ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

HIGH RESOLUTION FACE AGE EDITING

Thành viên nhóm:

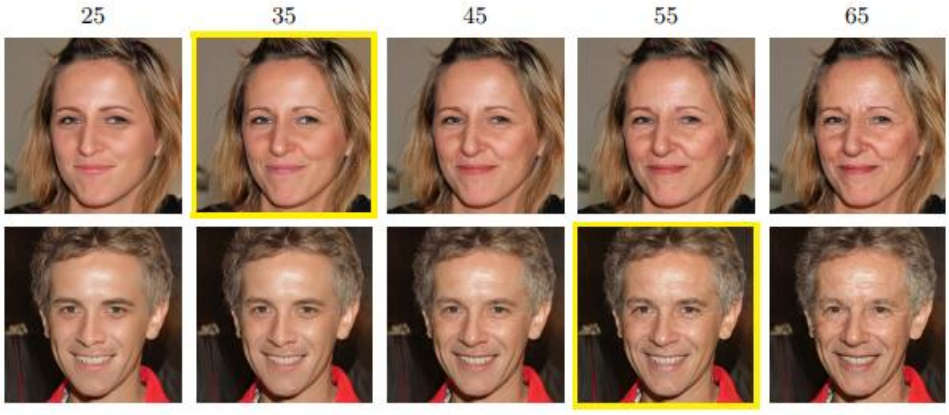
1. Trần Đắc Khoa-20120513
2. Phạm Trương Quang Khoa-20120512
3. Nguyễn Đình Lộc-20120522
4. Lê Mai Khôi Nguyên-20120538

MÔN: NHẬN DẠNG

Thành phố Hồ Chí Minh – 2023

1. Introduce

Điều chỉnh tuổi của khuôn mặt là một chủ đề quan trọng cả trong ngành công nghiệp lẫn học thuật. Trong ngành công nghiệp phim điện ảnh, nhiều diễn viên được chỉnh sửa theo một cách nào đó, có thể là để làm cho đẹp hơn hoặc chỉnh sửa cấu trúc của da. Cụ thể hơn, hiệu ứng làm già đi hoặc trẻ hóa thường được tạo ra bằng việc trang điểm hoặc hiệu ứng hình ảnh đặc biệt. Mặc dù có thể đạt được kết quả ấn tượng thông qua công nghệ kĩ thuật số, như trong bộ phim The Curious Case of Benjamin Button , Brad Pitt đã được trẻ hóa và già hóa khuôn mặt trong suốt bộ phim và việc này rất tốn thời gian. Vì vậy, cần một thuật toán mạnh mẽ và chất lượng cao để thực hiện điều chỉnh tuổi tự động là việc rất cần thiết. Tuy nhiên, chỉnh sửa khuôn mặt là một nhiệm vụ khó khăn trước giờ. Thực tế, não bộ của con người đặc biệt giỏi trong việc nhận biết các đặc điểm của khuôn mặt để phát hiện, nhận dạng hoặc phân tích chúng, ví dụ như suy luận về danh tính của một người hoặc cảm xúc của họ. Do đó, những vết xước nhỏ cũng được nhận biết ngay lập tức và làm hỏng cảm nhận về kết quả. Vì lý do này, mục tiêu của chúng ta là tạo ra kết quả không có vết xước, sắc nét và chân thực nhất trên hình ảnh khuôn mặt có độ phân giải cao.



Hình 1. Kết quả chỉnh sửa tuổi trên hình ảnh 1024 × 1024. Chúng ta đề xuất một mô hình học sâu biến đổi tuổi có thể thực hiện cả lão hóa và trẻ hoá khuôn mặt, tạo ra hình ảnh chất lượng cao, sắc nét và ít giả. Bằng cách sử dụng hình ảnh khuôn mặt được biểu thị bằng khung màu vàng làm đầu vào, mô hình của chúng ta có thể xuất ra hình ảnh chân thực của cùng một người ở bất kỳ độ tuổi mục tiêu bắt buộc nào trong phạm vi {20, . . . , 69}.

Với sự thành công của Generative Adversarial Networks (GANs) trong việc tạo ra hình ảnh chất lượng cao, các mô hình dựa trên GAN đã được sử dụng rộng rãi cho việc chuyển đổi hình ảnh từ hình ảnh này sang hình ảnh khác như [35,40]. Mặc dù đã đặt ra tiêu chuẩn mới cho việc tổng hợp hình ảnh tự nhiên, GANs được biết đến có hai nhược điểm chính: sự xuất hiện quá nhiều vết xước nhỏ và sự không ổn định của quá trình training. Các nghiên cứu về già hóa khuôn mặt

gần đây như [9,20,33,36,39] cũng áp dụng các mô hình dựa trên GANs. Cụ thể, họ chia tập dữ liệu khuôn mặt thành các nhóm tuổi khác nhau, đưa hình ảnh khi còn trẻ vào bộ tạo và dựa vào bộ phân biệt để ánh xạ hình ảnh đầu ra thành các nhóm tuổi lớn hơn. Phương pháp này có nhiều hạn chế. Đầu tiên, như dự kiến, các phương pháp này thừa hưởng các nhược điểm của các phương pháp dựa trên GANs - nền bị nhiễu mờ, cấu trúc phụ động nhỏ, không ổn định trong quá trình đào tạo. Thứ hai, vì hiệu ứng già hoá được tạo ra bằng cách phân phối hình ảnh đầu ra với nhóm tuổi mục tiêu, những phương pháp này chỉ giới hạn trong việc già hóa/trẻ hoá tương đối thô sơ. Để đạt được sự biến đổi tinh vi, cần huấn luyện một mô hình riêng biệt giữa mỗi cặp tuổi.

