, mô hình này vẫn yêu cầu huấn luyện riêng cho từng cặp miền. Khi số lượng miền tăng lên, số lượng mô hình cần huấn luyện tăng theo bậc hai, dẫn đến chi phí huấn luyện và lưu trữ rất lớn.

Trong bối cảnh bài toán thay đổi thuộc tính khuôn mặt, mỗi thuộc tính (ví dụ: tóc vàng, đeo kính, râu mép) có thể được xem là một miền ảnh riêng biệt. Do đó, nếu áp dụng trực tiếp các mô hình kiểu CycleGAN, số lượng mạng cần huấn luyện sẽ tăng nhanh theo số thuộc tính, gây khó khăn lớn trong thực tế triển khai. Ngoài ra, các mô hình này thường không tận dụng được mối quan hệ tiềm ẩn giữa các thuộc tính, dẫn đến hiệu quả học chưa tối ưu.

2.4.2 Ý tưởng của StarGAN

StarGAN (Star-shaped Generative Adversarial Network) được Choi và cộng sự đề xuất nhằm giải quyết bài toán dịch ảnh đa miền trong một mô hình duy nhất



Hình 2.1: Hình minh họa StarGAN thể hiện cấu trúc hình sao kết nối đa miền [7]

[7]. Thay vì huấn luyện nhiều mạng cho từng cặp miền, StarGAN sử dụng một generator và một discriminator chung, trong đó thông tin về miền đích được đưa trực tiếp vào quá trình sinh ảnh thông qua vector thuộc tính.

Ý tưởng trung tâm của StarGAN là mô hình hóa tất cả các miền ảnh trong một không gian chung, và điều khiển quá trình sinh ảnh bằng cách thay đổi vector điều kiện biểu diễn miền mong muốn (Hình 2.1). Cách tiếp cận này không chỉ giúp giảm đáng kể số lượng mô hình cần huấn luyện mà còn cho phép mô hình học được các đặc trưng chung giữa nhiều miền khác nhau, từ đó cải thiện khả năng tổng quát hóa.

Kiến trúc StarGAN bao gồm hai thành phần chính: bộ sinh (Generator) và bộ phân biệt (Discriminator). Bộ sinh nhận đầu vào là ảnh khuôn mặt và một vector biểu diễn thuộc tính đích, sau đó sinh ra ảnh mới có các thuộc tính tương ứng. Bộ phân biệt vừa đóng vai trò phân biệt ảnh thật và ảnh giả, vừa thực hiện nhiệm vụ phân loại miền của ảnh đầu vào.

Cụ thể, generator G nhận đầu vào là ảnh x và vector thuộc tính mục tiêu c , sinh ra ảnh $G(x, c)$. Vector c thường được biểu diễn dưới dạng one-hot hoặc multi-hot, tùy theo số lượng thuộc tính được xét. Discriminator D được thiết kế với hai nhánh đầu ra: một nhánh D_{adv} để phân biệt ảnh thật/giả và một nhánh D_{cls} để dự đoán thuộc tính của ảnh. Cách thiết kế này cho phép mô hình vừa học được phân phối ảnh, vừa học được mối quan hệ giữa ảnh và thuộc tính.

2.4.3 Các hàm mất mát trong StarGAN

StarGAN sử dụng nhiều thành phần hàm mất mát để đảm bảo chất lượng ảnh sinh ra cũng như tính chính xác của thuộc tính [7].

Trước hết, hàm mất mát đối kháng (adversarial loss) được sử dụng để đảm bảo ảnh sinh có phân phối gần với ảnh thật. Thành phần này giúp generator học cách tạo ra các ảnh có chất lượng thị giác cao và khó bị discriminator phân biệt.

Bên cạnh đó, StarGAN sử dụng hàm mất mát phân loại thuộc tính. Discriminator

được huấn luyện để dự đoán đúng thuộc tính của ảnh thật, trong khi generator được huấn luyện để tạo ra ảnh giả sao cho discriminator dự đoán đúng thuộc tính mục tiêu. Thành phần này đóng vai trò then chốt trong việc đảm bảo ảnh sinh ra phản ánh chính xác thuộc tính mong muốn.

Cuối cùng, StarGAN áp dụng hàm mất mát tái tạo (cycle consistency loss) nhằm bảo toàn nội dung không liên quan đến thuộc tính cần chỉnh sửa. Cụ thể, sau khi chuyển ảnh từ miền gốc sang miền đích, generator tiếp tục chuyển ngược ảnh đó về miền ban đầu, và sự khác biệt giữa ảnh ban đầu và ảnh tái tạo được tối thiểu hóa. Điều này giúp hạn chế hiện tượng biến dạng không mong muốn và giữ lại các đặc trưng nhận dạng của khuôn mặt.

2.4.4 Ưu điểm của StarGAN trong bài toán thay đổi thuộc tính khuôn mặt

Một trong những ưu điểm nổi bật của StarGAN là khả năng xử lý nhiều thuộc tính trong một mô hình duy nhất. Điều này đặc biệt phù hợp với các tập dữ liệu khuôn mặt phổ biến như CelebA, nơi mỗi ảnh được gán nhãn nhiều thuộc tính cùng lúc. Nhờ kiến trúc thống nhất, StarGAN cho phép thay đổi linh hoạt một hoặc nhiều thuộc tính chỉ bằng cách điều chỉnh vector điều kiện đầu vào.

Ngoài ra, việc chia sẻ tham số giữa các miền giúp mô hình học được các đặc trưng chung, từ đó cải thiện hiệu quả huấn luyện và giảm nguy cơ quá khớp. So với các mô hình huấn luyện riêng lẻ cho từng thuộc tính, StarGAN có khả năng tổng quát hóa tốt hơn và dễ dàng mở rộng khi bổ sung thuộc tính mới.

2.4.5 Các biến thể và nghiên cứu mở rộng từ StarGAN

Sau khi StarGAN được công bố, nhiều nghiên cứu đã tiếp tục mở rộng và cải tiến mô hình này. StarGAN v2 được đề xuất nhằm khắc phục hạn chế về tính đa dạng của ảnh sinh, bằng cách đưa vào khái niệm style embedding và cho phép sinh ra nhiều kết quả khác nhau cho cùng một thuộc tính [8]. Thay vì chỉ học ảnh xạ giữa các miền, StarGAN v2 còn học phân phối phong cách trong từng miền, giúp ảnh sinh đa dạng và tự nhiên hơn.

Bên cạnh đó, một số nghiên cứu kết hợp StarGAN với các ràng buộc ngữ nghĩa hoặc mặt nạ không gian để kiểm soát tốt hơn vùng ảnh bị chỉnh sửa. Những hướng tiếp cận này đặc biệt hữu ích trong việc giảm nhiễu và tránh thay đổi không mong muốn ở các vùng không liên quan đến thuộc tính mục tiêu.

2.4.6 Đánh giá tổng quan

Từ góc độ nghiên cứu và ứng dụng, StarGAN được xem là một trong những mô hình nền tảng cho bài toán thay đổi thuộc tính khuôn mặt đa miền. Mô hình này

không chỉ giải quyết hiệu quả vấn đề mở rộng mà còn đặt nền móng cho nhiều hướng nghiên cứu tiếp theo. Tuy nhiên, StarGAN vẫn tồn tại một số hạn chế, chẳng hạn như chất lượng ảnh suy giảm khi chỉnh sửa nhiều thuộc tính đồng thời hoặc khó kiểm soát mức độ thay đổi của từng thuộc tính.

Dù vậy, với sự cân bằng hợp lý giữa độ phức tạp, hiệu quả và khả năng mở rộng, StarGAN vẫn là lựa chọn phù hợp cho nhiều bài toán thực tế. Trong phạm vi đồ án này, StarGAN được lựa chọn làm mô hình cốt lõi để triển khai và đánh giá, đồng thời làm nền tảng cho việc xây dựng ứng dụng thử nghiệm trong các chương tiếp theo.

Chương này đã trình bày chi tiết mô hình StarGAN, từ bối cảnh ra đời, kiến trúc, các hàm mất mát cho đến những ưu điểm và hướng mở rộng liên quan. Qua đó có thể thấy StarGAN là một bước tiến quan trọng trong bài toán thay đổi thuộc tính khuôn mặt đa miền. Những phân tích lý thuyết trong chương này là cơ sở để đi vào phần phân tích thiết kế hệ thống và pipeline huấn luyện mô hình, sẽ được trình bày trong Chương ??.

2.5 Các chỉ số đánh giá chất lượng ảnh sinh

Đánh giá chất lượng ảnh sinh là một vấn đề quan trọng và không hề đơn giản trong các mô hình sinh ảnh nói chung và các mô hình GAN nói riêng. Khác với các bài toán phân loại hay hồi quy, đầu ra của bài toán sinh ảnh không có một nhãn “đúng” duy nhất để so sánh trực tiếp. Do đó, việc đánh giá chất lượng ảnh sinh thường dựa trên các chỉ số thống kê hoặc các độ đo phản ánh mức độ tương đồng giữa ảnh sinh và ảnh thật, cả về mặt cấu trúc lẫn phân bố dữ liệu. Trong phạm vi đồ án này, hai chỉ số được sử dụng để đánh giá mô hình là chỉ số tương đồng cấu trúc SSIM và chỉ số khoảng cách phân bố FID, đây cũng là hai chỉ số phổ biến và có ý nghĩa thực tiễn cao trong lĩnh vực sinh ảnh.

2.5.1 Chỉ số tương đồng cấu trúc SSIM

Chỉ số SSIM được đề xuất nhằm đánh giá mức độ tương đồng giữa hai ảnh dựa trên cách con người cảm nhận chất lượng ảnh, thay vì chỉ so sánh sai khác điểm ảnh đơn thuần. Ý tưởng cốt lõi của SSIM là cho rằng hệ thống thị giác của con người nhạy cảm hơn với sự thay đổi về cấu trúc của ảnh so với các sai khác tuyệt đối về cường độ sáng.

Về mặt toán học, SSIM đánh giá sự tương đồng giữa hai ảnh x và y thông qua ba thành phần: độ sáng (luminance), độ tương phản (contrast) và cấu trúc (structure). Chỉ số SSIM được định nghĩa như sau [9]:

$$\text{SSIM}(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma$$

trong đó $l(x, y)$, $c(x, y)$ và $s(x, y)$ lần lượt biểu diễn độ tương đồng về độ sáng, độ tương phản và cấu trúc giữa hai ảnh. Trong thực tế, các tham số α , β , γ thường được đặt bằng 1 để đơn giản hóa công thức.

Cụ thể, các thành phần này được xác định như sau [9]:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

trong đó μ_x, μ_y là giá trị trung bình, σ_x, σ_y là độ lệch chuẩn và σ_{xy} là hiệp phương sai của hai ảnh x và y . Các hằng số C_1, C_2, C_3 được đưa vào nhằm tránh hiện tượng chia cho 0 khi các mẫu có giá trị nhỏ.

Giá trị của SSIM nằm trong khoảng $[-1, 1]$, tuy nhiên trong thực tế thường nằm trong khoảng $[0, 1]$, với giá trị càng gần 1 thì hai ảnh càng tương đồng về mặt cấu trúc. Trong bài toán thay đổi thuộc tính khuôn mặt, SSIM thường được sử dụng để đánh giá mức độ bảo toàn danh tính và cấu trúc khuôn mặt gốc sau khi thực hiện biến đổi thuộc tính. Một mô hình tốt cần tạo ra ảnh có SSIM cao so với ảnh đầu vào, cho thấy rằng các đặc trưng không liên quan đến thuộc tính cần thay đổi vẫn được giữ lại.

Tuy nhiên, SSIM cũng tồn tại một số hạn chế. Chỉ số này yêu cầu ảnh sinh và ảnh gốc phải có sự tương ứng từng cặp, do đó không phản ánh được chất lượng tổng thể của phân bố ảnh sinh. Ngoài ra, SSIM không đánh giá được mức độ đa dạng của ảnh sinh, mà chỉ tập trung vào sự tương đồng cục bộ giữa hai ảnh cụ thể.

2.5.2 Chỉ số Fréchet Inception Distance (FID)

Khác với SSIM, chỉ số FID (Fréchet Inception Distance) được thiết kế để đánh giá sự tương đồng giữa hai phân bố ảnh: phân bố ảnh thật và phân bố ảnh sinh. FID được xem là một trong những chỉ số đánh giá chất lượng ảnh sinh hiệu quả và đáng tin cậy nhất hiện nay, đặc biệt trong các nghiên cứu về GAN.

Ý tưởng chính của FID là ánh xạ ảnh vào không gian đặc trưng có ý nghĩa ngữ nghĩa cao bằng cách sử dụng một mạng Inception v3 đã được huấn luyện trước trên tập ImageNet. Các đặc trưng trích xuất từ lớp trung gian của mạng này được giả định tuân theo phân bố Gaussian đa biến. Khi đó, khoảng cách Fréchet giữa hai

phân bố Gaussian tương ứng với ảnh thật và ảnh sinh được tính như sau [10]:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr} \left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2} \right)$$

trong đó (μ_r, Σ_r) và (μ_g, Σ_g) lần lượt là vector trung bình và ma trận hiệp phương sai của đặc trưng ảnh thật và ảnh sinh.

Giá trị FID càng nhỏ thì phân bố ảnh sinh càng gần với phân bố ảnh thật, đồng nghĩa với việc ảnh sinh có chất lượng cao và đa dạng tốt. Trong bài toán thay đổi thuộc tính khuôn mặt, FID không chỉ phản ánh mức độ chân thực của ảnh sinh mà còn gián tiếp đánh giá khả năng mô hình học được phân bố dữ liệu khuôn mặt thực tế.

Một ưu điểm lớn của FID là khả năng đánh giá đồng thời cả chất lượng và độ đa dạng của ảnh sinh, điều mà các chỉ số dựa trên so sánh từng cặp ảnh như SSIM không làm được. Tuy nhiên, FID cũng có một số hạn chế nhất định. Kết quả FID phụ thuộc vào số lượng mẫu đánh giá và có thể bị ảnh hưởng bởi nhiều thông kê khi tập mẫu quá nhỏ. Ngoài ra, việc sử dụng mạng Inception được huấn luyện trên ImageNet cũng có thể chưa hoàn toàn phù hợp với các miền dữ liệu đặc thù như khuôn mặt.

Trong đồ án này, SSIM và FID được sử dụng kết hợp nhằm đánh giá mô hình một cách toàn diện. SSIM phản ánh mức độ bảo toàn cấu trúc và danh tính khuôn mặt khi thay đổi thuộc tính, trong khi FID đánh giá chất lượng tổng thể và độ chân thực của ảnh sinh ở cấp độ phân bố. Việc phân tích đồng thời hai chỉ số này cho phép đưa ra nhận định khách quan hơn về hiệu quả của mô hình StarGAN trong bài toán thay đổi thuộc tính khuôn mặt.

CHƯƠNG 3. PHÂN TÍCH THIẾT KẾ HỆ THỐNG

Chương 2 đã trình bày cơ sở lý thuyết liên quan đến xử lý ảnh số, mạng sinh đối kháng GAN, bài toán thay đổi thuộc tính khuôn mặt, mô hình StarGAN cũng như các chỉ số đánh giá chất lượng ảnh sinh ra. Trên cơ sở các kiến thức nền tảng đó, chương này tập trung vào việc phân tích và thiết kế hệ thống cho bài toán thay đổi thuộc tính khuôn mặt dựa trên mô hình StarGAN.

Mục tiêu của Chương 3 là mô tả một cách có hệ thống kiến trúc tổng thể, các thành phần chức năng chính và luồng xử lý dữ liệu trong toàn bộ hệ thống, từ giai đoạn huấn luyện mô hình cho đến giai đoạn sinh ảnh, đánh giá và triển khai ứng dụng. Nội dung chương đóng vai trò cầu nối giữa phần lý thuyết và các chương triển khai, thực nghiệm ở phía sau.

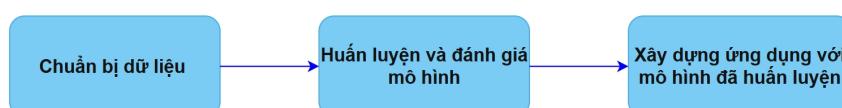
3.1 Kiến trúc tổng thể của hệ thống

3.1.1 Mô tả tổng quan

Hệ thống được xây dựng trong đồ án này nhằm mục tiêu thực hiện bài toán thay đổi thuộc tính khuôn mặt một cách linh hoạt, cho phép chỉnh sửa nhiều thuộc tính khác nhau trên cùng một ảnh đầu vào, đồng thời đảm bảo chất lượng ảnh sinh ra đạt mức chấp nhận được cả về mặt thị giác lẫn các chỉ số định lượng. Để đáp ứng yêu cầu đó, kiến trúc hệ thống được thiết kế theo hướng phân tầng, tách biệt rõ ràng giữa các khối chức năng, giúp việc huấn luyện, đánh giá và triển khai mô hình được thực hiện một cách độc lập nhưng vẫn đảm bảo tính liên kết tổng thể. Về mặt tổng quát, hệ thống bao gồm ba khối chính: (i) khối chuẩn bị dữ liệu và tiền xử lý, (ii) khối huấn luyện và đánh giá mô hình StarGAN, và (iii) khối ứng dụng và triển khai mô hình đã huấn luyện (Hình 3.1). Mỗi khối đảm nhiệm một vai trò riêng biệt trong toàn bộ pipeline xử lý, từ dữ liệu thô ban đầu cho đến kết quả cuối cùng là ảnh khuôn mặt đã được thay đổi thuộc tính theo yêu cầu của người dùng.

3.1.2 Khối chuẩn bị dữ liệu và tiền xử lý

Khối chuẩn bị dữ liệu đóng vai trò nền tảng cho toàn bộ hệ thống. Đầu vào của khối này là tập dữ liệu ảnh khuôn mặt kèm theo nhãn thuộc tính, điển hình là các tập dữ liệu phổ biến như CelebA. Dữ liệu thô ban đầu thường có kích thước, tỷ lệ và chất lượng không đồng nhất, do đó cần được tiền xử lý trước khi đưa vào huấn



Hình 3.1: Tổng quan các khối

luyện mô hình.