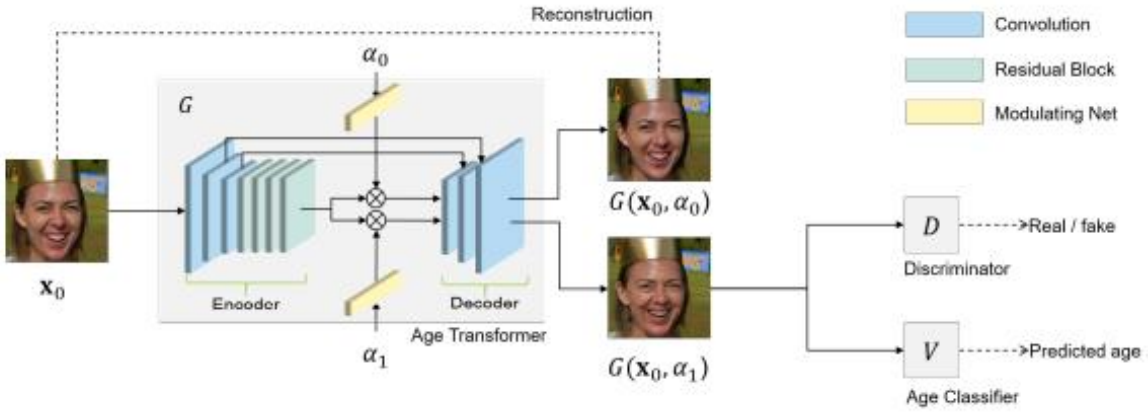
Trong đề tài này, chúng ta đề xuất một mô hình mã hóa-giải mã cho vấn đề chỉnh sửa tuổi của khuôn mặt với chất lượng hình ảnh cao trên hình ảnh có độ phân giải cao. Để giải quyết các hạn chế đã đề cập, như các nhược điểm tạo ra vết xước trên hình ảnh và không ổn định trong quá trình đào tạo, chúng ta đã cố gắng giữ cho cấu trúc đơn giản nhất có thể. Đầu tiên, chúng ta sử dụng một mạng duy nhất cho cả già hóa và trẻ hoá. Điều này là hợp lý vì phần mã hóa của mô hình chúng ta được cho là mã hóa danh tính, cảm xúc hoặc chi tiết trong hình ảnh đầu vào và không liên quan đến tuổi, để có thể sử dụng không gian tiềm ẩn giống nhau cho cả hai nhiệm vụ là già hóa và trẻ hoá. Thứ hai, chúng ta dựa vào một lớp điều chỉnh đặc trưng, có kích thước nhỏ, hoạt động trực tiếp trên không gian tiềm ẩn và cho phép chuyển đổi tuổi liên tục. Thứ ba, khác với các phương pháp cạnh tranh khác nơi bộ phân biệt được sử dụng trong quá trình huấn luyện giải quyết được điều kiện tuổi của mục tiêu, chúng ta sử dụng một bộ phân biệt không cần điều kiện và chỉ tập trung vào tính chân thực của hình ảnh đầu ra để giảm thiểu các vết xước trên ảnh. Bộ phân biệt có thể được coi là một bộ ràng buộc thông thường áp đặt tính chân thực ngoài việc cố gắng khớp hai phân phối lại với nhau. Nhờ thiết kế này, mô hình của chúng ta đạt được phân tách hiệu quả giữa thuộc tính tuổi và đặc điểm danh tính của khuôn mặt. Chúng ta trình bày kết quả thực nghiệm trên hình ảnh có độ phân giải cao với đánh giá chất lượng chính xác cả về định tính và định lượng. Đặc biệt, các thử nghiệm này cung cấp bằng chứng rõ ràng cho thấy chất lượng hình ảnh đạt được bởi kết quả của mô hình chúng ta vượt trội hơn so với các phương pháp hiện tại. Các thử nghiệm trên các tập dữ liệu thay thế cũng minh họa khả năng tổng quát hóa của phương pháp.

2. Related work

Lão hoá khuôn mặt: Công trình khảo sát [6] đã cung cấp một cái nhìn toàn diện về các thuật toán tổng hợp tuổi truyền thống. Tuy nhiên, những tiến bộ gần đây trong deep learning đã đạt được sự đáng kể trong việc làm già khuôn mặt bằng các phương pháp dựa trên deep learning. Một trong những phương pháp tiếp cận sớm nhất là mô hình GAN có điều kiện [24] cho nhiệm vụ làm già khuôn mặt, trong đó hình ảnh khuôn mặt được mã hóa vào không gian ẩn, được điều chỉnh và sau đó được giải mã để tạo ra một khuôn mặt già bằng bộ tạo. Tuy nhiên, quá trình này thường làm hỏng thông tin về danh tính của khuôn mặt. Các cải tiến tiếp theo như [36,38] đã thêm một thuật ngữ bảo toàn danh tính vào mục tiêu, nhưng kết quả của họ đã bị làm mờ so với hình ảnh ban đầu. Các mô hình tổng hợp dựa trên wavelet [19,20] cũng được giới thiệu để bắt đầu kỹ thuật chi tiết