Các bước tiền xử lý chính bao gồm chuẩn hóa kích thước ảnh, căn chỉnh khuôn mặt, chuẩn hóa giá trị pixel và chuyển đổi nhãn thuộc tính về dạng vector nhị phân phù hợp với đầu vào của mô hình StarGAN. Ngoài ra, dữ liệu cũng được chia thành các tập huấn luyện, tập kiểm tra và tập đánh giá nhằm đảm bảo tính khách quan trong quá trình huấn luyện và đánh giá mô hình.

Việc thiết kế khôi tiền xử lý độc lập giúp hệ thống dễ dàng mở rộng sang các tập dữ liệu khác trong tương lai mà không cần thay đổi cấu trúc của các khôi xử lý phía sau.

3.1.3 Khối huấn luyện mô hình StarGAN

Khối huấn luyện là thành phần cốt lõi của hệ thống, nơi mô hình StarGAN được triển khai và tối ưu hóa. Khối này bao gồm mạng sinh (Generator), mạng phân biệt (Discriminator) và các hàm mất mát tương ứng đã được trình bày trong Chương 2. Dữ liệu sau tiền xử lý được đưa vào khối huấn luyện dưới dạng các cặp ảnh và vector thuộc tính mục tiêu.

Trong quá trình huấn luyện, mô hình học cách sinh ra ảnh khuôn mặt mới sao cho vừa mang các thuộc tính mong muốn, vừa giữ được các đặc trưng nhận dạng quan trọng của ảnh gốc. Các checkpoint của mô hình được lưu lại định kỳ theo số vòng lặp huấn luyện (iteration) nhằm phục vụ cho việc đánh giá, so sánh và lựa chọn mô hình tối ưu.

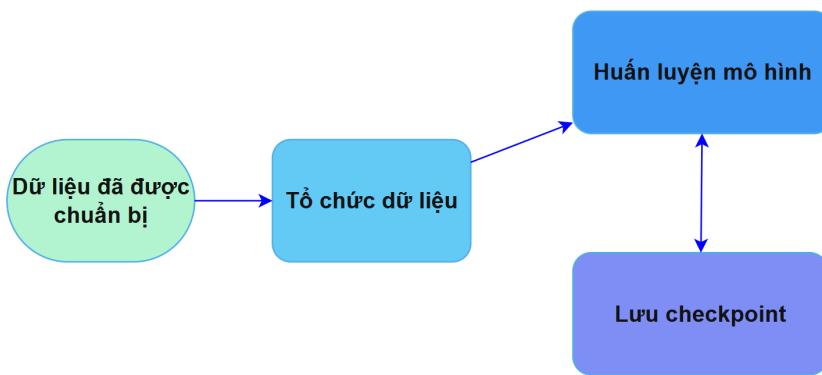
Khối huấn luyện được thiết kế để có thể chạy độc lập trên các nền tảng tính toán hiệu năng cao như GPU hoặc môi trường đám mây, điển hình là Kaggle hoặc Google Colab, giúp rút ngắn thời gian huấn luyện và thuận tiện cho việc thử nghiệm nhiều cấu hình khác nhau.

3.1.4 Khối ứng dụng và triển khai

Khối ứng dụng là lớp trên cùng của hệ thống, trực tiếp tương tác với người dùng cuối. Khối này sử dụng mô hình StarGAN đã được lựa chọn từ giai đoạn đánh giá để thực hiện sinh ảnh theo yêu cầu. Người dùng có thể cung cấp ảnh đầu vào và lựa chọn các thuộc tính cần thay đổi, hệ thống sẽ trả về ảnh kết quả tương ứng.

Trong đồ án này, khối ứng dụng được triển khai dưới nhiều hình thức khác nhau, bao gồm chatbot và giao diện web, nhằm chứng minh khả năng ứng dụng thực tế của mô hình đã huấn luyện. Việc tách biệt rõ ràng khối ứng dụng khỏi khối huấn luyện giúp hệ thống có thể dễ dàng bảo trì, nâng cấp hoặc thay thế mô hình mà không ảnh hưởng đến giao diện người dùng.

Tóm lại, kiến trúc tổng thể của hệ thống được thiết kế theo hướng mô-đun, phân



Hình 3.2: Các giai đoạn trong quá trình huấn luyện

tách rõ ràng các khối chức năng từ dữ liệu, huấn luyện, đánh giá cho đến triển khai ứng dụng. Cách tiếp cận này không chỉ giúp hệ thống hoạt động ổn định và dễ mở rộng mà còn tạo tiền đề thuận lợi cho việc phân tích, đánh giá và cải tiến mô hình trong các chương tiếp theo. Trên cơ sở kiến trúc đã đề xuất, các pipeline huấn luyện và sinh ảnh cụ thể sẽ được trình bày chi tiết trong các mục tiếp theo của Chương 3.

3.2 Thiết kế pipeline huấn luyện mô hình

Tiếp nối kiến trúc tổng thể của hệ thống đã được trình bày trong Mục 3.1, phần này tập trung mô tả chi tiết pipeline huấn luyện mô hình StarGAN được sử dụng trong đồ án. Pipeline huấn luyện đóng vai trò trung tâm trong toàn bộ hệ thống, quyết định trực tiếp đến chất lượng ảnh sinh ra cũng như khả năng thay đổi chính xác các thuộc tính khuôn mặt theo yêu cầu. Việc thiết kế pipeline một cách hợp lý giúp đảm bảo quá trình huấn luyện ổn định, có khả năng mở rộng, thuận tiện cho việc đánh giá và so sánh các checkpoint trong các giai đoạn khác nhau.

3.2.1 Tổng quan pipeline huấn luyện

Pipeline huấn luyện mô hình trong đồ án được thiết kế theo hướng tuần tự và khép kín, bao gồm các bước chính: tổ chức lại dữ liệu đã chuẩn bị, tiền xử lý và mã hóa nhãn thuộc tính, huấn luyện đồng thời hai mạng sinh (Generator) và phân biệt (Discriminator), lưu checkpoint định kỳ, và ghi nhận các thông tin phục vụ đánh giá mô hình (Hình 3.2). Mỗi bước trong pipeline đều được xây dựng với mục tiêu đảm bảo tính nhất quán giữa dữ liệu, nhãn và mô hình, đồng thời hỗ trợ việc phân tích kết quả sau huấn luyện.

Khác với các pipeline huấn luyện GAN truyền thống chỉ tập trung vào việc sinh ảnh, pipeline trong đồ án được mở rộng để phục vụ bài toán thay đổi nhiều thuộc tính khuôn mặt trên cùng một mô hình duy nhất. Do đó, ngoài mục tiêu sinh ảnh chân thực, pipeline còn phải đảm bảo rằng các thuộc tính mục tiêu được thay đổi chính xác, trong khi các đặc trưng không liên quan của khuôn mặt được giữ nguyên.

3.2.2 Tổ chức dữ liệu huấn luyện

Dữ liệu huấn luyện được tổ chức theo cấu trúc phù hợp với yêu cầu của StarGAN, bao gồm ảnh khuôn mặt và vector nhãn biểu diễn các thuộc tính tương ứng. Mỗi ảnh đầu vào được gán một vector nhị phân, trong đó mỗi phần tử biểu thị sự có hoặc không có của một thuộc tính khuôn mặt cụ thể (ví dụ: giới tính, đeo kính, râu, mỉm cười, v.v.).

Trong pipeline huấn luyện, dữ liệu được chia thành các minibatch, mỗi minibatch bao gồm ảnh gốc và nhãn thuộc tính ban đầu. Đồng thời, một nhãn thuộc tính mục tiêu mới được sinh ngẫu nhiên cho mỗi ảnh, đại diện cho các thuộc tính mà mô hình cần thay đổi trong lần sinh ảnh đó. Cách tiếp cận này cho phép mô hình học được nhiều ánh xạ thuộc tính khác nhau trong quá trình huấn luyện, từ đó nâng cao khả năng tổng quát hóa.

Trước khi đưa vào mô hình, ảnh khuôn mặt được chuẩn hóa về cùng kích thước và không gian màu thống nhất. Việc chuẩn hóa này giúp giảm sự khác biệt không cần thiết giữa các mẫu dữ liệu, đồng thời hỗ trợ quá trình hội tụ của mô hình trong huấn luyện sâu. Các phép tiền xử lý thường bao gồm thay đổi kích thước ảnh, chuẩn hóa giá trị pixel về một khoảng cố định, và có thể áp dụng một số phép biến đổi dữ liệu cơ bản nhằm tăng tính đa dạng của tập huấn luyện.

Đối với nhãn thuộc tính, pipeline đảm bảo rằng các vector nhãn được biểu diễn nhất quán về thứ tự và ý nghĩa của từng chiều. Điều này đặc biệt quan trọng trong bối cảnh huấn luyện đa thuộc tính, bởi bất kỳ sự sai lệch nào trong ánh xạ giữa nhãn và thuộc tính đều có thể dẫn đến kết quả sinh ảnh không chính xác.

3.2.3 Huấn luyện mô hình

Trong pipeline huấn luyện, Generator và Discriminator được huấn luyện đồng thời theo cơ chế đối kháng. Ở mỗi vòng lặp, Discriminator trước tiên được cập nhật để phân biệt giữa ảnh thật và ảnh sinh, đồng thời dự đoán các thuộc tính khuôn mặt của ảnh đầu vào. Sau đó, Generator được cập nhật nhằm sinh ra ảnh mới vừa có tính chân thực cao, vừa mang các thuộc tính mục tiêu mong muốn, đồng thời đánh lừa được Discriminator.

Pipeline được thiết kế để kiểm soát số lần cập nhật của từng mạng trong mỗi vòng huấn luyện, từ đó hạn chế hiện tượng mất cân bằng trong quá trình huấn luyện GAN. Ngoài ra, các thành phần hàm mất mát khác nhau (mất mát đối kháng, mất mát phân loại thuộc tính, và mất mát tái tạo ảnh) được kết hợp có trọng số, giúp Generator học được cả tính chân thực lẫn khả năng bảo toàn nội dung khuôn mặt ban đầu.

3.2.4 Lưu checkpoint và theo dõi quá trình huấn luyện

Một điểm quan trọng trong thiết kế pipeline là cơ chế lưu checkpoint định kỳ trong suốt quá trình huấn luyện. Tại mỗi mốc số vòng lặp nhất định, trạng thái của mô hình được lưu lại, bao gồm tham số của Generator và Discriminator. Điều này không chỉ giúp phòng tránh rủi ro gián đoạn huấn luyện mà còn tạo điều kiện thuận lợi cho việc đánh giá và so sánh chất lượng mô hình tại các giai đoạn khác nhau.

Song song với việc lưu checkpoint, pipeline cũng ghi nhận các giá trị mốc mác và một số chỉ số trung gian trong quá trình huấn luyện. Các thông tin này được sử dụng để theo dõi xu hướng hội tụ của mô hình, phát hiện sớm các dấu hiệu bất ổn, và làm cơ sở cho việc lựa chọn checkpoint tối ưu trong giai đoạn đánh giá thực nghiệm ở Chương 5.

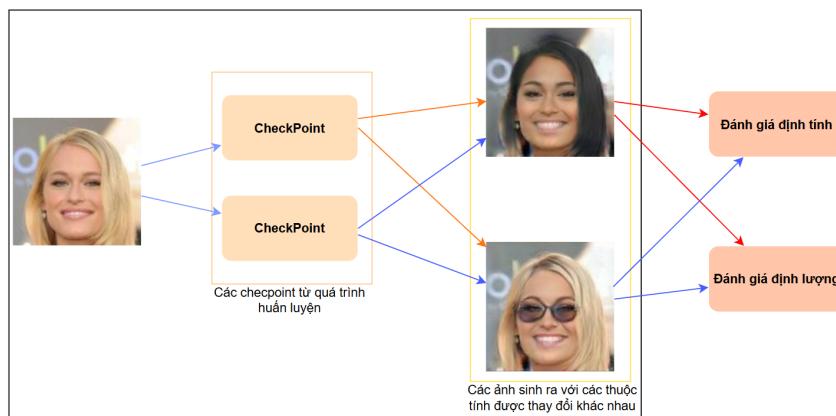
3.2.5 Tính mở rộng và khả năng tái sử dụng của pipeline

Pipeline huấn luyện trong đồ án được thiết kế theo hướng mô-đun, cho phép dễ dàng mở rộng hoặc điều chỉnh trong các nghiên cứu tiếp theo. Ví dụ, có thể thay đổi tập thuộc tính khuôn mặt, điều chỉnh cấu trúc mô hình, hoặc bổ sung các chiến lược huấn luyện nâng cao mà không cần xây dựng lại toàn bộ pipeline. Tính linh hoạt này giúp pipeline không chỉ phục vụ mục tiêu của đồ án hiện tại mà còn có giá trị tham khảo cho các nghiên cứu và ứng dụng liên quan trong tương lai.

Phần này đã trình bày chi tiết thiết kế pipeline huấn luyện mô hình StarGAN được sử dụng trong đồ án, từ khâu chuẩn bị dữ liệu, tiền xử lý, cơ chế huấn luyện đối kháng cho đến việc lưu checkpoint và theo dõi quá trình huấn luyện. Pipeline được xây dựng với mục tiêu đảm bảo tính ổn định, khả năng mở rộng và phục vụ hiệu quả cho việc đánh giá mô hình ở các chương sau. Trên cơ sở thiết kế này, Chương 4 sẽ đi sâu vào quá trình triển khai huấn luyện mô hình, bao gồm các thiết lập cụ thể và những vấn đề thực tế phát sinh trong quá trình huấn luyện.

3.3 Thiết kế pipeline sinh ảnh và đánh giá

Sau khi hoàn tất quá trình huấn luyện mô hình StarGAN như đã trình bày trong Mục 3.2, bước tiếp theo của hệ thống là thiết kế pipeline sinh ảnh và đánh giá chất lượng ảnh sinh. Pipeline này đóng vai trò then chốt trong việc kiểm tra khả năng tổng quát hóa của mô hình, cũng như đo lường một cách định lượng và định tính mức độ thành công của bài toán thay đổi thuộc tính khuôn mặt. Nội dung của mục này trình bày chi tiết quy trình sinh ảnh từ mô hình đã huấn luyện và phương pháp đánh giá kết quả sinh ảnh theo nhiều góc độ khác nhau.



Hình 3.3: Các giai đoạn của việc sinh ảnh và đánh giá

3.3.1 Pipeline sinh ảnh từ mô hình đã huấn luyện

Pipeline sinh ảnh (Mô tả trong trung đen tại Hình 3.3) được thiết kế nhằm đảm bảo hai mục tiêu chính: (i) cho phép mô hình thực hiện thay đổi từng thuộc tính khuôn mặt một cách độc lập hoặc kết hợp nhiều thuộc tính, và (ii) tạo ra tập ảnh đầu ra có cấu trúc thống nhất, thuận tiện cho việc đánh giá và so sánh giữa các checkpoint huấn luyện khác nhau.

Đầu vào của pipeline sinh ảnh bao gồm ảnh khuôn mặt gốc x và vector thuộc tính mục tiêu c_{trg} . Ảnh đầu vào được tiền xử lý theo đúng quy trình đã sử dụng trong giai đoạn huấn luyện, bao gồm thay đổi kích thước về kích thước chuẩn của mô hình, chuẩn hóa giá trị điểm ảnh về miền $[-1, 1]$, và chuyển đổi về tensor phù hợp với framework học sâu sử dụng. Vector thuộc tính mục tiêu c_{trg} được xây dựng dựa trên nhãn thuộc tính mong muốn, trong đó mỗi chiều tương ứng với một thuộc tính khuôn mặt như đã định nghĩa trong tập dữ liệu [7].

Mô hình sinh G sau khi được huấn luyện sẽ nhận cặp đầu vào (x, c_{trg}) và sinh ra ảnh kết quả $\hat{x} = G(x, c_{trg})$. Đối với bài toán thay đổi từng thuộc tính riêng lẻ, pipeline lần lượt tạo các vector c_{trg} khác nhau bằng cách thay đổi một chiều thuộc tính trong khi giữ nguyên các chiều còn lại [7]. Điều này cho phép đánh giá mức độ kiểm soát của mô hình đối với từng thuộc tính khuôn mặt cụ thể. Ngoài ra, pipeline cũng hỗ trợ sinh ảnh với nhiều thuộc tính được thay đổi đồng thời nhằm kiểm tra khả năng xử lý các tổ hợp thuộc tính phức tạp.

Ảnh đầu ra sau khi sinh được hậu xử lý bằng cách đưa giá trị điểm ảnh trở lại miền $[0, 255]$ và chuyển đổi về định dạng ảnh chuẩn để lưu trữ [7]. Các ảnh sinh được tổ chức theo cấu trúc thư mục rõ ràng, phân tách theo checkpoint huấn luyện và theo từng thuộc tính, nhằm phục vụ cho các bước đánh giá tự động ở giai đoạn tiếp theo.

3.3.2 Đánh giá định lượng

Để đánh giá chất lượng ảnh sinh một cách khách quan, pipeline đánh giá định lượng được xây dựng dựa trên các chỉ số phổ biến trong lĩnh vực sinh ảnh, cụ thể là SSIM và FID, đã được trình bày chi tiết trong Mục 2.5. Pipeline đánh giá được thiết kế theo hướng tự động hóa, cho phép xử lý số lượng lớn ảnh sinh tương ứng với nhiều checkpoint huấn luyện khác nhau.

Đối với chỉ số SSIM, pipeline tiến hành so sánh ảnh gốc x và ảnh sinh \hat{x} nhằm đánh giá mức độ bảo toàn cấu trúc khuôn mặt sau khi thay đổi thuộc tính. Việc tính SSIM được thực hiện riêng cho từng cặp ảnh và sau đó lấy giá trị trung bình trên toàn bộ tập dữ liệu. Ngoài ra, pipeline cũng cho phép tính SSIM theo từng thuộc tính để phân tích chi tiết mức độ ảnh hưởng của từng loại thay đổi lên cấu trúc khuôn mặt.

Đối với chỉ số FID, pipeline đánh giá sự khác biệt phân phối giữa tập ảnh sinh và tập ảnh thật tương ứng. Đặc trưng của ảnh được trích xuất thông qua mạng Inception được huấn luyện sẵn, sau đó sử dụng để tính toán khoảng cách Fréchet. Pipeline được thiết kế để tính FID cho từng thuộc tính riêng biệt, cũng như FID trung bình trên toàn bộ các thuộc tính, nhằm phản ánh chất lượng tổng thể của mô hình tại mỗi checkpoint huấn luyện.

Một điểm quan trọng trong thiết kế pipeline đánh giá là đảm bảo tính nhất quán giữa các checkpoint. Cùng một tập ảnh đầu vào và cùng một cấu hình đánh giá được sử dụng cho tất cả các checkpoint, giúp việc so sánh kết quả trở nên công bằng và có ý nghĩa.

3.3.3 Đánh giá định tính

Bên cạnh đánh giá định lượng, pipeline sinh ảnh và đánh giá còn hỗ trợ đánh giá định tính thông qua quan sát trực quan. Các ảnh sinh được sắp xếp theo từng thuộc tính và theo thứ tự checkpoint, cho phép người đánh giá dễ dàng so sánh mức độ thay đổi thuộc tính, tính tự nhiên của ảnh sinh và khả năng giữ nguyên các đặc điểm không liên quan của khuôn mặt.

Đánh giá định tính đóng vai trò bổ trợ quan trọng cho đánh giá định lượng, đặc biệt trong các trường hợp chỉ số số học chưa phản ánh đầy đủ cảm nhận thị giác của con người. Việc kết hợp cả hai hình thức đánh giá giúp đưa ra nhận xét toàn diện hơn về hiệu quả của mô hình StarGAN trong bài toán thay đổi thuộc tính khuôn mặt.

3.3.4 Phân tích và so sánh giữa các checkpoint huấn luyện

Dựa trên kết quả SSIM và FID thu được, pipeline tiếp tục hỗ trợ phân tích sự thay đổi chất lượng mô hình theo số vòng lặp huấn luyện. Các giá trị đánh giá được tổng hợp và lưu trữ dưới dạng bảng dữ liệu, từ đó cho phép vẽ các biểu đồ thể hiện xu hướng biến đổi của SSIM và FID theo checkpoint.

Việc phân tích theo checkpoint giúp làm rõ mối quan hệ giữa quá trình huấn luyện và chất lượng ảnh sinh, đồng thời hỗ trợ lựa chọn checkpoint tối ưu cho việc triển khai ứng dụng thực tế. Đặc biệt, pipeline được thiết kế để phát hiện các hiện tượng như overfitting hoặc suy giảm chất lượng ảnh khi huấn luyện quá lâu, thông qua việc quan sát sự biến động của các chỉ số đánh giá.

Chương này đã trình bày chi tiết thiết kế pipeline sinh ảnh và đánh giá chất lượng ảnh sinh của hệ thống. Pipeline được xây dựng theo hướng tự động, có khả năng sinh ảnh cho nhiều thuộc tính và nhiều checkpoint khác nhau, đồng thời hỗ trợ đánh giá cả định lượng lẫn định tính. Kết quả thu được từ pipeline này là cơ sở quan trọng cho các phân tích thực nghiệm và so sánh mô hình, sẽ được trình bày chi tiết trong hai chương tiếp theo.

CHƯƠNG 4. HUẤN LUYỆN MÔ HÌNH

Chương 3 đã trình bày kiến trúc tổng thể của hệ thống cũng như thiết kế các pipeline chính cho quá trình huấn luyện, sinh ảnh và đánh giá mô hình. Trên cơ sở thiết kế đó, Chương 4 tập trung vào quá trình hiện thực hoá pipeline huấn luyện trong thực tế, bao gồm việc lựa chọn và sử dụng bộ dữ liệu, thiết lập mô hình StarGAN, cấu hình các tham số huấn luyện và triển khai huấn luyện trên môi trường tính toán.

Trong chương này, trước hết đồ án trình bày về bộ dữ liệu CelebA được sử dụng cho bài toán thay đổi thuộc tính khuôn mặt, cùng với cách thức dữ liệu được tổ chức và xử lý đầu vào trong khuôn khổ mã nguồn StarGAN. Tiếp theo, quá trình thiết lập mô hình và các tham số huấn luyện sẽ được mô tả chi tiết, làm cơ sở cho các thực nghiệm và đánh giá được trình bày trong chương tiếp theo.

4.1 Chuẩn bị dữ liệu và thiết lập mô hình

4.1.1 Giới thiệu bộ dữ liệu CelebA

Trong đồ án này, mô hình StarGAN được huấn luyện và đánh giá trên bộ dữ liệu CelebA (CelebFaces Attributes Dataset), một trong những bộ dữ liệu chuẩn và phổ biến nhất trong lĩnh vực xử lý ảnh khuôn mặt và học sâu. Bộ dữ liệu CelebA được công bố bởi nhóm nghiên cứu tại Đại học Trung văn Hồng Kông và hiện được cung cấp rộng rãi thông qua nhiều kho dữ liệu công cộng, trong đó có nền tảng Kaggle¹.

CelebA bao gồm hơn 200,000 ảnh khuôn mặt của khoảng 10,000 danh tính khác nhau, được thu thập từ các nguồn ảnh trong điều kiện tự nhiên. Mỗi ảnh trong bộ dữ liệu đều được gán nhãn cho 40 thuộc tính khuôn mặt ở dạng nhị phân, phản ánh các đặc trưng hình thái và phong cách phổ biến như giới tính, kiểu tóc, tình trạng râu, trạng thái biểu cảm, phụ kiện và các yếu tố trang điểm. Các thuộc tính này được gán nhãn thủ công và đã được kiểm chứng rộng rãi trong nhiều công trình nghiên cứu trước đây. Một số đặc điểm nổi bật của bộ dữ liệu CelebA có thể kể đến như:

¹<https://www.kaggle.com/datasets/jessicali9530/celeba-dataset>

CelebFaces Attributes (CelebA) Dataset

Over 200k images of celebrities with 40 binary attribute annotations



Hình 4.1: Giới thiệu về bộ dữ liệu CelebA

- Đa dạng cao về khuôn mặt, góc nhìn, ánh sáng và biểu cảm, phản ánh tương đối tốt các điều kiện trong thế giới thực.
- Hệ thống nhãn thuộc tính phong phú, cho phép xây dựng và đánh giá các bài toán chỉnh sửa thuộc tính khuôn mặt đa miền.
- Quy mô lớn, phù hợp cho việc huấn luyện các mô hình học sâu có số lượng tham số lớn như GAN.

Nhờ những đặc điểm trên, CelebA đã trở thành bộ dữ liệu chuẩn cho nhiều nghiên cứu về sinh ảnh khuôn mặt, thay đổi thuộc tính khuôn mặt, phân tích biểu cảm và nhận dạng khuôn mặt. Việc sử dụng CelebA trong đồ án này giúp đảm bảo tính so sánh và đổi chiều với các nghiên cứu liên quan đã được công bố trước đó.

4.1.2 Lựa chọn và sử dụng tập thuộc tính

Mặc dù CelebA cung cấp tới 40 thuộc tính khuôn mặt, trong phạm vi đồ án, em đã lựa chọn một tập con gồm 17 thuộc tính tiêu biểu để huấn luyện và đánh giá mô hình StarGAN. Việc lựa chọn này dựa trên hai tiêu chí chính: (i) các thuộc tính có ý nghĩa đổi với người dùng, và (ii) các thuộc tính thường được sử dụng trong các nghiên cứu về StarGAN và các mô hình chỉnh sửa khuôn mặt đa miền.

Danh sách các thuộc tính được sử dụng bao gồm: Hói đầu, Tóc mái, Tóc đen, Tóc vàng, Mũm mĩm, Đeo kính, Râu dê, Tóc bạc, Trang điểm đậm, Nam giới, Miệng hói mở, Râu mép, Không râu, Da nhợt nhạt, Má hồng, Mỉm cười và Đánh son. Mỗi ảnh đầu vào được biểu diễn bởi một vector thuộc tính nhị phân có chiều tương ứng với số lượng thuộc tính được chọn, đóng vai trò là điều kiện miền (domain condition) cho mô hình StarGAN.

Việc giảm số lượng thuộc tính giúp quá trình huấn luyện trở nên ổn định hơn, đồng thời cho phép tập trung phân tích sâu chất lượng sinh ảnh và khả năng thay đổi từng thuộc tính cụ thể của mô hình.

4.1.3 Tiền xử lý dữ liệu

Trong đồ án này, em không xây dựng một quy trình tiền xử lý dữ liệu độc lập mà sử dụng trực tiếp pipeline tiền xử lý đã được cung cấp trong mã nguồn chính thức của StarGAN². Pipeline này đã được kiểm chứng qua nhiều nghiên cứu và thực nghiệm, đảm bảo tính uy tín.

Cụ thể, các bước tiền xử lý chính bao gồm:

- Chuẩn hóa kích thước ảnh về cùng một độ phân giải cố định (thường là 128×128 hoặc 256×256 pixel và sẽ được xét từ tham số truyền vào ở bước huấn

²<https://github.com/yunjey/stargan>

luyện) để phù hợp với kiến trúc mạng.

- Cắt và căn chỉnh khuôn mặt nhằm đảm bảo vùng khuôn mặt chiếm tỷ lệ hợp lý trong ảnh đầu vào.
- Chuẩn hóa giá trị điểm ảnh về khoảng $[-1, 1]$ để phù hợp với hàm kích hoạt *tanh* ở đầu ra của bộ sinh (Generator).
- Tạo các cặp dữ liệu gồm ảnh gốc và vector thuộc tính tương ứng để phục vụ huấn luyện có điều kiện.

4.1.4 Thiết lập mô hình StarGAN

Mô hình StarGAN được sử dụng trong đồ án dựa trên kiến trúc chuẩn do tác giả gốc đề xuất [6], bao gồm hai thành phần chính là bộ sinh (Generator) và bộ phân biệt (Discriminator). Bộ sinh nhận ảnh đầu vào cùng với vector thuộc tính mục tiêu và sinh ra ảnh mới với các thuộc tính được thay đổi tương ứng. Bộ phân biệt đồng thời thực hiện hai nhiệm vụ: phân biệt ảnh thật và ảnh sinh, và dự đoán thuộc tính của ảnh đầu vào.

Trong quá trình thiết lập mô hình, các thành phần kiến trúc chính như số lớp tích chập, số lượng kênh, hàm kích hoạt và cơ chế chuẩn hóa đều được giữ nguyên theo cấu hình gốc của StarGAN. Điều này giúp tập trung đánh giá ảnh hưởng của quá trình huấn luyện và lựa chọn checkpoint, thay vì thay đổi cấu trúc mạng.

Các tham số huấn luyện quan trọng như tốc độ học (learning rate), kích thước batch, số bước huấn luyện của bộ sinh và bộ phân biệt, cũng như các hệ số trọng số cho hàm mất mát đối kháng, mất mát phân loại và mất mát tái tạo chu trình đều được thiết lập theo khuyến nghị trong bài báo gốc. Trong một số trường hợp, các tham số này được điều chỉnh nhẹ để phù hợp với tài nguyên tính toán và điều kiện thực nghiệm trên nền tảng Kaggle.

Tóm lại, phần chuẩn bị dữ liệu và thiết lập mô hình trong đồ án được thực hiện dựa trên các thực hành chuẩn đã được kiểm chứng trong cộng đồng nghiên cứu. Việc sử dụng bộ dữ liệu CelebA cùng với pipeline và kiến trúc StarGAN gốc giúp đảm bảo tính khoa học, tính so sánh và độ tin cậy của các kết quả thực nghiệm. Những thiết lập này tạo nền tảng quan trọng cho quá trình huấn luyện mô hình được trình bày trong phần tiếp theo của chương.

4.2 Triển khai huấn luyện mô hình

Sau khi hoàn tất công tác chuẩn bị dữ liệu và thiết lập cấu hình mô hình ở Mục 4.1, chương này trình bày chi tiết quá trình triển khai huấn luyện mô hình StarGAN cho bài toán thay đổi thuộc tính khuôn mặt. Nội dung tập trung vào quy trình huấn luyện thực tế, chiến lược tối ưu được sử dụng, các siêu tham số quan

trọng, cũng như những quyết định kỹ thuật nhằm đảm bảo mô hình học ổn định và đạt hiệu quả tốt trong thực nghiệm.

4.2.1 Quy trình huấn luyện tổng thể

Quá trình huấn luyện mô hình được triển khai theo pipeline chuẩn của StarGAN, bao gồm hai mạng thành phần chính là mạng sinh (Generator) và mạng phân biệt (Discriminator). Hai mạng này được huấn luyện luân phiên theo cơ chế đối kháng, trong đó Generator học cách sinh ra ảnh giả có chất lượng cao và mang đúng thuộc tính mong muốn, còn Discriminator học cách phân biệt ảnh thật – ảnh giả, đồng thời dự đoán chính xác các thuộc tính của ảnh đầu vào.

Tại mỗi vòng lặp huấn luyện (iteration), mô hình thực hiện các bước chính sau:

- Lấy một minibatch ảnh thật từ tập huấn luyện cùng với vector thuộc tính gốc.
- Sinh vector thuộc tính mục tiêu bằng cách hoán đổi hoặc chọn ngẫu nhiên các nhãn thuộc tính.
- Cập nhật Discriminator thông qua các hàm mất mát đối kháng và mất mát phân loại thuộc tính.
- Cập nhật Generator nhằm đánh lừa Discriminator, đồng thời đảm bảo ảnh sinh giữ được nội dung gốc và thay đổi đúng thuộc tính.

Quy trình này được lặp lại liên tục cho đến khi đạt số vòng huấn luyện định trước. Các checkpoint trung gian được lưu định kỳ để phục vụ đánh giá và phân tích chất lượng mô hình theo thời gian.

4.2.2 Hàm mất mát và chiến lược tối ưu

Trong quá trình huấn luyện, StarGAN sử dụng tổ hợp nhiều hàm mất mát nhằm đảm bảo cả tính chân thực của ảnh sinh lẫn khả năng kiểm soát thuộc tính. Cụ thể, hàm mất mát tổng cho Generator bao gồm ba thành phần chính: mất mát đối kháng, mất mát phân loại thuộc tính và mất mát tái tạo chu trình (cycle consistency loss). Thành phần tái tạo đóng vai trò quan trọng trong việc giữ lại các đặc trưng không liên quan đến thuộc tính cần thay đổi, chẳng hạn như danh tính khuôn mặt hay cấu trúc tổng thể của ảnh.

Đối với Discriminator, hàm mất mát bao gồm mất mát phân biệt thật – giả và mất mát phân loại thuộc tính. Việc kết hợp phân loại thuộc tính trực tiếp trong Discriminator giúp mô hình học được mối quan hệ giữa nội dung ảnh và nhãn thuộc tính một cách hiệu quả hơn.

Mô hình được tối ưu bằng thuật toán Adam với các hệ số momentum được thiết lập theo khuyến nghị trong công trình StarGAN gốc. Tốc độ học (learning rate)

được giữ cố định trong giai đoạn đầu của huấn luyện, sau đó giảm dần theo lịch trình nhằm giúp mô hình hội tụ ổn định và tránh dao động mạnh ở các giai đoạn sau.

4.2.3 Thiết lập siêu tham số và cấu hình huấn luyện

Lệnh huấn luyện mô hình StarGAN được cấu hình như sau:

```
python main.py --mode train
--dataset CelebA
--image_size 156
--c_dim 17
--attr_path data/celeba/list_attr_celeba.txt
--celeba_image_dir /kaggle/input/celeba-dataset
--log_dir stargan_celeba/logs
--model_save_dir stargan_celeba/models
--result_dir stargan_celeba/results
--sample_dir stargan_celeba/samples
--batch_size 16
--g_lr 0.00005
--d_lr 0.00005
--num_iters 650000
--num_iters_decay 80000
--resume_iters 500000
--sample_step 2000
--model_save_step 50000
--log_step 2000
--selected_attrs Bald Bangs Black_Hair Blond_Hair
                    Chubby Eyeglasses Goatee Gray_Hair
                    Heavy_Makeup Male Mouth_Slightly_Open
                    Mustache No_Beard Pale_Skin
                    Rosy_Cheeks Smiling Wearing_Lipstick
```

Trong đó, tham số `image_size = 156` được lựa chọn nhằm cân bằng giữa chất lượng ảnh sinh và giới hạn tài nguyên GPU của môi trường Kaggle. Số chiều điều kiện `c_dim = 17` tương ứng với 17 thuộc tính khuôn mặt được sử dụng xuyên suốt trong toàn bộ quá trình huấn luyện và đánh giá mô hình.

Kích thước `batch_size = 16` được xác định sau nhiều lần thử nghiệm với các giá trị khác nhau. Giá trị này cho phép mô hình huấn luyện ổn định trong điều kiện bộ nhớ GPU hạn chế, đồng thời không làm suy giảm đáng kể tốc độ hội

tụ.

Tốc độ học của bộ sinh và bộ phân biệt được thiết lập đồng nhất với $g_{lr} = d_{lr} = 5 \times 10^{-5}$. Qua thực nghiệm, mức học này cho thấy sự cân bằng tốt giữa hai mạng, hạn chế hiện tượng mất ổn định trong huấn luyện GAN như dao động mạnh hoặc sụp đổ mô hình. Các giá trị tốc độ học lớn hơn thường dẫn đến hiện tượng nhiễu và làm giảm chất lượng ảnh sinh, trong khi các giá trị nhỏ hơn khiến quá trình hội tụ diễn ra chậm và kéo dài đáng kể thời gian huấn luyện.