về cấu trúc, nhưng các mô hình phức tạp này tăng độ khó trong quá trình huấn luyện và vẫn tạo ra nhiều lỗi. Hầu hết các mô hình hiện có cho việc già hoá khuôn mặt chỉ cho phép làm già từ một nhóm tuổi này sang nhóm tuổi khác, ví dụ từ 20 đến 40 tuổi, nó bị thiếu tính linh hoạt. Gần đây, [9] đề xuất một mạng mã hóa-giải mã, trong đó một cơ sở già hóa được cá nhân hóa và một phép biến đổi tuổi cụ thể được áp dụng. Mô hình của họ cũng dựa trên một bộ phân biệt có điều kiện để phân biệt các mẫu già hóa giữa các nhóm tuổi khác nhau. Khác với các phương pháp khác, mô hình của chúng ta được thiết kế cho việc chỉnh sửa tuổi với một độ tuổi mục tiêu ngẫu nhiên. Hơn nữa, phương pháp của chúng ta tạo ra ít lỗi hơn, làm cho việc chỉnh sửa tuổi trên hình ảnh có độ phân giải cao cụ thể là kích thước (1024 × 1024) trở nên khả thi.

Chuyển từ hình ảnh sang hình ảnh: Làm già khuôn mặt có thể được coi là một vấn đề dịch hình ảnh sang hình ảnh, tức là chuyển đổi hình ảnh giữa các miền tuổi: trẻ và già. Một số phương pháp đã sử dụng tiếp cận dựa trên tối ưu hóa như [34], tận dụng sự nội suy tuyến tính của các đặc trưng sâu từ các mạng convnets được tiền huấn luyện để biến đổi hình ảnh. Các phương pháp dựa trên GAN khác như [13,40,11] cho phép dịch hình ảnh thời gian thực bằng cách huấn luyện một bộ tạo feed-forward. Các nghiên cứu hiện có về dịch hình ảnh sang hình ảnh [3,4,18,29,30,37] trên hình ảnh khuôn mặt cũng đạt được kết quả ấn tượng trong việc điều chỉnh các đặc điểm khuôn mặt. Ví dụ, Lample et al. [18] thiết kế một kiến trúc autoencoder để tái tạo hình ảnh và phân tách các đặc điểm hình ảnh đơn lẻ trong thành phần ẩn bằng một bộ phân biệt. Những đặc điểm này sau đó có thể được chỉnh sửa trực tiếp trong không gian ẩn. Choi et al. [4] đề xuất một phương pháp để thực hiện dịch hình ảnh sang hình ảnh cho nhiều miền bằng chỉ một mô hình duy nhất. Pumarola et al. [29] giới thiệu một mô hình dựa trên sự chú ý, cho phép tạo hiệu ứng hoạt hình khuôn mặt thông qua việc nội suy đơn giản.



Hình 2. ***Quá trình train:*** *ảnh đầu vào x0 được chỉnh sửa bởi generator G. xử dụng tuổi ban đầu a0 (nhiệm vụ tái tạo) và a1 (nhiệm vụ điều chỉnh). Ảnh tái tạo e G(x0, α0) nên giống như ảnh đầu vào. Ảnh chỉnh sửa G(x0, α1) được chuyển tiếp vào bộ phân biệt D để đảm bảo tính chất quang học của hình ảnh và bộ phân lớp theo độ tuổi V đảm bảo chuyển đổi chỉnh xác theo độ tuổi.* ***Mô hình chuyển đổi tuổi mạng sinh dữ liệu*** *G gồm 3 lớp mạng một lớp mã hóa, một lớp mạng hiệu chỉnh(*) và một lớp mạng giải mã. Bộ mã hóa ánh xạ ảnh x0 vào một không gian đặc trưng không liên quan đến tuổi. Bộ hiệu chỉnh ánh xạ độ tuổi a đến 1 vector điều chỉnh có 128 chiều. Vector này được dùng để điều chỉnh từng kênh của các đặc trưng vừa được mã hóa ở bộ trước, từ đó áp dụng sự biến đổi độ tuổi tương ứng. Các đặc trưng đã được hiệu chỉnh, cuối cùng đi qua lớp giải mã để cho ra những bức ảnh được thay đổi. Hai kết nối tắt giữa bộ mã hóa và bộ giải mã để giữ các đặc trưng không liên quan đến tuổi tốt hơn.*

(*) Mạng hiệu chỉnh trong học sâu là một kỹ thuật liên quan đến việc sửa đổi kích hoạt các nơ-ron trong mạng nơ-ron để cải thiện hiệu suất của nó. Việc điều chỉnh có thể được thực hiện bằng cách thêm một thông tin bổ sung vào mạng, thông tin này có thể được sử dụng để kiểm soát việc kích hoạt các nơ-ron.

Tổng hợp hình ảnh độ phân giải cao: Mặc dù các phương pháp gần đây đã tiến bộ trong việc chỉnh sửa hình ảnh tự nhiên và chỉnh sửa hình ảnh có độ phân giải cao vẫn là một thách thức lớn. Tuy nhiên, trong việc tạo ra hình ảnh, đã có bước tiến lớn đó là đạt được hình ảnh chất lượng cao ở độ phân giải cao. Ví dụ, sự phát triển tiến bộ của các mạng GAN theo tiến trình [15] đã tạo rs hình ảnh ở độ phân giải 1024×1024. Hoặc StyleGAN [16,17] cải thiện chất lượng của hình ảnh được tạo bằng cách tự động học tách biệt các thuộc tính cấp cao trong quá trình huấn luyện. Dựa trên công việc này, Shen et al. [32] đề xuất một cách hiệu quả để giải thích không gian ẩn được học bởi bộ tạo và đạt được sự chỉnh sửa khuôn mặt có độ trung thực cao trên hình ảnh tổng hợp. Tuy nhiên, theo những thử nghiệm của chúng ta, chỉ một phần nhỏ hình ảnh tự nhiên có thể tái tạo chính xác với mã ẩn, điều này làm cho loại phương pháp này trở nên không thực tế. Ngược lại, phương pháp mà chúng ta đề xuất đạt được chỉnh sửa tuổi trên hình ảnh 1024 × 1024, với cấu trúc và thiết kế bớt phức tạp. Chỉnh sửa tuổi được thực hiện chỉ bằng một mạng phụ trợ modulating, có thể tổng quát cho các nhiệm vụ chỉnh sửa khuôn mặt khác.

3. Phương pháp

Trong phần này, chúng ta sẽ trình bày chi tiết về vấn đề chỉnh sửa khuôn mặt theo độ tuổi và trình bày chi tiết mô hình. Hình 2 minh hóa khá đầy đủ quá trình chuyển đổi và huấn luyện.

3.1. Tổng quan:

X0 là một bức ảnh ngẫu nhiên được lấy từ tập dữ liệu. Ta chỉ ra độ tuổi α0 của người x0 này. Mục đích là chuyển đổi x0 sao cho khuôn mặt trong bức ảnh này giống như người đang ở độ tuổi a1. Chúng ta đồng thời mong muốn hai bức ảnh này có nhiều nét tương đồng nhất (có chung càng nhiều đặc trưng không liên quan đến tưởi càng tốt), chẳng hạn như giới tính, biểu cảm, kiểu tóc, nền, ... Điều đó có nghĩa rằng: những đặc trưng của khuôn mặt không liên quan tới độ tuổi, cũng như nền, sẽ được giữ trong suốt quá trình chuyển đổi. Vì vậy ta có thể giả định rằng, mô hình biến đổi và khử biến đổi khuôn mặt sẽ có hầu hết những tham số chung. Trong cài đặt này, chúng ta

xem mạng sinh dữ liệu (G) của mo hình chuyển đổi có thể biến đổi ảnh đến bất kì độ tuổi a nào. Đầu vào của mô hình là hình ảnh khuôn mặt x0 và độ tuổi mục tiêu α1. Đầu ra được ký hiệu bởi G(x0, α1), mô tả x0 ở độ tuổi mục tiêu α1.

3.2. Biến đổi tuổi

Mô hình 2 biểu diễn một kiến trúc auto-encoder (tự động mã hóa) được tạo từ một bộ encoder, một khối điều chỉnh đặc trưng, và bộ giải mã.

Bộ mã hóa gồm 3 lớp tích chập có bước nhảy (lớp đầu tiên là stride 1, 2 lớp sau là stride 2) và 4 residual block .0

Strided convolutional layers (tầng tích chập có bước nhảy) là một phần quan trọng trong mạng nơ-ron tích chập (CNN) được sử dụng rộng rãi trong lĩnh vực xử lý ảnh và thị giác máy tính. Các tầng tích chập được thiết kế để học các đặc trưng cục bộ từ dữ liệu đầu vào.

**

Kernel trong mạng CNN (Convolutional Neural Network) là một ma trận có kích thước nhỏ được sử dụng để thực hiện phép tích chập trên đầu vào. Kernel, còn được gọi là ma trận hạt nhân, đóng vai trò quan trọng trong việc trích xuất đặc trưng từ dữ liệu đầu vào.

Khi áp dụng phép tích chập giữa kernel và đầu vào, kernel sẽ trượt qua các vùng cục bộ của đầu vào, nhân các phần tử tương ứng và tính tổng để tạo ra giá trị đầu ra tương ứng. Việc áp dụng phép tích chập này giúp mô hình học được các đặc trưng cục bộ từ dữ liệu, bao gồm cạnh, góc, màu sắc, texture và các thông tin quan trọng khác.

Kernel thường có kích thước nhỏ hơn so với kích thước đầu vào và được thiết kế để tìm ra các mẫu cụ thể trong dữ liệu. Ví dụ, trong xử lý ảnh, kernel có thể được thiết kế để phát hiện cạnh, đường cong, điểm nổi bật, hay các đặc trưng hình học khác.

Mạng CNN sử dụng nhiều kernel khác nhau để học các đặc trưng ở nhiều mức độ và tạo ra các bản đồ đặc trưng khác nhau. Các kernel này được học trong quá trình huấn luyện mạng CNN để tìm ra các đặc trưng quan trọng cho nhiệm vụ của mô hình, ví dụ như phân lớp ảnh, nhận dạng vật thể, hay xử lý ảnh y tế.

Tổ hợp các kernel và các tầng tích chập khác trong mạng CNN tạo ra sự hiệu quả trong việc trích xuất và học các đặc trưng từ dữ liệu đầu vào, làm cho mạng có khả năng nhìn thấy và hiểu dữ liệu ở cấu trúc và mức độ cao hơn.