Tổng số vòng lặp huấn luyện được đặt ở mức `num_iters` = 650000, cho phép mô hình học đầy đủ các biến đổi thuộc tính khuôn mặt phức tạp. Cơ chế suy giảm tuyến tính tốc độ học được kích hoạt tại `num_iters_decay` = 80000 vòng lặp cuối, giúp mô hình tinh chỉnh và ổn định hơn ở giai đoạn huấn luyện sau.

Tập các thuộc tính `selected_attrs` được lựa chọn dựa trên mức độ phổ biến trong tập dữ liệu CelebA cũng như khả năng thể hiện rõ ràng trên ảnh khuôn mặt.

Qua nhiều lần thử nghiệm với các cấu hình khác nhau, bộ tham số trên được xác định là cấu hình tối ưu nhất trong điều kiện tài nguyên hiện có, đảm bảo sự cân bằng giữa chất lượng ảnh sinh, độ ổn định huấn luyện và thời gian thực thi.

4.2.4 Môi trường triển khai

Quá trình huấn luyện được triển khai trên nền tảng điện toán đám mây Kaggle, sử dụng GPU để tăng tốc tính toán. Môi trường phần mềm bao gồm Python, PyTorch và các thư viện xử lý ảnh phổ biến. Việc sử dụng Kaggle giúp đảm bảo tính tái lập của thực nghiệm, đồng thời tạo điều kiện thuận lợi cho việc quản lý dữ liệu và checkpoint huấn luyện.

Trong quá trình huấn luyện kéo dài, mô hình được giám sát thông qua các chỉ số mất mát và kết quả sinh ảnh trung gian. Việc theo dõi này giúp phát hiện sớm các dấu hiệu bất ổn như mất mát dao động mạnh hoặc hiện tượng mode collapse, từ đó có thể điều chỉnh kịp thời các tham số huấn luyện.

4.2.5 Lưu checkpoint và chiến lược đánh giá trung gian

Một điểm quan trọng trong quá trình triển khai huấn luyện là chiến lược lưu và đánh giá checkpoint. Thay vì chỉ sử dụng mô hình ở vòng huấn luyện cuối cùng, các checkpoint trung gian được trích xuất và đánh giá độc lập bằng các chỉ số SSIM và FID. Cách tiếp cận này cho phép phân tích sâu hơn quá trình học của mô hình, đồng thời xác định những thời điểm huấn luyện mà mô hình đạt chất lượng sinh ảnh tốt nhất.

Việc đánh giá theo checkpoint cũng cho thấy mô hình không nhất thiết đạt kết quả tối ưu ở iteration cuối, mà có thể đạt cực trị chất lượng tại một giai đoạn trung

gian. Điều này đặc biệt quan trọng đối với các mô hình GAN, vốn nhạy cảm với quá trình huấn luyện kéo dài và dễ bị suy giảm chất lượng nếu huấn luyện quá mức.

Tổng thể, quá trình huấn luyện mô hình StarGAN được triển khai theo đúng pipeline chuẩn, kết hợp với chiến lược giám sát và đánh giá chặt chẽ. Mô hình thể hiện khả năng học ổn định, sinh ảnh có chất lượng tốt và thay đổi thuộc tính tương đối chính xác trên nhiều nhóm đặc trưng khuôn mặt khác nhau. Những kết quả thu được từ quá trình huấn luyện này là cơ sở cho các thực nghiệm đánh giá chi tiết được trình bày trong Chương 5, cũng như cho việc xây dựng ứng dụng thử nghiệm ở Chương 6.

4.3 Các vấn đề gặp phải và cách khắc phục

Trong quá trình triển khai huấn luyện mô hình StarGAN cho bài toán thay đổi thuộc tính khuôn mặt, bên cạnh những kết quả đạt được, đồ án cũng gặp phải nhiều khó khăn và thách thức đến từ cả yếu tố kỹ thuật lẫn điều kiện thực nghiệm. Các vấn đề này xuất phát chủ yếu từ đặc thù của mô hình học sâu sinh ảnh, yêu cầu tài nguyên tính toán lớn, cũng như những giới hạn cố hữu của môi trường huấn luyện được sử dụng. Việc nhận diện rõ các vấn đề và đưa ra giải pháp phù hợp đóng vai trò quan trọng trong việc đảm bảo tính khả thi và độ tin cậy của kết quả huấn luyện.

4.3.1 Hạn chế về tài nguyên tính toán

Một trong những khó khăn lớn nhất trong quá trình huấn luyện mô hình là hạn chế về tài nguyên phần cứng, đặc biệt là năng lực xử lý của GPU. Môi trường Kaggle cung cấp GPU miễn phí nhưng chỉ ở mức NVIDIA Tesla T4 hoặc P100, với dung lượng bộ nhớ giới hạn. Trong khi đó, StarGAN là mô hình sinh ảnh đa miền có kiến trúc phức tạp, yêu cầu xử lý đồng thời nhiều thuộc tính khuôn mặt và thực hiện các phép toán tích chập với ảnh độ phân giải tương đối cao.

Hạn chế này dẫn đến việc không thể tăng batch size lớn như mong muốn, làm ảnh hưởng đến độ ổn định của quá trình huấn luyện và tốc độ hội tụ của mô hình. Ngoài ra, việc huấn luyện trong thời gian dài với số lượng iteration lớn cũng dễ gặp tình trạng tràn bộ nhớ hoặc gián đoạn tiến trình.

Để khắc phục vấn đề này, đồ án đã tiến hành thử nghiệm nhiều cấu hình siêu tham số khác nhau, bao gồm batch size, learning rate, hệ số cân bằng giữa các hàm mất mát và tần suất cập nhật discriminator/generator. Thay vì lựa chọn một cấu hình cố định ngay từ đầu, quá trình huấn luyện được chia thành các giai đoạn thử nghiệm ngắn, trong đó từng tập siêu tham số được đánh giá thông qua các chỉ số định lượng như FID và SSIM. Từ kết quả thu được, bộ siêu tham số cho hiệu năng tốt nhất trong điều kiện tài nguyên cho phép được lựa chọn để tiếp tục huấn luyện dài hạn. Cách tiếp cận này giúp tận dụng hiệu quả tài nguyên GPU hạn chế mà vẫn

đảm bảo chất lượng mô hình ở mức chấp nhận được.

4.3.2 Hạn chế về thời gian huấn luyện

Một thách thức quan trọng khác đến từ giới hạn thời gian sử dụng GPU liên tục trên Kaggle: Kaggle chỉ cho phép tối đa khoảng 12 giờ cho mỗi phiên làm việc. Trong khi đó, để mô hình StarGAN hội tụ ổn định, số lượng iteration cần thiết thường lên tới hàng trăm nghìn bước huấn luyện.

Để giải quyết vấn đề này, quá trình huấn luyện được chia nhỏ thành nhiều giai đoạn, mỗi giai đoạn tương ứng với một khoảng iteration xác định và kết thúc bằng việc lưu lại checkpoint của mô hình. Các checkpoint này bao gồm trọng số của generator và discriminator, cho phép tiếp tục huấn luyện từ trạng thái trước đó thay vì phải khởi động lại từ đầu. Cách làm này không chỉ giúp vượt qua giới hạn thời gian của mỗi phiên Kaggle mà còn tạo điều kiện thuận lợi cho việc đánh giá chất lượng mô hình tại các mốc huấn luyện khác nhau.

Bên cạnh đó, đồ án cũng tận dụng việc tạo và sử dụng nhiều tài khoản Kaggle hợp lệ để phân bổ quá trình huấn luyện theo thời gian, từ đó tối ưu hóa việc khai thác tài nguyên GPU miễn phí. Mặc dù cách tiếp cận này không làm giảm tổng thời gian huấn luyện thực tế, nhưng giúp đảm bảo tiến trình huấn luyện không bị gián đoạn và tận dụng tối đa hạ tầng sẵn có.

4.3.3 Hạn chế trong đánh giá định lượng

Mặc dù các chỉ số như FID và SSIM cung cấp cơ sở định lượng quan trọng để đánh giá chất lượng ảnh sinh, việc tính toán các chỉ số này cũng tiêu tốn đáng kể tài nguyên và thời gian, đặc biệt khi cần đánh giá trên nhiều thuộc tính và nhiều checkpoint khác nhau. Do đó, đồ án không tính toán các chỉ số này tại mọi iteration mà chỉ thực hiện tại các mốc checkpoint quan trọng.

Cách tiếp cận này giúp giảm chi phí tính toán trong khi vẫn đảm bảo có đủ dữ liệu để phân tích xu hướng hội tụ của mô hình và so sánh chất lượng giữa các giai đoạn huấn luyện.

Từ những vấn đề và giải pháp nêu trên có thể thấy rằng, mặc dù bị giới hạn bởi môi trường huấn luyện và tài nguyên tính toán, đồ án vẫn đạt được các kết quả đáng tin cậy thông qua việc lựa chọn chiến lược huấn luyện phù hợp, tận dụng checkpoint, tối ưu siêu tham số và đánh giá có chọn lọc. Những kinh nghiệm rút ra trong quá trình khắc phục các khó khăn này không chỉ giúp hoàn thiện đồ án hiện tại mà còn là cơ sở thực tiễn quan trọng cho các nghiên cứu và ứng dụng tiếp theo trong lĩnh vực sinh ảnh bằng mô hình GAN.

CHƯƠNG 5. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Chương này trình bày quá trình thiết lập môi trường thực nghiệm, phương pháp đánh giá và các kết quả thu được khi áp dụng mô hình StarGAN đã huấn luyện cho bài toán thay đổi thuộc tính khuôn mặt. Toàn bộ các thí nghiệm được thiết kế nhằm đánh giá một cách định lượng và có hệ thống chất lượng ảnh sinh ra theo từng checkpoint huấn luyện cũng như theo từng thuộc tính khuôn mặt. Các nội dung trong chương bao gồm mô tả môi trường thực nghiệm, phương pháp đánh giá dựa trên các chỉ số SSIM và FID, phân tích kết quả tổng thể và phân tích chi tiết theo từng thuộc tính, từ đó rút ra nhận xét về quá trình huấn luyện và lựa chọn mô hình phù hợp cho ứng dụng thực tế.

5.1 Thiết lập môi trường thực nghiệm

Các thí nghiệm trong đồ án được triển khai trên nền tảng Kaggle Notebook, một môi trường điện toán đám mây phổ biến cho nghiên cứu và thử nghiệm các mô hình học sâu. Việc lựa chọn Kaggle xuất phát từ khả năng cung cấp miễn phí tài nguyên GPU, hỗ trợ sẵn các thư viện phổ biến cho học máy và thị giác máy tính, đồng thời thuận tiện cho việc tổ chức các thử nghiệm mà không cần tới thiết bị có GPU thật.

Về phần cứng, em sử dụng môi trường Kaggle với GPU P100, đi kèm với bộ nhớ đồ họa giới hạn và thời gian chạy tối đa cho mỗi phiên làm việc là 12 giờ liên tục. Các hạn chế này ảnh hưởng trực tiếp đến chiến lược huấn luyện và đánh giá mô hình, đặc biệt đối với các mô hình sinh ảnh có thời gian huấn luyện dài như StarGAN. Do đó, quá trình thực nghiệm được thiết kế theo hướng chia nhỏ thành nhiều phiên.

Về phần mềm, các thí nghiệm được thực hiện với code bằng Python với các thư viện chính bao gồm PyTorch cho việc xây dựng và sử dụng mô hình học sâu, torchvision cho các mô hình và phép biến đổi ảnh chuẩn hóa, OpenCV và Pillow cho xử lý ảnh, cùng với NumPy và SciPy cho các phép toán số học và thống kê. Mã nguồn StarGAN được sử dụng trong đồ án được kệ thừa và điều chỉnh từ phiên bản công khai trên GitHub.

Dữ liệu thực nghiệm được lấy từ bộ dữ liệu CelebA, em lấy ra 1000 ảnh khuôn mặt ngẫu nhiên và cố định để làm tập kiểm tra. Việc sử dụng cùng một tập ảnh đầu vào cho tất cả các checkpoint nhằm đảm bảo tính công bằng và nhất quán khi so sánh kết quả sinh ảnh giữa các giai đoạn huấn luyện khác nhau. Tập ảnh này được giữ nguyên trong suốt quá trình thực nghiệm và được sử dụng để sinh ảnh



Hình 5.1: Ví dụ về ảnh sinh ra

với 17 thuộc tính khuôn mặt khác nhau theo cấu hình của mô hình StarGAN train được. Quy trình thực nghiệm được tổ chức theo dạng pipeline khép kín. Với mỗi checkpoint huấn luyện, mô hình StarGAN được sử dụng ở chế độ kiểm tra để sinh ảnh từ tập 1000 ảnh đầu vào. Kết quả sinh ảnh có dạng ảnh ghép ngang, trong đó ảnh đầu tiên là ảnh gốc và 17 ảnh tiếp theo tương ứng với từng thuộc tính khuôn mặt được thay đổi (Hình 5.1). Các ảnh ghép này sau đó được tự động tách thành các ảnh thành phần và lưu trữ vào các thư mục riêng biệt theo từng thuộc tính, tạo tiền đề cho bước đánh giá định lượng.

Việc đánh giá chất lượng ảnh sinh được thực hiện hoàn toàn tự động thông qua các chỉ số SSIM và FID. Trong đó, SSIM được sử dụng để đo mức độ tương đồng về cấu trúc giữa ảnh sinh và ảnh gốc, còn FID được sử dụng để đánh giá mức độ khác biệt phân phối đặc trưng giữa tập ảnh sinh và tập ảnh thật. Toàn bộ quá trình đánh giá được triển khai song song nhằm giảm thời gian tính toán, phù hợp với giới hạn tài nguyên của môi trường Kaggle.

Nhìn chung, môi trường thực nghiệm được thiết lập theo hướng cân bằng giữa điều kiện tài nguyên thực tế và yêu cầu đánh giá khoa học. Các lựa chọn về phần cứng, phần mềm, dữ liệu và quy trình thực nghiệm đều nhằm đảm bảo rằng kết quả thu được phản ánh đúng chất lượng của mô hình StarGAN tại các giai đoạn huấn luyện khác nhau, đồng thời có thể tái hiện và mở rộng trong các nghiên cứu hoặc ứng dụng tiếp theo.

5.2 Phương pháp đánh giá mô hình

Đánh giá chất lượng mô hình sinh ảnh nói chung và mô hình thay đổi thuộc tính khuôn mặt nói riêng là một bài toán không đơn giản, do kết quả đầu ra không chỉ cần đảm bảo tính chân thực về mặt thị giác mà còn phải phản ánh đúng mức độ thay đổi của các thuộc tính mong muốn. Trong bối cảnh bài toán thay đổi thuộc tính khuôn mặt đa miền với StarGAN, việc đánh giá mô hình cần được thực hiện một cách toàn diện, kết hợp giữa các chỉ số định lượng khách quan và phân tích định tính dựa trên quan sát trực quan.

Xuất phát từ mục tiêu và phạm vi của đồ án đã trình bày trong Chương 1, phương

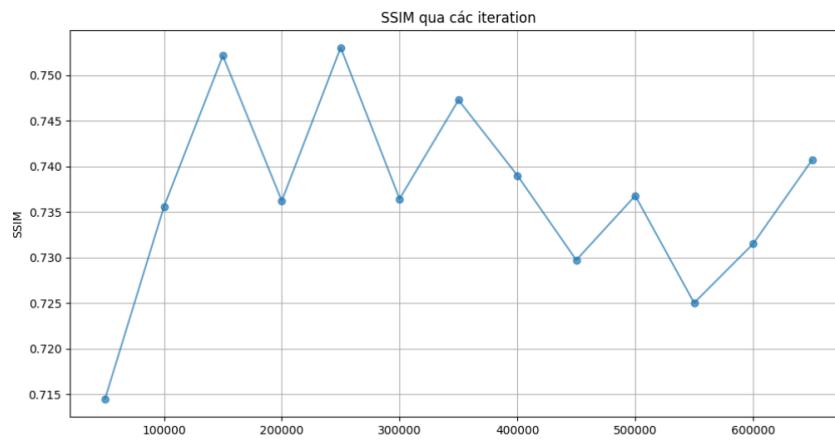
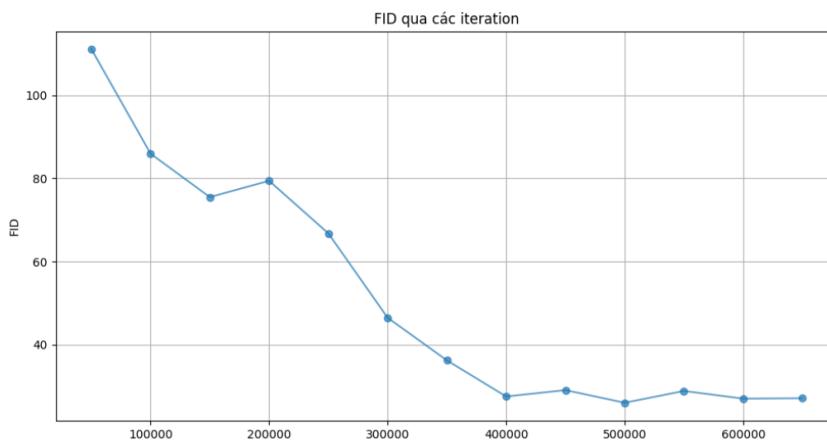
pháp đánh giá trong đồ án này tập trung vào ba khía cạnh chính: (i) mức độ bảo toàn cấu trúc khuôn mặt gốc sau khi thay đổi thuộc tính, (ii) chất lượng và độ chân thực của ảnh sinh so với phân bố ảnh thật, và (iii) khả năng kiểm soát và ổn định của mô hình đối với từng thuộc tính khuôn mặt trong quá trình huấn luyện.

Trước hết, để đánh giá mức độ bảo toàn cấu trúc và nội dung chính của khuôn mặt gốc, em sử dụng chỉ số SSIM (Structural Similarity Index Measure). Chỉ số này đo lường mức độ tương đồng cấu trúc giữa ảnh đầu vào và ảnh sinh sau khi thay đổi thuộc tính, dựa trên ba thành phần là độ sáng, độ tương phản và cấu trúc không gian. Trong ngữ cảnh của bài toán, SSIM được tính giữa ảnh gốc và ảnh sinh ứng với từng thuộc tính, qua đó phản ánh mức độ mà mô hình giữ được danh tính và hình dạng khuôn mặt ban đầu trong khi vẫn thực hiện thay đổi thuộc tính. Giá trị SSIM cao cho thấy mô hình hạn chế được hiện tượng méo hình, mất chi tiết hoặc thay đổi không mong muốn ở các vùng không liên quan.

Tiếp theo, để đánh giá chất lượng tổng thể và độ chân thực của ảnh sinh, em sử dụng chỉ số FID (Fréchet Inception Distance). FID đo khoảng cách giữa phân bố đặc trưng của ảnh thật và ảnh sinh trong không gian đặc trưng được trích xuất bởi mạng Inception. Khác với các chỉ số dựa trên từng cặp ảnh, FID đánh giá ở mức phân bố, do đó phản ánh tốt hơn khả năng của mô hình trong việc sinh ra các ảnh có tính đa dạng và gần với dữ liệu thực. Trong đồ án này, FID được tính cho toàn bộ tập ảnh sinh tại các mốc huấn luyện khác nhau, đồng thời được phân tích riêng cho từng thuộc tính khuôn mặt, nhằm đánh giá mức độ ổn định và hiệu quả của mô hình đối với từng miền thuộc tính.

Bên cạnh việc đánh giá tại một thời điểm huấn luyện cố định, đồ án còn tiến hành theo dõi sự thay đổi của các chỉ số SSIM và FID theo số vòng lặp huấn luyện (iteration). Cách tiếp cận này cho phép quan sát trực tiếp quá trình hội tụ của mô hình, phát hiện sớm các hiện tượng như quá khớp (overfitting), suy giảm chất lượng ảnh sinh hoặc mất cân bằng giữa các thuộc tính. Các biểu đồ SSIM và FID theo iteration đóng vai trò quan trọng trong việc lựa chọn checkpoint huấn luyện phù hợp để sử dụng cho các thí nghiệm tiếp theo và cho ứng dụng thử nghiệm.

Ngoài các chỉ số định lượng, đánh giá định tính thông qua quan sát trực quan cũng được sử dụng như một phương pháp bổ trợ. Các ảnh sinh được so sánh trực tiếp với ảnh gốc và với các ảnh sinh tại những checkpoint khác nhau, tập trung vào các yếu tố như mức độ rõ ràng của thuộc tính thay đổi, tính tự nhiên của khuôn mặt, cũng như các lỗi thường gặp như nhiễu hoặc thay đổi sai thuộc tính. Phân tích định tính này đặc biệt quan trọng trong các trường hợp mà chỉ số định lượng chưa phản ánh đầy đủ cảm nhận thị giác của con người.

**Hình 5.2:** Biểu đồ chỉ số SSIM qua các iteration**Hình 5.3:** Biểu đồ chỉ số FID qua các iteration

Tóm lại, phương pháp đánh giá mô hình trong đồ án được xây dựng theo hướng đa chiều, kết hợp giữa các chỉ số định lượng phổ biến trong lĩnh vực sinh ảnh và phân tích định tính dựa trên quan sát thực tế. Cách tiếp cận này không chỉ cho phép đánh giá khách quan chất lượng mô hình StarGAN được huấn luyện, mà còn cung cấp cơ sở vững chắc cho việc phân tích sâu kết quả thực nghiệm và so sánh các checkpoint huấn luyện trong các mục tiếp theo của chương này.

5.3 Kết quả đánh giá tổng thể

Dựa trên phương pháp đánh giá đã trình bày ở Mục 5.2, mô hình StarGAN được huấn luyện và đánh giá định kỳ tại nhiều mốc checkpoint khác nhau cách nhau 50 000 iters, từ 50 000 đến 650 000 vòng lặp. Hai chỉ số chính được sử dụng để đánh giá chất lượng ảnh sinh là SSIM và FID, phản ánh lần lượt mức độ bảo toàn cấu trúc so với ảnh gốc và mức độ tương đồng phân bố giữa ảnh sinh và ảnh thật.

Xét trên phương diện tổng thể, kết quả thực nghiệm cho thấy mô hình có xu hướng hội tụ ổn định khi số vòng lặp huấn luyện tăng lên. Ở các checkpoint ban đầu (50 000–100 000 iterations), giá trị SSIM trung bình còn ở mức tương đối thấp,

trong khi FID trung bình khá cao. Điều này cho thấy tại giai đoạn đầu, mô hình chưa học được đầy đủ các đặc trưng hình ảnh khuôn mặt và việc thay đổi thuộc tính còn gây ra nhiều biến dạng không mong muốn. Khi quá trình huấn luyện tiếp tục, đặc biệt từ khoảng 150 000 đến 300 000 iterations, SSIM trung bình tăng dần, đồng thời FID giảm rõ rệt, phản ánh sự cải thiện đáng kể về cả mức độ bảo toàn cấu trúc khuôn mặt lẫn tính chân thực của ảnh sinh (Hình 5.2, Hình 5.3).

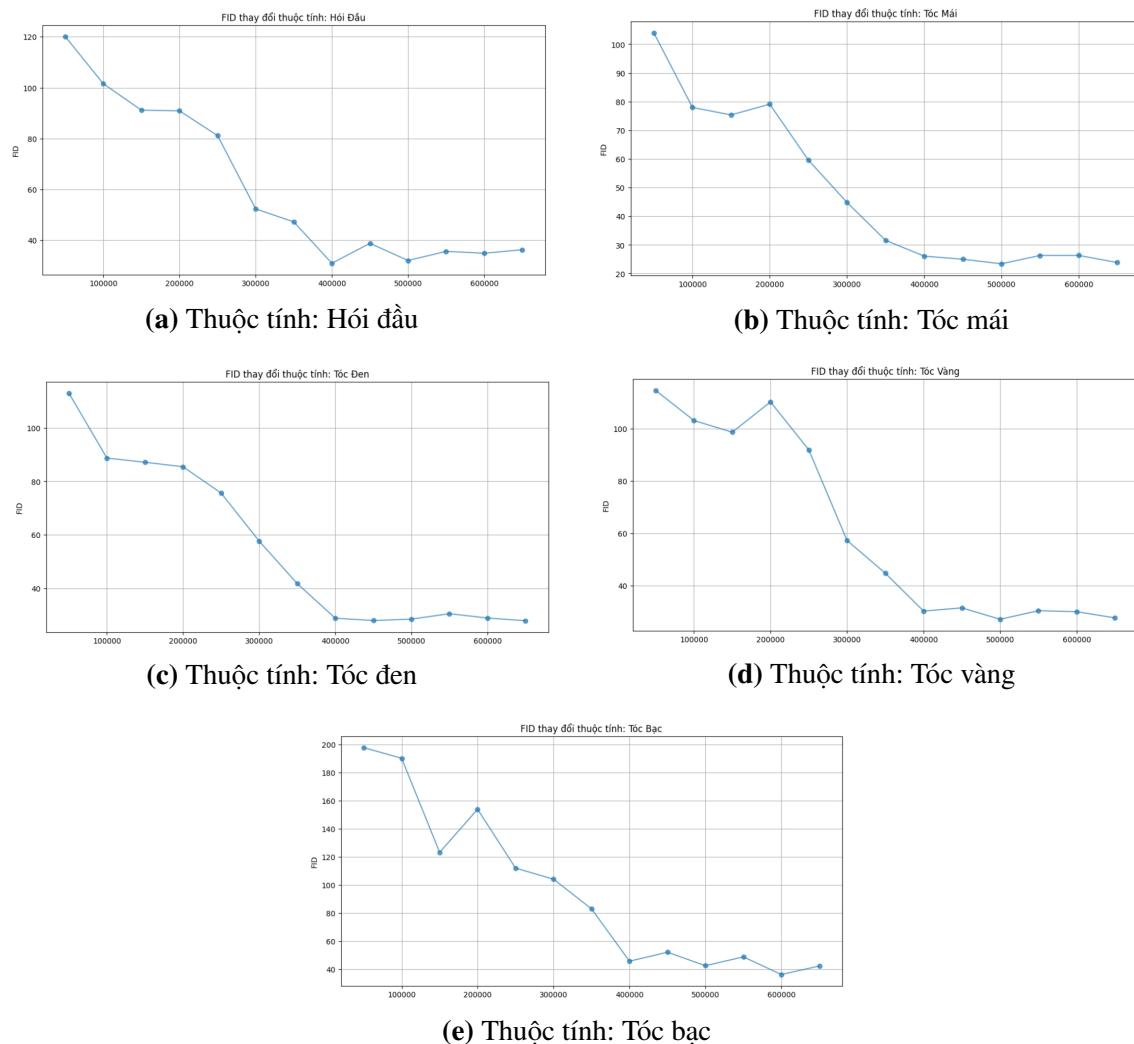
Tại các checkpoint từ 300 000 đến 400 000 iterations, mô hình đạt được sự cân bằng tốt giữa hai tiêu chí đánh giá. SSIM trung bình duy trì ở mức cao và ổn định, cho thấy các chi tiết hình học và cấu trúc khuôn mặt của ảnh đầu vào được giữ lại tương đối tốt sau khi thay đổi thuộc tính. Đồng thời, FID trung bình giảm xuống mức thấp, phản ánh việc phân bố đặc trưng của ảnh sinh ngày càng tiệm cận với ảnh thật trong tập dữ liệu CelebA. Đây có thể xem là giai đoạn mô hình đạt hiệu năng tổng thể tốt nhất, phù hợp để lựa chọn làm checkpoint đại diện cho quá trình huấn luyện.

Ở các checkpoint muộn hơn (từ 450 000 iterations trở đi), kết quả đánh giá cho thấy SSIM trung bình không còn tăng đáng kể và có xu hướng dao động nhẹ, trong khi FID không tiếp tục giảm mà duy trì quanh một ngưỡng nhất định. Hiện tượng này cho thấy mô hình đã tiệm cận trạng thái hội tụ, và việc huấn luyện thêm không mang lại cải thiện rõ rệt về chất lượng ảnh sinh. Trong một số trường hợp, việc huấn luyện quá lâu còn có thể dẫn đến hiện tượng quá khớp nhẹ, khiến chất lượng tổng thể không ổn định bằng giai đoạn trước đó.

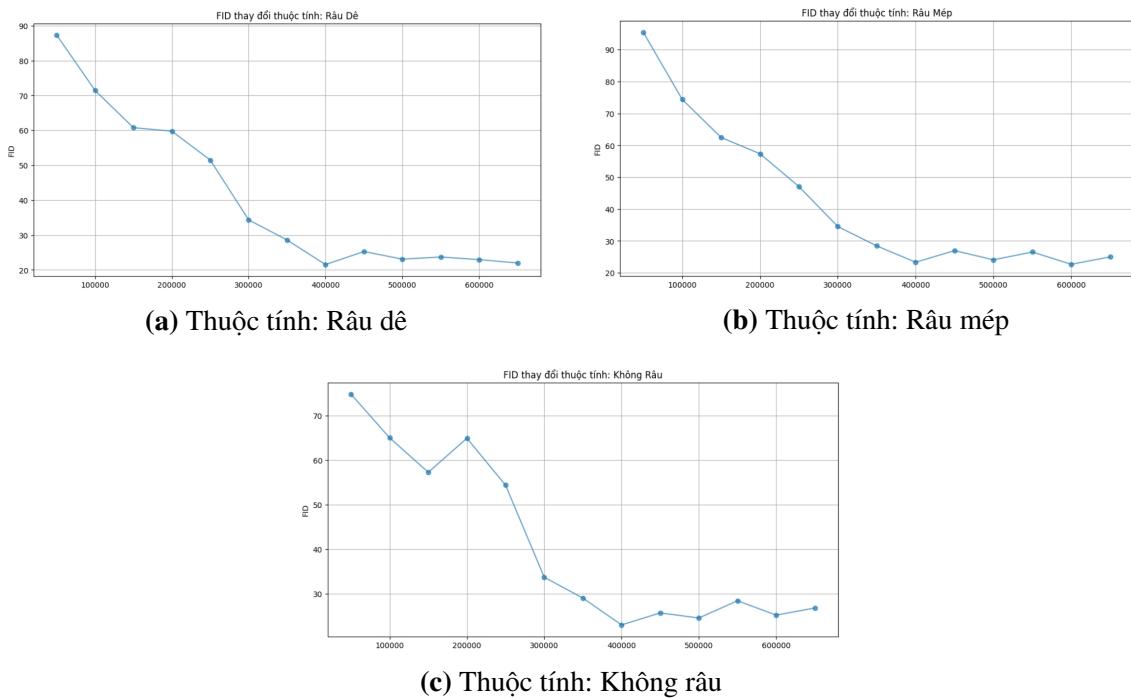
Tổng hợp các kết quả trên, có thể nhận thấy mô hình StarGAN trong đồ án đã học được biểu diễn hiệu quả cho bài toán thay đổi thuộc tính khuôn mặt trên tập dữ liệu CelebA. Các chỉ số SSIM và FID trung bình đều cho thấy xu hướng cải thiện rõ ràng theo thời gian huấn luyện và đạt giá trị ổn định ở các checkpoint trung gian. Những kết quả này là cơ sở quan trọng để tiếp tục phân tích sâu hơn hiệu năng của mô hình theo từng thuộc tính khuôn mặt cụ thể, nội dung sẽ được trình bày chi tiết trong Mục 5.4.

5.4 Phân tích kết quả theo từng thuộc tính thay đổi trên ảnh khuôn mặt

Bên cạnh việc đánh giá tổng thể mô hình thông qua các chỉ số trung bình, việc phân tích chi tiết kết quả theo từng thuộc tính khuôn mặt là cần thiết nhằm làm rõ khả năng học và mức độ ổn định của mô hình StarGAN đối với từng loại biến đổi cụ thể. Do các thuộc tính khuôn mặt có bản chất, mức độ biểu hiện và độ khó khác nhau, kết quả SSIM và FID thu được giữa các thuộc tính cũng thể hiện sự khác biệt đáng kể. Phần này tập trung phân tích xu hướng biến đổi của hai chỉ số SSIM và FID theo số vòng lặp huấn luyện đối với từng thuộc tính trong tập CelebA.



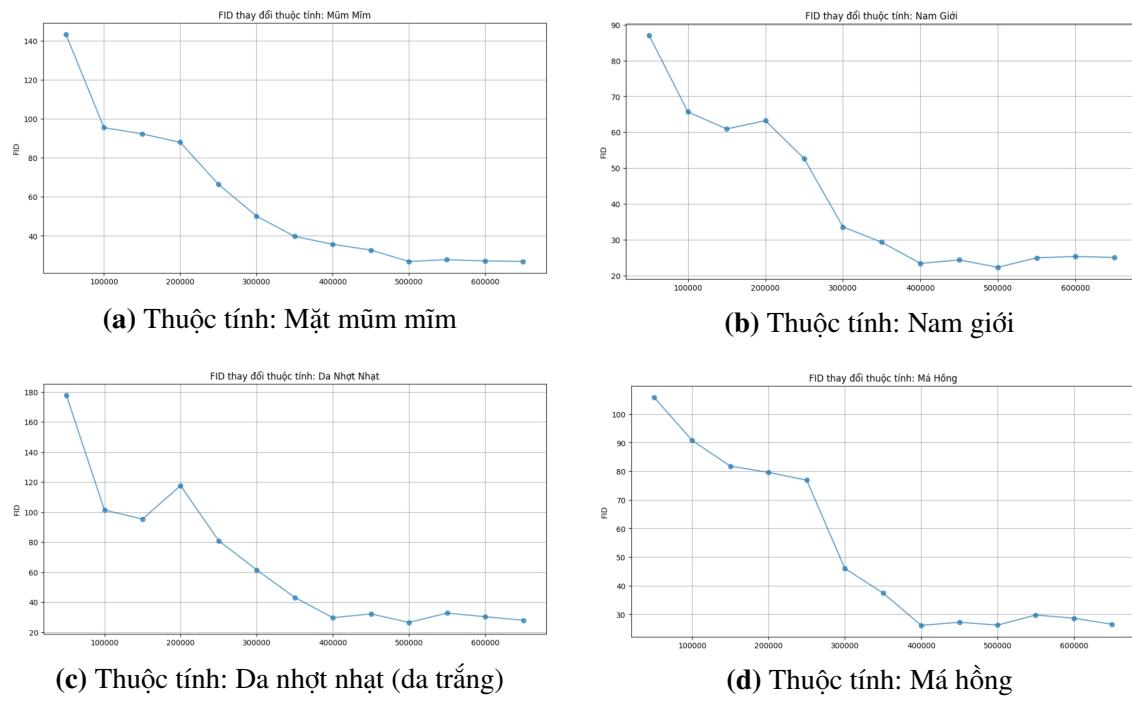
Hình 5.4: Biểu đồ FID của nhóm ảnh liên quan đến tóc

**Hình 5.5:** Biểu đồ FID của nhóm ảnh liên quan đến râu

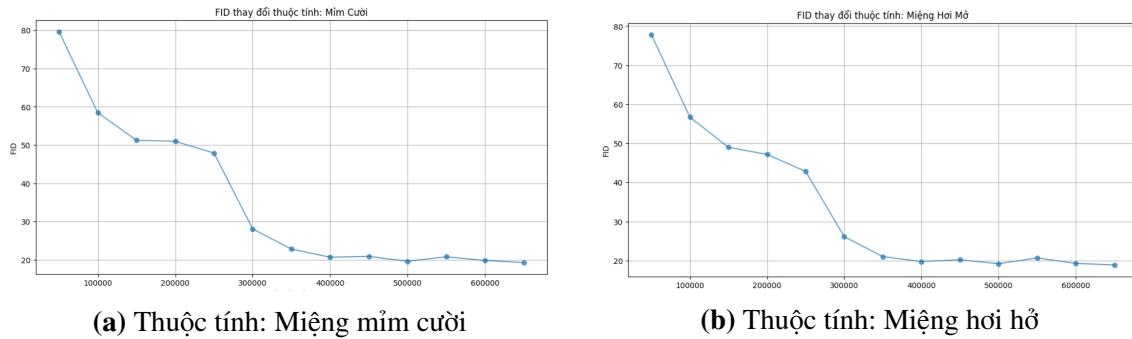
Trước hết, nhóm các thuộc tính liên quan đến kiểu tóc và màu tóc, bao gồm *Hói Đầu*, *Tóc Mái*, *Tóc Đen*, *Tóc Vàng* và *Tóc Bạc*, cho thấy khả năng học khá tốt của mô hình (Hình 5.4). Các thuộc tính này thường có sự thay đổi rõ rệt về mặt thị giác, tập trung chủ yếu vào vùng tóc, do đó mô hình dễ dàng nắm bắt được các đặc trưng không gian liên quan. Chỉ số SSIM của nhóm này tăng dần theo số vòng lặp huấn luyện và đạt giá trị ổn định từ các checkpoint trung và muộn, trong khi FID giảm mạnh trong giai đoạn đầu và tiếp tục cải thiện ở giai đoạn sau. Điều này cho thấy ảnh sinh ra ngày càng giữ được cấu trúc tổng thể của ảnh gốc, đồng thời phân bố ảnh sinh tiệm cận hơn với phân bố ảnh thật.