**

Trong một tầng tích chập thông thường, một ma trận hạt nhân (kernel) được trượt qua các vùng cục bộ trên đầu vào để tính toán đầu ra. Strided convolutional layers mở rộng ý tưởng này bằng cách thêm một bước nhảy (stride) trong quá trình trượt của kernel. Thay vì di chuyển một bước duy nhất, kernel sẽ di chuyển với một bước nhảy lớn hơn.

Bằng cách sử dụng stride lớn hơn 1, strided convolutional layers có thể giảm kích thước không gian của đầu ra so với đầu vào. Điều này có thể hữu ích để giảm số lượng tham số và tính toán trong mạng nơ-ron tích chập, đồng thời tạo ra các bản đồ đặc trưng cụ thể hơn với thông tin tóm tắt.

Strided convolutional layers có thể được sử dụng để thay thế các lớp pooling truyền thống trong một mạng nơ-ron tích chập. Thay vì sử dụng pooling để giảm kích thước, strided convolutional layers có thể thực hiện cả việc giảm kích thước và tính toán đặc trưng.

Việc sử dụng strided convolutional layers trong mạng nơ-ron tích chập có thể cải thiện hiệu suất và hiệu quả tính toán, đồng thời duy trì thông tin quan trọng về đặc trưng trong quá trình học.

Stride 1 và stride 2 trong strided convolutional layers có ý nghĩa khác nhau.

Stride 1: Khi sử dụng stride 1 trong strided convolutional layers, kernel sẽ di chuyển một bước duy nhất trong quá trình tính toán đầu ra. Điều này có nghĩa là mỗi vùng cục bộ của đầu vào sẽ được xử lý và tính toán đầu ra tương ứng. Stride 1 giúp duy trì kích thước không gian của đầu vào và đầu ra, giữ nguyên thông tin chi tiết và tạo ra các bản đồ đặc trưng chi tiết.

Stride 2: Khi sử dụng stride 2 trong strided convolutional layers, kernel sẽ di chuyển hai bước trong quá trình tính toán đầu ra. Điều này có nghĩa là các vùng cục bộ trên đầu vào sẽ được xử lý không liên tục và có khoảng cách lớn hơn. Stride 2 giúp giảm kích thước không gian của đầu ra so với đầu vào, giảm độ phân giải và số lượng tính toán. Khi kích thước đầu vào được giảm, thông tin không gian được tóm tắt và các bản đồ đặc trưng trở nên toàn cục hơn.

Sự lựa chọn stride 1 hoặc stride 2 phụ thuộc vào mục tiêu của mô hình và đặc điểm của dữ liệu. Stride 1 thường được sử dụng để duy trì thông tin chi tiết và phù hợp với việc xử lý ảnh có kích thước nhỏ hoặc cần giữ lại nhiều chi tiết. Trong khi đó, stride 2 thường được sử dụng để giảm kích thước không gian và tăng tốc quá trình tính toán, phù hợp với việc xử lý ảnh có kích thước lớn hơn hoặc khi thông tin không gian ít quan trọng hơn so với thông tin tổng quát.

**

Residual block (khối dư thừa) là một thành phần cơ bản trong kiến trúc mạng nơ-ron sâu gọi là ResNet (Residual Network). Kiến trúc ResNet được đề xuất nhằm khắc phục hiện tượng sao mạng nơ-ron sâu càng sâu, hiệu suất càng giảm đi. Residual block đóng vai trò quan trọng trong việc xây dựng kiến trúc này.

Trong một mạng nơ-ron thông thường, đầu ra của một tầng được tính bằng cách áp dụng một phép biến đổi phi tuyến (như hàm kích hoạt ReLU) vào đầu vào. Trong Residual block, thay vì tính toán trực tiếp đầu ra, chúng ta tính toán sự khác biệt (độ lệch) giữa đầu ra dự đoán và đầu vào gốc. Công thức toán học của Residual block có thể được biểu diễn như sau:

$$output = input + F(input)$$

Trong đó, input là đầu vào ban đầu, F(input) là phép biến đổi được áp dụng lên input để tạo ra đầu ra dự đoán. Công thức trên có ý nghĩa là đầu ra cuối cùng của Residual block là sự cộng dồn giữa đầu vào và độ lệch, cho phép mạng học được những thay đổi nhỏ dựa trên đầu vào ban đầu.

Ý tưởng của Residual block là cho phép mạng học các đặc trưng thừa kế từ đầu vào ban đầu thông qua độ lệch. Nếu phép biến đổi F(input) xấp xỉ bằng 0, thì đầu ra sẽ gần như bằng input, và mạng có khả năng học các đặc trưng nhỏ và tinh vi từ dữ liệu đầu vào.

Sự sử dụng của Residual block trong kiến trúc ResNet đã giúp giải quyết vấn đề gradient bị mất mát khi mạng nơ-ron sâu. Nó cũng cho phép xây dựng các mô hình mạnh mẽ và dễ huấn luyện hơn bằng cách cho phép lan truyền ngược dễ dàng và xây dựng các mô hình có độ sâu lớn mà vẫn duy trì hiệu suất cao.

- Bộ giải mã bao gồm 2 lớp nâng cấp lân cận và 3 lớp tích chập.