Đối với các thuộc tính liên quan đến râu như *Râu Dê*, *Râu Mép* và *Không Râu*, mô hình cũng đạt kết quả tương đối tốt, tuy nhiên tốc độ hội tụ có phần chậm hơn so với nhóm thuộc tính về tóc (Hình 5.5). Nguyên nhân có thể đến từ việc các đặc trưng này thường chỉ chiếm diện tích nhỏ trên khuôn mặt và dễ bị ảnh hưởng bởi tư thế đầu, ánh sáng cũng như chất lượng ảnh. Dù vậy, các giá trị FID của nhóm này vẫn giảm đều theo quá trình huấn luyện, cho thấy mô hình dần học được cách biến đổi các chi tiết tinh vi mà không làm phá vỡ cấu trúc tổng thể của khuôn mặt.

Nhóm các thuộc tính mang tính hình thái khuôn mặt hoặc đặc điểm sinh học, bao gồm *Mũm Mũm*, *Nam Giới*, *Da Nhợt Nhạt* và *Má Hồng*, thể hiện mức độ khó cao hơn (Hình 5.6). Các thuộc tính này không chỉ ảnh hưởng cục bộ mà còn liên quan đến nhiều vùng trên khuôn mặt, thậm chí là toàn bộ bộ cục ảnh. Do đó, SSIM của nhóm này có xu hướng tăng chậm hơn và dao động nhẹ giữa các checkpoint,



Hình 5.6: Biểu đồ FID của nhóm ảnh liên quan đến hình thái khuôn mặt, màu da mặt

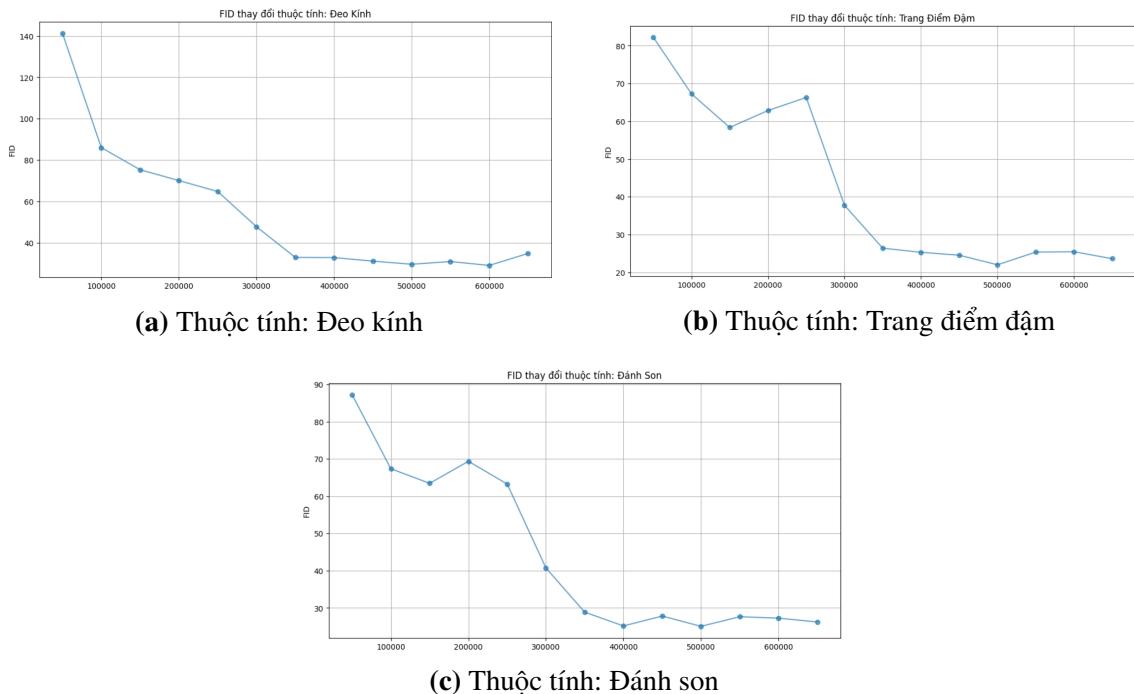


Hình 5.7: Biểu đồ FID của nhóm ảnh liên quan đến biểu cảm khuôn mặt, hình miệng

trong khi FID vẫn duy trì xu hướng giảm nhưng với biên độ không lớn. Điều này phản ánh thách thức của mô hình trong việc thay đổi các đặc điểm mang tính toàn cục mà vẫn bảo toàn danh tính và các chi tiết quan trọng của khuôn mặt.

Đối với các thuộc tính biểu cảm và trạng thái khuôn mặt như *Mỉm Cười* và *Miệng Hơi Mở*, kết quả cho thấy mô hình có khả năng học tương đối tốt, đặc biệt ở các checkpoint muộn (Hình 5.7). Đây là các thuộc tính có sự thay đổi hình học rõ ràng ở vùng miệng, giúp mô hình dễ dàng phát hiện và tái tạo. Chỉ số SSIM của các thuộc tính này đạt giá trị cao và ổn định, trong khi FID giảm xuống mức thấp, cho thấy ảnh sinh ra vừa giữ được cấu trúc khuôn mặt, vừa thể hiện đúng biểu cảm mong muốn.

Cuối cùng, các thuộc tính liên quan đến trang điểm và phụ kiện, bao gồm *Trang Điểm Đậm*, *Dánh Son* và *Đeo Kính*, cho thấy sự cải thiện rõ rệt theo quá trình huấn



Hình 5.8: Biểu đồ FID của nhóm ảnh liên quan đến trang điểm, phụ kiện

luyện (Hình 5.8). Đây là những thuộc tính thường gắn với các vùng cụ thể như mắt, môi hoặc toàn bộ khuôn mặt, đồng thời có sự tương phản màu sắc cao. Nhờ đó, mô hình StarGAN có thể học được các đặc trưng tương ứng khá hiệu quả. Các chỉ số FID của nhóm này giảm mạnh ở giai đoạn giữa và duy trì ổn định ở giai đoạn sau, trong khi SSIM đạt mức tương đối cao so với nhiều nhóm thuộc tính khác.

Tổng hợp lại, kết quả phân tích theo từng thuộc tính cho thấy mô hình StarGAN hoạt động hiệu quả hơn đối với các thuộc tính có sự thay đổi rõ ràng về mặt thị giác và mang tính cục bộ, trong khi gặp nhiều thách thức hơn với các thuộc tính mang tính toàn cục hoặc có biểu hiện tinh vi. Những quan sát này không chỉ giúp đánh giá sâu hơn chất lượng mô hình mà còn là cơ sở quan trọng để lựa chọn checkpoint phù hợp và đề xuất các hướng cải tiến trong các nghiên cứu và ứng dụng tiếp theo.

5.5 Nhận xét và so sánh các checkpoint huấn luyện

Dựa trên các kết quả đánh giá tổng thể ở Mục 5.3 và phân tích chi tiết theo từng thuộc tính khuôn mặt ở Mục 5.4, có thể nhận thấy quá trình huấn luyện mô hình StarGAN diễn ra theo xu hướng hợp lý và phản ánh rõ sự tiến bộ của mô hình theo số vòng lặp huấn luyện (iteration).

Trong giai đoạn đầu của quá trình huấn luyện (từ 50 000 đến khoảng 150 000 iteration), các chỉ số đánh giá cho thấy mô hình còn chưa ổn định. Giá trị FID trung bình ở mức cao, phản ánh sự khác biệt lớn giữa phân phối ảnh sinh và ảnh thật, trong khi SSIM tuy có xu hướng tăng nhưng vẫn còn dao động đáng kể giữa

các thuộc tính. Điều này là phù hợp với đặc điểm của mô hình GAN, khi bộ sinh và bộ phân biệt vẫn đang trong quá trình học các đặc trưng cơ bản của dữ liệu khuôn mặt và mối quan hệ giữa các thuộc tính.

Từ khoảng 200 000 đến 350 000 iteration, mô hình bắt đầu thể hiện sự cải thiện rõ rệt. Giá trị FID trung bình giảm mạnh, cho thấy chất lượng ảnh sinh được nâng cao đáng kể và phân phối ảnh sinh tiến gần hơn tới phân phối ảnh thật. Đồng thời, SSIM duy trì ở mức tương đối ổn định và cao hơn so với giai đoạn đầu, phản ánh khả năng bảo toàn cấu trúc khuôn mặt gốc khi thay đổi thuộc tính. Ở giai đoạn này, hầu hết các thuộc tính như *Smiling*, *Male*, *Wearing Lipstick* hay *Black Hair* đều đạt được sự cân bằng tốt giữa mức độ biến đổi thuộc tính và tính nhất quán về hình dạng khuôn mặt.

Trong giai đoạn sau (từ 400 000 đến 650 000 iteration), mặc dù SSIM không còn tăng mạnh mà có xu hướng dao động nhẹ quanh một giá trị trung bình, chỉ số FID tiếp tục giảm và đạt mức thấp nhất tại checkpoint 500 000. Cụ thể, checkpoint này cho giá trị FID trung bình nhỏ nhất trong toàn bộ quá trình huấn luyện, đồng thời SSIM vẫn được duy trì ở mức ổn định, không xuất hiện dấu hiệu suy giảm nghiêm trọng. Điều này cho thấy mô hình đã đạt được trạng thái cân bằng tương đối tốt giữa hai mục tiêu: tạo ảnh có chất lượng cao và bảo toàn cấu trúc khuôn mặt ban đầu.

So sánh chi tiết giữa các checkpoint lân cận cho thấy, mặc dù các checkpoint sau 500 000 iteration (ví dụ 550 000, 600 000 và 650 000) vẫn duy trì chất lượng ảnh ở mức tốt, song không mang lại sự cải thiện đáng kể về FID, thậm chí trong một số thuộc tính còn xuất hiện hiện tượng FID tăng nhẹ. Điều này có thể được lý giải bởi hiện tượng bão hòa trong huấn luyện GAN, khi mô hình đã học được đầy đủ các đặc trưng chính của dữ liệu và việc huấn luyện tiếp chủ yếu chỉ mang lại những cải thiện nhỏ hoặc dao động ngẫu nhiên.

Từ các phân tích trên, checkpoint tại 500 000 iteration được lựa chọn là mô hình tối ưu để sử dụng trong các bước tiếp theo của đồ án. Việc lựa chọn này dựa trên ba cơ sở chính: (i) đạt giá trị FID trung bình thấp nhất, phản ánh chất lượng ảnh sinh tốt nhất; (ii) duy trì SSIM ở mức ổn định và phù hợp, đảm bảo tính nhất quán của khuôn mặt gốc; và (iii) thể hiện sự cân bằng tốt giữa các thuộc tính khuôn mặt khác nhau, không xảy ra hiện tượng suy giảm chất lượng rõ rệt ở một nhóm thuộc tính cụ thể.

Checkpoint 500 000 do đó được sử dụng làm mô hình cuối cùng cho quá trình xây dựng ứng dụng thử nghiệm ở Chương 6. Việc sử dụng checkpoint này giúp đảm bảo rằng hệ thống ứng dụng không chỉ tạo ra ảnh có chất lượng thị giác cao,

mà còn thể hiện rõ ràng và nhất quán các thuộc tính khuôn mặt theo yêu cầu của người dùng.

CHƯƠNG 6. XÂY DỰNG ỨNG DỤNG THỬ NGHIỆM

Sau quá trình huấn luyện, đánh giá và lựa chọn mô hình phù hợp ở các chương trước, đặc biệt là việc xác định checkpoint tối ưu tại mốc 500 000 vòng lặp huấn luyện, chương này tập trung vào việc xây dựng một ứng dụng minh họa nhằm khai thác kết quả của mô hình StarGAN trong một kịch bản sử dụng thực tế.

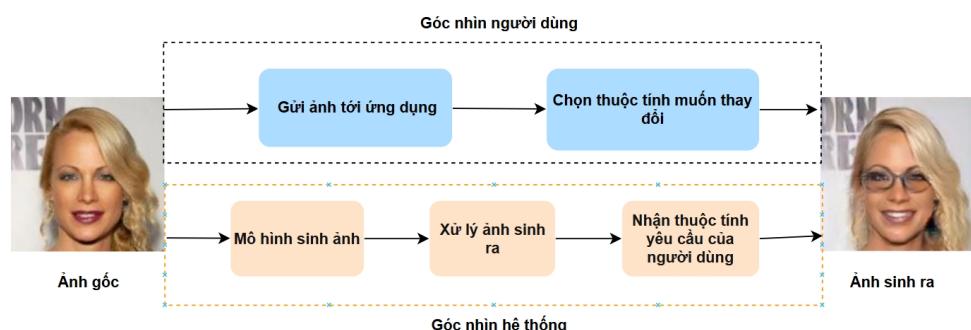
Mục tiêu của chương không chỉ dừng lại ở việc trình diễn khả năng sinh ảnh của mô hình, mà còn cho thấy tính khả dụng của hệ thống khi tích hợp vào một pipeline hoàn chỉnh, từ tiếp nhận dữ liệu đầu vào, xử lý tiền xử lý, sinh ảnh với các thuộc tính khuôn mặt mong muốn, cho đến trả về kết quả trực quan cho người dùng cuối. Ứng dụng được xây dựng theo hướng gọn nhẹ, dễ sử dụng và phù hợp cho mục đích trình diễn, thử nghiệm cũng như mở rộng trong các nghiên cứu hoặc sản phẩm sau này.

6.1 Chức năng của ứng dụng

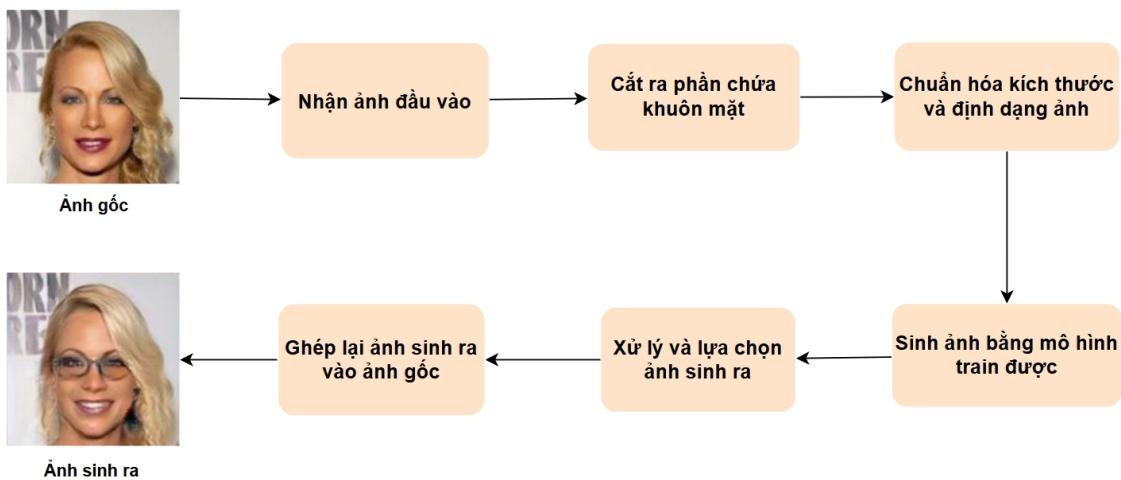
Ứng dụng được xây dựng nhằm minh họa trực tiếp khả năng thay đổi thuộc tính khuôn mặt của mô hình StarGAN đã huấn luyện. Về mặt chức năng, ứng dụng đóng vai trò là một giao diện trung gian, cho phép người dùng tương tác với mô hình sinh ảnh mà không cần can thiệp trực tiếp vào mã nguồn huấn luyện hoặc các tham số nội bộ phức tạp (Minh họa trong Hình 6.1).

Đầu vào của ứng dụng là một ảnh khuôn mặt do người dùng cung cấp. Ảnh này có thể được chụp từ thiết bị cá nhân hoặc lấy từ các nguồn ảnh thông thường, với điều kiện khuôn mặt xuất hiện rõ ràng trong khung hình. Sau khi tiếp nhận ảnh đầu vào, hệ thống thực hiện các bước xử lý cần thiết để đưa ảnh về định dạng phù hợp với mô hình đã được huấn luyện.

Bên cạnh ảnh đầu vào, người dùng lựa chọn một thuộc tính khuôn mặt mong muốn trong tập các thuộc tính đã được mô hình hỗ trợ, bao gồm các đặc trưng như



Hình 6.1: Minh họa chức năng ứng dụng

**Hình 6.2:** Các bước xử lý của hệ thống sinh ảnh

kiểu tóc, giới tính, trạng thái biểu cảm hoặc các yếu tố trang điểm. Mỗi lựa chọn tương ứng với một hướng biến đổi cụ thể trong không gian thuộc tính mà StarGAN đã học được từ dữ liệu CelebA.

Đầu ra của ứng dụng là một ảnh khuôn mặt đã được chỉnh sửa, trong đó thuộc tính được chọn đã được thay đổi theo hướng mong muốn, trong khi các đặc trưng còn lại của khuôn mặt gốc được giữ ở mức ổn định. Kết quả này cho phép người dùng quan sát trực quan hiệu quả của mô hình, đồng thời đánh giá chất lượng sinh ảnh trong bối cảnh sử dụng thực tế, thay vì chỉ thông qua các chỉ số định lượng như ở chương đánh giá.

6.2 Cốt lõi của ứng dụng

Cốt lõi của ứng dụng được xây dựng dựa trên một pipeline xử lý ảnh áp dụng chung cho cả hai hình thức triển khai (ứng dụng web và chatbot trên telegram). Pipeline này bao gồm đầy đủ các bước từ tiếp nhận ảnh đầu vào, tiền xử lý, sinh ảnh bằng mô hình StarGAN cho đến hậu xử lý và trả kết quả cho người dùng (Hình 6.2). Sự khác biệt giữa các ứng dụng chỉ nằm ở cách tổ chức giao diện và môi trường triển khai, trong khi toàn bộ luồng xử lý chính được giữ nguyên.

Tiếp nhận ảnh đầu vào

Ứng dụng cho phép người dùng tải lên một ảnh chân dung chứa khuôn mặt người. Ảnh đầu vào được tiếp nhận dưới dạng tệp ảnh chuẩn (JPEG hoặc PNG) và được lưu tạm thời trong hệ thống để phục vụ cho các bước xử lý tiếp theo. Trong quá trình này, thư viện OpenCV được sử dụng để đọc và chuyển đổi ảnh sang định dạng mảng số (numpy array), đảm bảo khả năng xử lý linh hoạt và hiệu quả.

Phát hiện và cắt khuôn mặt

Sau khi ảnh được tải lên, bước đầu tiên trong pipeline là phát hiện khuôn mặt. Ứng dụng sử dụng thư viện MediaPipe Face Detection để xác định vị trí khuôn mặt trong ảnh đầu vào. MediaPipe được lựa chọn nhờ khả năng phát hiện khuôn mặt nhanh, ổn định và hoạt động tốt trong nhiều điều kiện ánh sáng khác nhau.

Dựa trên hộp giới hạn (bounding box) do MediaPipe trả về, vùng khuôn mặt được cắt ra khỏi ảnh gốc. Để đảm bảo giữ lại đầy đủ các đặc trưng cần thiết (tóc, cằm, trán), hộp giới hạn được mở rộng thêm một tỷ lệ nhất định so với kích thước ban đầu. Việc mở rộng này giúp giảm hiện tượng mất thông tin khi đưa ảnh vào mô hình sinh.

Chuẩn hóa kích thước và định dạng ảnh

Ảnh khuôn mặt sau khi cắt được chuẩn hóa về kích thước theo đúng yêu cầu của tập dữ liệu CelebA, cụ thể là kích thước 178×218 pixel. Đây là bước quan trọng nhằm đảm bảo sự tương thích giữa ảnh đầu vào và mô hình StarGAN đã được huấn luyện.

Quá trình chuẩn hóa được thực hiện bằng thư viện OpenCV, kết hợp với các phép nội suy để hạn chế biến dạng hình học. Ảnh sau đó được lưu lại trong thư mục dữ liệu để mô hình StarGAN có thể đọc trực tiếp trong chế độ suy luận (test mode).

Sinh ảnh bằng mô hình StarGAN

Ở bước cốt lõi nhất của pipeline, mô hình StarGAN v1 được sử dụng để thực hiện bài toán thay đổi thuộc tính khuôn mặt. Mô hình được triển khai thông qua mã nguồn gốc từ repository chính thức và chạy ở chế độ test.

StarGAN nhận vào ảnh khuôn mặt đã chuẩn hóa cùng với vector thuộc tính mục tiêu, trong đó mỗi thuộc tính biểu diễn một đặc điểm khuôn mặt cụ thể (ví dụ: hói đầu, đeo kính, mỉm cười, đánh son, ...). Với mỗi thuộc tính, mô hình sinh ra một ảnh tương ứng, đảm bảo chỉ thay đổi đặc điểm được chỉ định trong khi vẫn giữ nguyên danh tính khuôn mặt.

Kết quả đầu ra của mô hình là một ảnh tổng hợp, trong đó bao gồm ảnh gốc và các ảnh sinh tương ứng với từng thuộc tính. Ảnh này sau đó được tách thành các phần riêng lẻ để phục vụ cho bước xử lý tiếp theo.

Xử lý và lựa chọn ảnh sinh

Ảnh tổng hợp đầu ra từ StarGAN được chia thành các ảnh con, mỗi ảnh biểu diễn kết quả thay đổi của một thuộc tính khuôn mặt. Trong pipeline, thư viện OpenCV

và PIL được sử dụng để cắt, chuyển đổi và lưu trữ các ảnh này.

Dựa trên lựa chọn của người dùng (ví dụ: chọn một thuộc tính cụ thể trong giao diện), hệ thống sẽ trích xuất đúng ảnh tương ứng và chuyển sang bước hậu xử lý.

Định vị và ghép ảnh vào ảnh gốc

Để tạo ra kết quả trực quan và tự nhiên, ảnh khuôn mặt sinh ra cần được ghép ngược trở lại ảnh gốc ban đầu. Do kích thước và tỷ lệ ảnh có thể đã thay đổi, hệ thống sử dụng kỹ thuật khớp đặc trưng (feature matching) dựa trên thuật toán ORB để xác định vị trí chính xác của khuôn mặt trong ảnh gốc.

Sau khi xác định được vùng tương ứng, ảnh khuôn mặt sinh được co giãn và chồng lên ảnh gốc tại đúng vị trí. Quá trình này giúp giữ nguyên bối cảnh xung quanh (nền, trang phục) và chỉ thay đổi khuôn mặt, từ đó nâng cao tính trực quan của kết quả.

Trả kết quả cho người dùng

Ảnh sau khi ghép hoàn chỉnh được chuyển sang định dạng phù hợp để hiển thị. Trong ứng dụng web, ảnh được chuyển đổi sang định dạng PIL Image và hiển thị trực tiếp trên giao diện thông qua thư viện Gradio. Người dùng có thể xem kết quả ngay sau khi chọn thuộc tính mong muốn.

Tổng thể pipeline trên đảm bảo tính thống nhất, dễ mở rộng và có thể tái sử dụng cho nhiều nền tảng khác nhau, đồng thời tận dụng hiệu quả các thư viện mã nguồn mở phổ biến trong xử lý ảnh và học sâu.

6.3 Xây dựng chatbot sử dụng mô hình được huấn luyện

Dựa trên pipeline xử lý cốt lõi đã được trình bày ở Mục 6.2, mô hình StarGAN sau huấn luyện được tích hợp vào một ứng dụng chatbot trên nền tảng Telegram nhằm minh họa khả năng ứng dụng thực tế của hệ thống. Việc lựa chọn Telegram xuất phát từ các ưu điểm như khả năng triển khai nhanh, hỗ trợ tốt cho việc truyền nhận ảnh, giao diện tương tác thân thiện và mức độ phổ biến cao đối với người dùng cuối.

Về mặt kiến trúc, chatbot được xây dựng theo mô hình client–server, trong đó Telegram đóng vai trò là giao diện tương tác phía người dùng, còn toàn bộ logic xử lý được triển khai tại phía máy chủ. Bot Telegram có nhiệm vụ tiếp nhận ảnh đầu vào và lựa chọn thuộc tính từ người dùng, sau đó chuyển dữ liệu này về máy chủ để xử lý. Các bước tiền xử lý ảnh, sinh ảnh bằng mô hình StarGAN và hậu xử lý kết quả đều được thực hiện hoàn toàn ở phía server, đảm bảo khả năng kiểm soát tài nguyên tính toán cũng như tính ổn định của hệ thống.

Quá trình xây dựng chatbot bắt đầu bằng việc đăng ký bot thông qua dịch vụ *BotFather* của Telegram để nhận *API Token*. Token này cho phép máy chủ giao tiếp với Telegram thông qua Telegram Bot API. Trên phía server, một thư viện Python chuyên dụng cho Telegram được sử dụng để lắng nghe và xử lý các sự kiện gửi tin nhắn, đặc biệt là tin nhắn chứa ảnh, đồng thời quản lý luồng hội thoại và phản hồi kết quả cho người dùng. Trong phạm vi triển khai của đồ án, cơ chế *long polling* được ưu tiên sử dụng do cấu hình đơn giản và phù hợp với môi trường thử nghiệm.

Luồng tương tác giữa người dùng và chatbot được thiết kế theo hướng tối giản. Người dùng gửi một ảnh chân dung chứa khuôn mặt người tới chatbot, sau đó lựa chọn một thuộc tính khuôn mặt cần thay đổi thông qua danh sách các thuộc tính được bot cung cấp. Ngay khi nhận được ảnh đầu vào, chatbot tải ảnh về máy chủ và chuyển ảnh này vào pipeline xử lý cốt lõi. Pipeline này bao gồm các bước phát hiện khuôn mặt bằng thư viện MediaPipe, cắt và mở rộng vùng khuôn mặt, chuẩn hóa kích thước ảnh theo định dạng của bộ dữ liệu CelebA, sinh ảnh bằng mô hình StarGAN đã được huấn luyện với checkpoint tối ưu, và cuối cùng ghép khuôn mặt đã chỉnh sửa trở lại ảnh gốc.

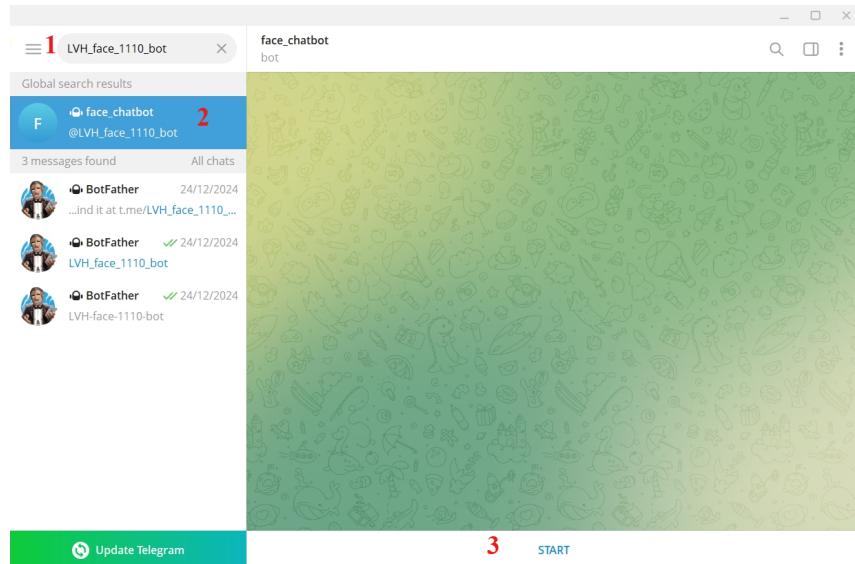
Thuộc tính khuôn mặt do người dùng lựa chọn được ánh xạ trực tiếp tới vector điều kiện đầu vào của StarGAN. Mỗi lựa chọn tương ứng với một hướng biến đổi cụ thể trong không gian thuộc tính, đảm bảo rằng mô hình chỉ thay đổi thuộc tính mong muốn trong khi vẫn giữ được các đặc trưng nhận dạng chính của khuôn mặt. Việc tái sử dụng trực tiếp pipeline đã xây dựng giúp chatbot kế thừa đầy đủ chất lượng sinh ảnh và tính ổn định của hệ thống, đồng thời giảm thiểu đáng kể công sức phát triển.

Sau khi quá trình sinh ảnh hoàn tất, chatbot gửi ảnh kết quả về cho người dùng dưới dạng ảnh tiêu chuẩn của Telegram, kèm theo mô tả ngắn gọn về thuộc tính đã được áp dụng. Thiết kế này hướng tới trải nghiệm người dùng trực quan, không yêu cầu kiến thức chuyên sâu về trí tuệ nhân tạo hay xử lý ảnh, trong khi toàn bộ quá trình xử lý phức tạp được ẩn phía sau hệ thống.

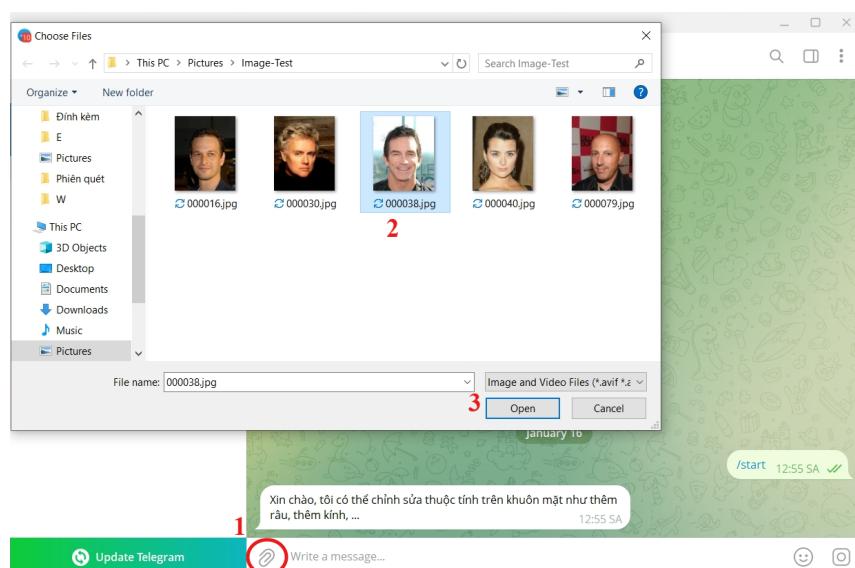
Dưới đây sẽ là phần hình ảnh minh họa cho việc sử dụng chatbot.

Người dùng có thể tìm kiếm chatbot trên Telegram với tên: LVH_face_1110_bot và bắt đầu trải nghiệm bot chỉnh sửa ảnh khuôn mặt (Hình 6.3). Người dùng có thể gửi ảnh chứa khuôn mặt từ thiết bị cho bot (Hình 6.4), sau đó bot tự động sinh ra kho ảnh với 17 thuộc tính khác nhau. Người dùng có thể chọn ảnh trả về từ danh sách các tiêu chí hiển thị (Hình 6.5). Cuối cùng, chatbot trả về ảnh kết quả cho người dùng (Hình 6.6).

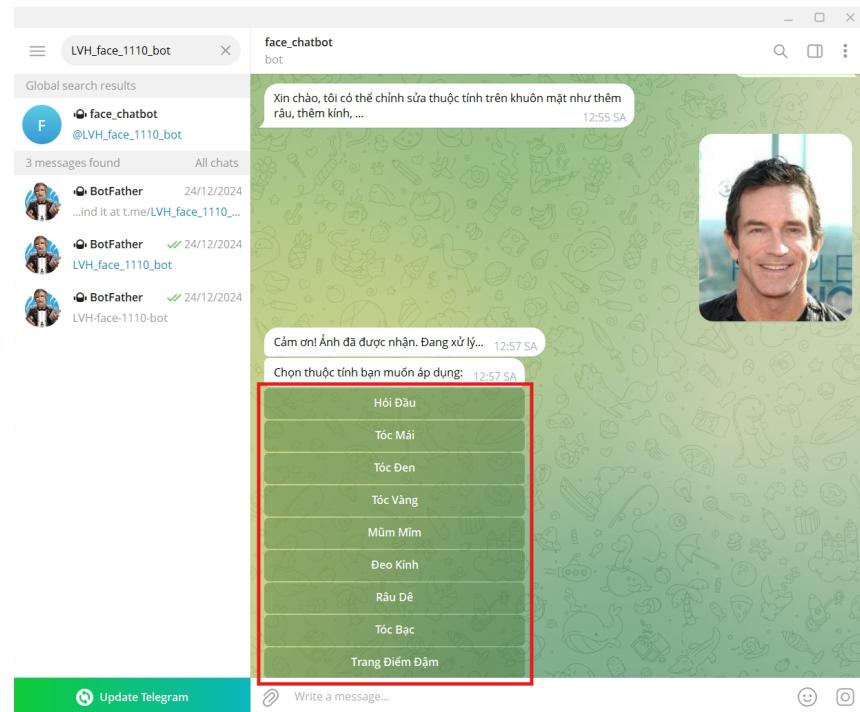
Việc triển khai chatbot Telegram cho thấy mô hình StarGAN sau huấn luyện



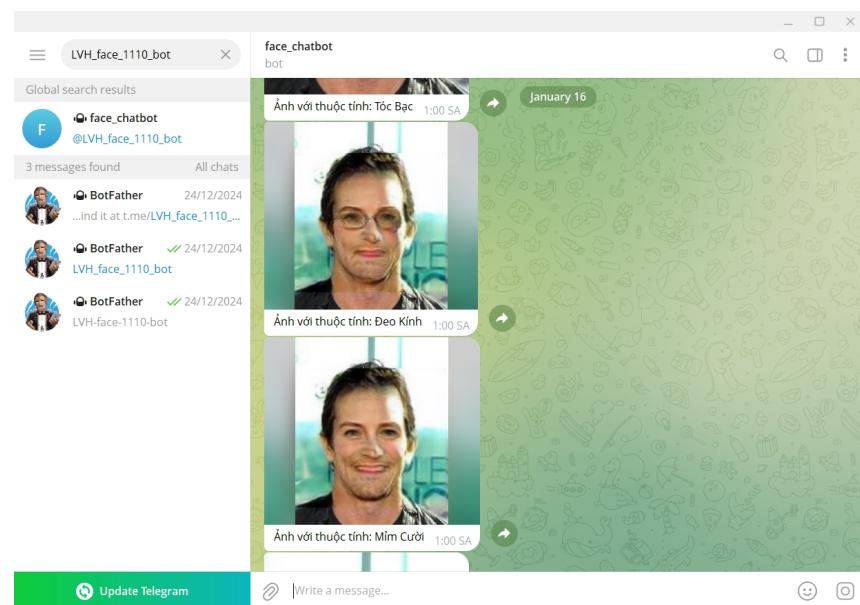
Hình 6.3: Minh họa sử dụng chatbot: tìm kiếm bot trên Telegram.