**

Nearest-neighbor upsampling layers (tầng nâng cấp lân cận) là một phương pháp trong xử lý ảnh để tăng kích thước của hình ảnh. Đây là một phương pháp nâng cấp đơn giản, không yêu cầu thêm thông tin từ các tầng trước đó và không tính toán phức tạp.

Khi sử dụng nearest-neighbor upsampling, mỗi điểm ảnh trong hình ảnh gốc sẽ được nhân bản thành một khối (block) lớn hơn trong hình ảnh đầu ra. Cách làm này đơn giản là sao chép giá trị của điểm ảnh gốc và điền vào các vị trí mới.

Để minh hoạ cách nearest-neighbor upsampling hoạt động, giả sử chúng ta có một hình ảnh gốc có kích thước 2x2 như sau:

[A B]

[C D]

Nếu chúng ta muốn tăng kích thước của hình ảnh lên 4x4 bằng nearest-neighbor upsampling, kết quả sẽ là:

[A A B B]

[A A B B]

[C C D D]

[C C D D]

Như vậy, giá trị của mỗi điểm ảnh trong hình ảnh gốc được sao chép và điền vào các vị trí mới để tạo ra hình ảnh đầu ra.

Nearest-neighbor upsampling là một phương pháp nâng cấp đơn giản và nhanh chóng, nhưng nó có một số hạn chế. Nó không khai thác thông tin từ các điểm ảnh xung quanh để tạo ra các giá trị ảnh mới, do đó có thể dẫn đến hiện tượng pixelated hoặc mờ nếu tỷ lệ nâng cấp lớn. Ngoài ra, nó không giải quyết được các vấn đề như mất mát chi tiết hoặc làm mịn hình ảnh trong quá trình nâng cấp. Do đó, trong các tác vụ yêu cầu chất lượng cao và khả năng tái tạo chi tiết, các phương pháp nâng cấp khác phức tạp hơn như bilinear hoặc cubic interpolation thường được sử dụng.

Sự khác biệt so với những công trình trước nằm ở những khối điều chỉnh đặc trưng. đặc trưng được xuất ra được điều chỉnh bởi vector xác định tuổi(chi tiết bên dưới). được dùng bởi việc chuyển

đổi style, cái mà cho thấy khả năng đại diện các trạng thái khác nhau [5,10] bằng cách sử dụng các tham số của các lớp chuẩn hóa.

- **ENCODER**: ảnh khuôn mặt x0 là đầu vào của lớp mã hóa, đặc trưng output được kí hiệu bởi $C \in R^{n \times c}$, c= 128 là số kênh và n là tích của 2 chiều không gian của vector ảnh
- **FEATURE MODULATION FOR AGE SELECTION**: (điều chỉnh đặc trưng cho việc chọn tuổi): mục tiêu tuổi a1 được mã hóa bằng one-hot encoder,kí hiệu bằng z1, qua một mạng hiệu chỉnh. mạng này gồm một lớp mạng neural với hàm kích hoạt là sigmoid. Đầu ra là một vector $w \in [0, 1]^c$, được dùng để tìm w của C trước khi đưa vào lớp giải mã và có được hình ảnh khuôn mặt tương ứng với độ tuổi a1. Những tính năng được điều chỉnh là C nhân ma trận đường chéo của w
- **DECODER:** C diag(w) là đầu vào, bỏ qua hai kết nối,sử dụng để bảo đảm những đặc trưng tốt của ảnh đầu vào. output cuối cùng là G(x0, α1).

3.3. Huấn luyện

Như đã minh họa trong hình hai, ta huấn luyện mô hình chuyển đổi tuổi bằng một bộ phân lớp tuổi nhằm đảm bảo độ chính xác khi chuyển đổi và một bộ phân biệt (discriminator) để duy trì tính chân thực của bức ảnh.

Tuổi a0 của ảnh ban đầu x0 khá dễ để ước lượng dựa vào một mô hình phân lớp tuổi trước đó [31]. Vì vậy ta không dùng dùng tập dữ liệu được chú thích theo độ tuổi để huấn luyện. Độ tuổi gốc của tập dữ liệu được kí hiệu bởi $Q \subset N$. Tại thời điểm kiểm tra, độ tuổi mục tiêu có thể được chọn là bất kỳ độ tuổi nào trong Q. Tại thời điểm huấn luyện, có vẻ hợp lý khi ta chọn bất kỳ giá trị nào trong Q một cách ngẫu nhiên. Tuy nhiên, ta nhận thấy rằng các chi tiết xuất hiện trong quá trình biến đổi tuổi lớn đã thay đổi tốt hơn khi chọn tuổi mục tiêu α1 đủ xa so với α0 trong quá trình đào tạo. Ta đề xuất lấy mẫu α1 từ tập $Q_{\alpha 0} = \{\alpha \in Q : |\alpha - \alpha 0| \geq \alpha *\}$ tại thời điểm huấn luyện, trong đó $\alpha *$ là hằng số được xác định trước biểu thị khoảng biến đổi tuổi tối thiểu. Ta biểu diễn bằng $q(\alpha|\alpha 0)$ phân phối đều trên $Q_{\alpha 0}$.