Hình 6.4: Minh họa sử dụng chatbot: chọn ảnh đầu vào.



Hình 6.5: Minh họa sử dụng chatbot: chọn tiêu chí.



Hình 6.6: Minh họa sử dụng chatbot: nhận ảnh kết quả.

không chỉ dừng lại ở mức nghiên cứu học thuật mà hoàn toàn có thể tích hợp vào các ứng dụng tương tác thực tế. Với cùng một cốt lõi xử lý, hệ thống có thể dễ dàng mở rộng sang các nền tảng khác như ứng dụng web hoặc ứng dụng di động, chỉ bằng cách thay đổi lớp giao diện và cơ chế giao tiếp với người dùng. Qua đó, chatbot đóng vai trò như một minh chứng rõ ràng cho tính ứng dụng và khả năng triển khai thực tiễn của mô hình được đề xuất trong đồ án.

6.4 Xây dựng web site sử dụng mô hình được huấn luyện

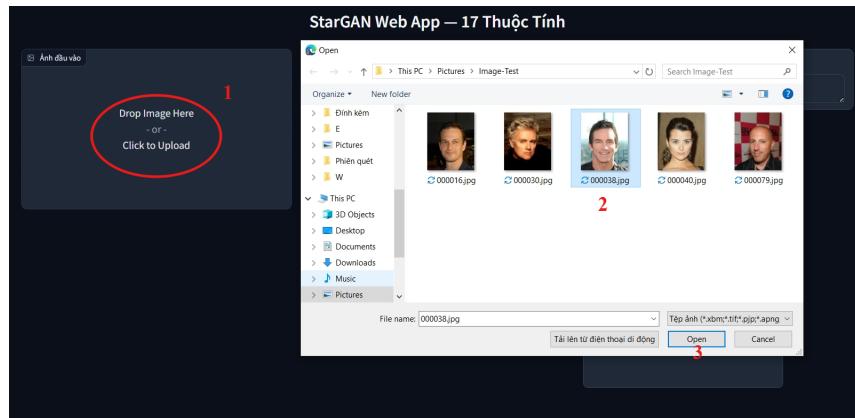
Bên cạnh hình thức triển khai dưới dạng chatbot, mô hình StarGAN sau khi huấn luyện cũng được tích hợp vào một ứng dụng web nhằm cung cấp giao diện trực quan hơn cho người dùng cuối. Ứng dụng web cho phép người dùng tương tác trực tiếp với hệ thống thông qua trình duyệt, không phụ thuộc vào nền tảng nhắn tin, đồng thời phù hợp cho mục đích trình diễn, kiểm thử và mở rộng trong tương lai.

Về tổng thể, website được xây dựng dựa trên cùng pipeline xử lý cốt lõi đã trình bày trong Mục 6.2. Sự khác biệt chính nằm ở lớp giao diện và cơ chế tương tác với người dùng. Thay vì nhận dữ liệu đầu vào từ tin nhắn Telegram, hệ thống web tiếp nhận ảnh thông qua chức năng tải ảnh (upload) và lựa chọn thuộc tính khuôn mặt thông qua các thành phần giao diện đồ họa.

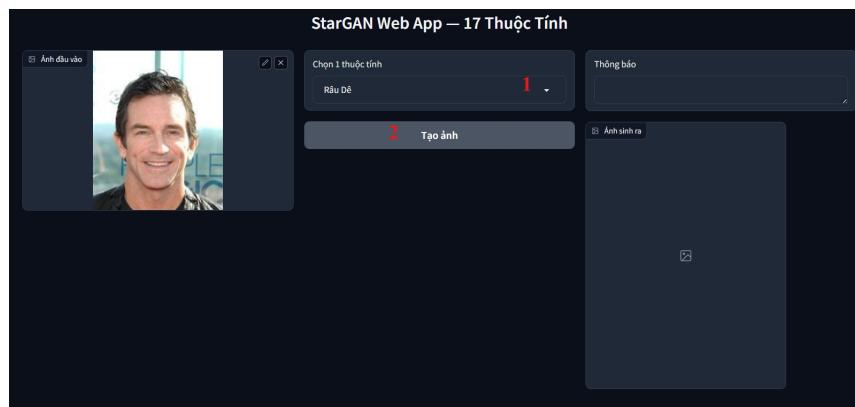
Cụ thể, ứng dụng web được triển khai bằng thư viện Gradio, một framework hỗ trợ xây dựng nhanh các giao diện web cho mô hình học sâu. Gradio cho phép kết nối trực tiếp giữa giao diện người dùng và các hàm xử lý Python, nhờ đó quá trình tích hợp mô hình StarGAN trở nên đơn giản và linh hoạt. Người dùng chỉ cần tải lên một ảnh khuôn mặt và chọn một thuộc tính mong muốn trong danh sách các thuộc tính hỗ trợ; hệ thống sẽ tự động kích hoạt pipeline xử lý phía sau và trả về kết quả tương ứng.

Quy trình xử lý trong website có thể được mô tả theo các bước sau. Trước hết, ảnh đầu vào do người dùng tải lên được lưu tạm thời trên máy chủ và chuyển vào pipeline tiền xử lý. Tại đây, thư viện MediaPipe được sử dụng để phát hiện khuôn mặt trong ảnh và xác định vùng bao quanh khuôn mặt. Vùng này được cắt ra, mở rộng theo một tỷ lệ nhất định để đảm bảo giữ lại các đặc trưng cần thiết, sau đó được chuẩn hóa kích thước theo định dạng dữ liệu huấn luyện của CelebA.

Tiếp theo, ảnh khuôn mặt đã chuẩn hóa được đưa vào mô hình StarGAN đã huấn luyện sẵn, với checkpoint được lựa chọn là checkpoint tại 500 000 vòng lặp huấn luyện. Việc sử dụng checkpoint này dựa trên kết quả đánh giá ở Chương 5, trong đó mô hình đạt giá trị FID thấp nhất và chất lượng ảnh sinh ổn định nhất. StarGAN thực hiện sinh ra một tập các ảnh tương ứng với các thuộc tính khuôn mặt khác



Hình 6.7: Minh họa sử dụng web: chọn ảnh đầu vào.



Hình 6.8: Minh họa sử dụng web: chọn tiêu chí và sinh ảnh.

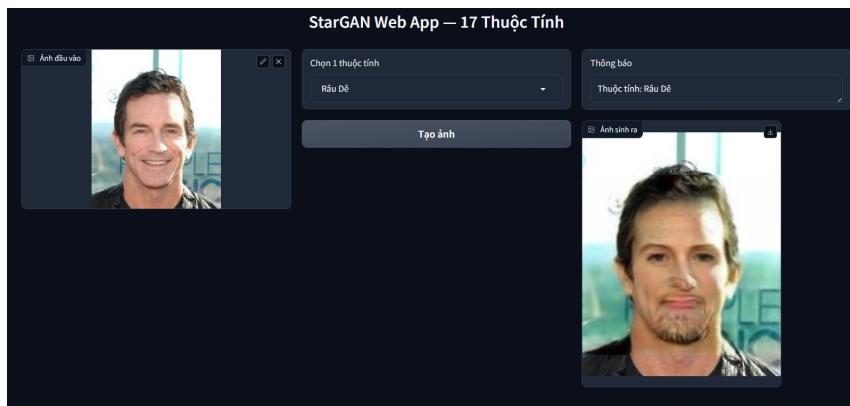
nhau, trong đó mỗi ảnh biểu diễn sự thay đổi của một thuộc tính cụ thể so với ảnh gốc.

Sau khi quá trình sinh ảnh hoàn tất, hệ thống lựa chọn ảnh tương ứng với thuộc tính mà người dùng đã chọn trên giao diện web. Ảnh sinh này sau đó được xử lý hậu kỳ để ghép ngược trở lại ảnh gốc. Quá trình ghép sử dụng các phép biến đổi hình học nhằm đảm bảo vị trí và tỷ lệ khuôn mặt sinh ra phù hợp với ảnh ban đầu, từ đó tạo ra kết quả cuối cùng có tính tự nhiên cao.

Cuối cùng, ảnh kết quả được chuyển đổi về định dạng phù hợp để hiển thị trên trình duyệt và trả về cho người dùng thông qua giao diện Gradio. Toàn bộ quá trình, từ lúc người dùng tải ảnh đến khi nhận được ảnh kết quả, được thực hiện một cách tự động và gần như theo thời gian thực, mang lại trải nghiệm tương tác trực quan và thuận tiện.

Dưới đây sẽ là phần hình ảnh minh họa cho việc sử dụng trang web ứng dụng.

Như trên Hình 6.7, người dùng có thể chọn ảnh từ thiết bị hoặc kéo ảnh vào vùng nhận ảnh đầu vào. Sau đó chọn tiêu chí sinh ảnh và ấn nút tạo ảnh để bắt đầu quá trình sinh ảnh (Hình 6.8). Cuối cùng, ảnh được sinh ra với đúng tiêu chí người



Hình 6.9: Minh họa sử dụng web: nhận thông báo và ảnh sinh ra.

dùng lựa chọn và ảnh có thể tải về thiết bị (Hình 6.9).

Như vậy, việc xây dựng website sử dụng mô hình đã huấn luyện không chỉ giúp minh họa rõ ràng khả năng của StarGAN trong bài toán thay đổi thuộc tính khuôn mặt, mà còn chứng minh tính ứng dụng thực tiễn của mô hình trong các hệ thống tương tác trực tiếp với người dùng. Đồng thời, cách triển khai này cho thấy pipeline cốt lõi có thể dễ dàng được tái sử dụng và mở rộng sang nhiều nền tảng khác nhau mà không cần thay đổi bản chất của mô hình.

CHƯƠNG 7. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

7.1 Kết luận

Trong đồ án này, hệ thống thay đổi thuộc tính khuôn mặt dựa trên mô hình mạng sinh đồi kháng StarGAN đã được nghiên cứu, triển khai và đánh giá một cách toàn diện. Từ việc khảo sát bài toán chỉnh sửa thuộc tính khuôn mặt đa miền, đồ án đã lựa chọn StarGAN như một mô hình phù hợp nhờ khả năng xử lý nhiều thuộc tính trong một kiến trúc thông nhất, giảm đáng kể chi phí huấn luyện so với các mô hình GAN truyền thống.

Quá trình huấn luyện mô hình được thực hiện trên bộ dữ liệu CelebA với 17 thuộc tính khuôn mặt phổ biến. Mô hình được huấn luyện trong nhiều giai đoạn, sử dụng cơ chế checkpoint nhằm thích nghi với các ràng buộc về tài nguyên tính toán. Kết quả thực nghiệm cho thấy mô hình có khả năng học được các đặc trưng chính của từng thuộc tính, tạo ra các ảnh khuôn mặt biến đổi hợp lý trong khi vẫn duy trì được cấu trúc khuôn mặt ban đầu.

Bên cạnh đánh giá định tính thông qua quan sát trực quan, đồ án đã sử dụng các chỉ số đánh giá định lượng phổ biến là SSIM và FID để phân tích chất lượng ảnh sinh ra. Kết quả cho thấy checkpoint tại mốc 500 000 bước huấn luyện đạt được sự cân bằng tốt giữa tính chân thực và mức độ bảo toàn nội dung ảnh, từ đó được lựa chọn làm mô hình cuối cùng phục vụ cho việc xây dựng ứng dụng.

Trên cơ sở mô hình đã huấn luyện, đồ án tiếp tục triển khai hai ứng dụng minh họa gồm chatbot trên nền tảng Telegram và ứng dụng web. Cả hai ứng dụng đều sử dụng chung một pipeline xử lý cốt lõi, bao gồm phát hiện khuôn mặt, tiền xử lý ảnh, sinh ảnh bằng StarGAN và hậu xử lý kết quả. Việc triển khai thành công các ứng dụng này cho thấy mô hình không chỉ có giá trị nghiên cứu mà còn có khả năng ứng dụng thực tế.

Nhìn chung, đồ án đã đạt được mục tiêu đề ra ban đầu, từ nghiên cứu lý thuyết, triển khai mô hình, đánh giá kết quả cho tới xây dựng các ứng dụng minh họa, qua đó khẳng định tính khả thi của việc áp dụng StarGAN cho bài toán thay đổi thuộc tính khuôn mặt.

7.2 Hạn chế của đồ án hiện tại

Mặc dù đạt được những kết quả nhất định, đồ án vẫn còn tồn tại một số hạn chế.

Trước hết, chất lượng ảnh sinh ra trong một số trường hợp vẫn chưa thực sự thuyết phục, đặc biệt đối với các thuộc tính có sự thay đổi mạnh về hình dạng hoặc màu sắc như tóc, râu hoặc trang điểm đậm. Một số ảnh sinh xuất hiện hiện tượng

nhiều nhẹ, chi tiết khuôn mặt chưa sắc nét, hoặc mất tính tự nhiên khi so sánh với ảnh gốc.

Bên cạnh đó, do pipeline xử lý bao gồm bước cắt khuôn mặt và ghép ngược kết quả vào ảnh ban đầu, nên trong một số trường hợp có thể quan sát được viền ghép tương đối rõ. Điều này đặc biệt dễ nhận thấy khi điều kiện ánh sáng của ảnh đầu vào phức tạp hoặc khi khuôn mặt bị xoay lệch nhiều so với tư thế chuẩn trong tập huấn luyện.

Một hạn chế khác đến từ điều kiện huấn luyện mô hình. Việc sử dụng môi trường Kaggle với giới hạn về thời gian và tài nguyên GPU khiến quá trình thử nghiệm và tối ưu siêu tham số chưa thể thực hiện một cách đầy đủ. Mô hình chưa được huấn luyện với độ phân giải cao hơn hoặc với số vòng lặp lớn hơn để khai thác hết tiềm năng của kiến trúc StarGAN.

Ngoài ra, hệ thống hiện tại chỉ hỗ trợ thay đổi từng thuộc tính đơn lẻ tại một thời điểm. Việc kết hợp đồng thời nhiều thuộc tính vẫn còn hạn chế và có thể dẫn tới hiện tượng xung đột giữa các thuộc tính trong ảnh sinh ra.

7.3 Hướng phát triển

Trong tương lai, đồ án có thể được mở rộng và cải thiện theo nhiều hướng khác nhau.

Một hướng phát triển quan trọng là nâng cao chất lượng ảnh sinh ra. Điều này có thể đạt được bằng cách sử dụng các biến thể cải tiến của StarGAN như StarGAN v2, hoặc kết hợp thêm các kỹ thuật huấn luyện hiện đại như perceptual loss, adaptive instance normalization hoặc các cơ chế attention để mô hình tập trung tốt hơn vào vùng khuôn mặt cần chỉnh sửa.

Bên cạnh đó, pipeline hậu xử lý có thể được cải tiến nhằm giảm hiện tượng viền ghép và tăng tính liền mạch giữa khuôn mặt đã chỉnh sửa và ảnh gốc. Việc áp dụng các phương pháp làm mịn biên, hòa trộn màu sắc hoặc sử dụng mô hình segmentation chính xác hơn là những hướng nghiên cứu tiềm năng.

Về mặt dữ liệu, hệ thống có thể được mở rộng sang các bộ dữ liệu lớn hơn và đa dạng hơn, bao gồm nhiều độ tuổi, chủng tộc và điều kiện ánh sáng khác nhau. Điều này giúp mô hình có khả năng tổng quát tốt hơn khi áp dụng vào các tình huống thực tế.

Cuối cùng, về mặt ứng dụng, hệ thống có thể được phát triển thành các nền tảng hoàn chỉnh hơn như ứng dụng di động hoặc dịch vụ trực tuyến, cho phép người dùng tương tác linh hoạt với nhiều thuộc tính cùng lúc. Việc tích hợp thêm các cơ chế cá nhân hóa hoặc chỉnh sửa mức độ tác động của từng thuộc tính cũng là một

hướng phát triển đáng chú ý.

Tổng thể, mặc dù còn một số hạn chế, đồ án đã đặt nền tảng vững chắc cho việc nghiên cứu và ứng dụng các mô hình GAN trong bài toán chỉnh sửa khuôn mặt, đồng thời mở ra nhiều hướng phát triển tiềm năng trong tương lai.

TÀI LIỆU THAM KHẢO

- [1] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. Pearson, 2002.
- [2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, “Generative adversarial nets,” *arXiv preprint arXiv:1406.2661*, 2014, Available at: <https://arxiv.org/pdf/1406.2661.pdf>.
- [3] e. a. Trần, “Ten years of generative adversarial nets (gans): A survey of the state of the art,” *arXiv preprint arXiv:2308.16316*, 2023, Available at: <https://arxiv.org/abs/2308.16316>.
- [4] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, “Attgan: Facial attribute editing by only changing what you want,” *arXiv preprint arXiv:1711.10678*, 2017, Available at: <https://arxiv.org/pdf/1711.10678.pdf>.
- [5] J. Tang, R. Wang, X. Liang, P. Yan, and J. Feng, “Maggan: High-resolution face attribute editing with mask-guided generative adversarial network,” *arXiv preprint arXiv:1909.07330*, 2019, Available at: <https://arxiv.org/pdf/1909.07330.pdf>.
- [6] J.-Y. Zhu and et al., “Unpaired image-to-image translation using cycle-consistent adversarial networks,” *arXiv preprint*, vol. arXiv:1703.10593, 2017.
- [7] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” *arXiv preprint*, vol. arXiv:1711.09020, 2018. [Online]. Available: <https://arxiv.org/abs/1711.09020>.
- [8] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, “Stargan v2: Diverse image synthesis for multiple domains,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [10] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *arXiv preprint arXiv:1706.08500*, 2017, Available at: <https://arxiv.org/pdf/1706.08500.pdf>.