Classification loss: Để đo tuổi của G(x0, α1), chúng ta sử dụng cùng bộ phân lớp tuổi với bộ phân lớp được sử dụng để ước tính α0. Trong quá trình huấn luyện, chúng ta không cập nhật trọng số của bộ phân lớp này. Bộ phân lớp, ký hiệu là V , lấy G(x0, α1) làm đầu vào và tạo ra một

phân phối xác suất rời rạc trên tập hợp các độ tuổi {0, 1, . . . , 100}. Hàm mất mát của bộ phân lớp như sau:

$$Lclass = Ex0\sim p(x)\ E\alpha 1\sim q(\alpha|\alpha 0)\ [l(z1, V\ (G(x0, \alpha 1)))]$$

Trong đó:

- *Ex0∼p(x) là kỳ vọng của đầu vào x0 theo phân phối p(x), biểu thị việc lấy mẫu ngẫu nhiên x0 từ phân phối này.*
- *Eα1∼q(α|α0) là kỳ vọng của biến ngẫu nhiên α1 theo phân phối q(α|α0), biểu thị việc lấy mẫu ngẫu nhiên α1 từ phân phối này dựa trên giá trị α0.*
- *l(z1, V(G(x0, α1))) là hàm mất mát categorical cross-entropy giữa vector one-hot z1 và phân bố dự đoán của bộ phân loại V cho đầu vào G(x0, α1).*
- *Hàm mất mát này đo lường sự khác biệt giữa phân bố đúng và phân bố dự đoán của bộ phân loại tuổi khi đánh giá tuổi của ảnh được tạo ra bởi mô hình G. Mục tiêu của quá trình huấn luyện là tối thiểu hóa hàm mất mát này để mgiảm thiểu sai khác giữa tuổi dự đoán và tuổi thực tế của ảnh.*

Adversarial loss: Adversarial loss (mất mát đối nghich) là một loại hàm mất mát được sử dụng trong mạng sinh (generative network) và mạng phân biệt (discriminative network) trong mô hình GAN (Generative Adversarial Network). Mục tiêu của adversarial loss là đạt được sự cân bằng giữa mạng sinh và mạng phân biệt thông qua quá trình đấu tranh.

Trong mô hình GAN, mạng sinh cố gắng tạo ra các mẫu giống như dữ liệu thật, trong khi mạng phân biệt cố gắng phân biệt giữa các mẫu thật và các mẫu được tạo ra bởi mạng sinh. Adversarial loss được sử dụng để huấn luyện mạng sinh bằng cách khuyến khích nó tạo ra các mẫu mà mạng phân biệt không thể phân biệt được với các mẫu thật.

Để tăng cường tính tính chân thật của hình ảnh được biến đổi G(x0, α1), ta xây dựng một hàm mất mát đối nghịch sử dụng PatchGAN [13] với mục tiêu LSGAN objective [22]. Khác với những công trình trước đó, mạng phân loại này được dùng để phân biệt ảnh thật và ảnh được tạo mà không tính đến tác động độ tuổi. Trong bài này, việc thay đổi tuổi từ bức ảnh gốc được xét riêng biệt với hàm mất mát phân loại độ tuổi.

Mạng phân loại được ký hiệu là D, kiến trúc giống như tài liệu ở mục 13. Sử dụng patch size 142×142 cho ảnh 1024×1024. Ảnh được tạo G(x0, α1) nên khác biệt với mẫu thật. Vì vậy ta dùng hàm mất mát sau:

$$L_{GAN}(G) = E_{x0\sim p(x)}E_{\alpha1\sim q(\alpha|\alpha0)} [(D(G(x0, \alpha1)) - 1)2],$$

khi huấn luyện G, trong công thức này:

- $E_{x0\sim p(x)}$ là kỳ vọng của đầu vào x0 theo phân phối p(x), biểu thị việc lấy mẫu ngẫu nhiên x0 từ phân phối này.
- $E_{\alpha1\sim q(\alpha|\alpha0)}$ là kỳ vọng của biến ngẫu nhiên α1 theo phân phối q(α|α0), biểu thị việc lấy mẫu ngẫu nhiên α1 từ phân phối này dựa trên giá trị α0.
- D(G(x0, α1)) là đầu ra của mạng phân biệt D khi nhận đầu vào là G(x0, α1).
- $(D(G(x0, \alpha1)) - 1)^2$ là hiệu số bình phương giữa đầu ra của mạng phân biệt và giá trị 1. Mục tiêu là làm cho hiệu số này càng gần 0, tức là đầu ra của mạng phân biệt gần với giá trị 1.

và

$$L_{GAN}(D) = E_{x0\sim p(x)} E_{\alpha1\sim q(\alpha|\alpha0)} [(D(G(x0, \alpha1)))2] + E_{y\sim p(x)} [(D(y) - 1)2]$$

khi huấn luyện D, trong đó:

- $E_{x0\sim p(x)}$ là kỳ vọng của đầu vào x0 theo phân phối p(x), biểu thị việc lấy mẫu ngẫu nhiên x0 từ phân phối này.
- $E_{\alpha1\sim q(\alpha|\alpha0)}$ là kỳ vọng của biến ngẫu nhiên α1 theo phân phối q(α|α0), biểu thị việc lấy mẫu ngẫu nhiên α1 từ phân phối này dựa trên giá trị α0.
- D(G(x0, α1)) là đầu ra của mạng phân biệt D khi nhận đầu vào là G(x0, α1), tức là đầu ra của mạng phân biệt khi phân biệt ảnh được chỉnh sửa.
- D(y) là đầu ra của mạng phân biệt D khi nhận đầu vào là ảnh thật y.
- $(D(G(x0, \alpha1)))^2$ là hiệu số bình phương giữa đầu ra của mạng phân biệt khi phân biệt ảnh được chỉnh sửa và giá trị 0. Mục tiêu là làm cho hiệu số này càng gần 0, tức là mạng phân biệt không thể phân biệt được giữa ảnh được chỉnh sửa và ảnh thật.
- $(D(y) - 1)^2$ là hiệu số bình phương giữa đầu ra của mạng phân biệt khi phân biệt ảnh thật và giá trị 1. Mục tiêu là làm cho hiệu số này càng gần 0, tức là mạng phân biệt phân biệt chính xác ảnh thật.

LGAN(D) cung cấp hàm mất mát để huấn luyện mạng phân biệt D sao cho nó có khả năng phân biệt chính xác giữa ảnh được chỉnh sửa và ảnh thật.

Reconstruction loss: Khi mô hình chuyển đổi tuổi nhận x0 và a0 là dữ liệu đầu vào. ảnh đầu ra được tạo từ G(x0, α0) phải tương đồng với ảnh đầu vào. Vì vậy chúng ta nên cực tiểu hóa hàm mất mát của nó.

$$Lrecon = Ex0\sim p(x) [||G(x0, \alpha0) - x0||1]$$

Trong đó:

- Ex0~p(x) là kỳ vọng của đầu vào x0 theo phân phối p(x), biểu thị việc lấy mẫu ngẫu nhiên x0 từ phân phối này.
- G(x0, α0) là ảnh được tạo ra bởi mô hình khi nhận đầu vào x0 và α0.
- x0 là ảnh đầu vào.
- Norm L1 của hiệu giữa ảnh được tạo ra và ảnh đầu vào (G(x0, α0) - x0) đo lường tổng các chênh lệch tuyệt đối giữa các điểm ảnh tại cùng vị trí trên ảnh được tạo ra và ảnh đầu vào. Bằng cách tối thiểu hóa khoảng cách L1 này, mô hình được đánh giá và huấn luyện để tạo ra ảnh có độ giống nhau cao nhất với ảnh đầu vào.

Do đó, Lrecon đảm bảo rằng mô hình có khả năng tái tạo lại ảnh đầu vào một cách chính xác, đồng thời cung cấp một mất mát để huấn luyện mô hình sao cho nó tạo ra các biến thể tuổi thích hợp mà vẫn giữ được đặc trưng quan trọng của ảnh gốc.

Full loss: Ta huấn luyện mô hình chuyển đổi tuổi và phân biệt bằng cách tối thiểu hóa đối tượng tổng thể hàm mất mát. gồm ba thành phần chính: reconstruction loss (Lrecon), classification loss (Lclass) và adversarial loss (LGAN). Công thức mất mát được định nghĩa như sau::

$$L = \lambda reconLrecon + \lambda classLclass + LGAN$$

Trong đó:

- λrecon và λclass là các trọng số để cân bằng ảnh hưởng của từng thành phần mất mát. Chúng xác định mức độ quan trọng của mỗi thành phần trong quá trình huấn luyện. Bằng

cách điều chỉnh các giá trị của λrecon và λclass, ta có thể ưu tiên một thành phần mất mát so với các thành phần khác.

- Lrecon là reconstruction loss, như đã giải thích trước đó, đo lường sự khác biệt giữa ảnh được tạo ra và ảnh đầu vào, với mục tiêu tạo ra ảnh giống hệt ảnh gốc.
- Lclass là classification loss, như đã giải thích trước đó, đo lường độ chính xác của mô hình phân loại tuổi khi đánh giá đầu ra của age transformer.
- LGAN là adversarial loss, như đã giải thích trước đó, đảm bảo tính thực tế cao hơn của ảnh được chỉnh sửa bằng cách đánh giá sự khác biệt giữa đầu ra của mạng phân biệt và giá trị mục tiêu.

Bằng cách kết hợp các thành phần mất mát này với các trọng số tương ứng, công thức mất mát đầy đủ giúp huấn luyện mô hình age transformer và mạng phân biệt sao cho đạt được sự cân bằng giữa việc tái tạo ảnh gốc và tạo ra các biến thể tuổi phù hợp, đồng thời đảm bảo tính thực tế cao của ảnh được chỉnh sửa.

4. Các thử nghiệm

Trong báo cáo này, ta đã đề xuất một kiến trúc age transformer, cho phép chỉnh sửa tuổi của khuôn mặt một cách liên tục với một mạng duy nhất, cố gắng giữ cho mô hình đơn giản nhất có thể. Với phương pháp này, kết hợp với một kiến trúc mã hóa-giải mã, thay vì dựa vào một mạng GAN phức tạp, là con đường tốt nhất để đạt được kết quả chỉnh sửa khuôn mặt với chất lượng và độ phân giải cao. Ta đã chứng minh được khả năng của mô hình khi tạo ra kết quả chân thực và sắc nét, mà không gây ra các hiện tượng nhiễu, trên ảnh có độ phân giải 1024 × 1024. Khối feature modulation làm việc hiệu quả trong phân tách thông tin về tuổi và thông tin về định danh. Với hiệu suất đạt được, thiết kế này có thể hữu ích cho các bài tập hay yêu cầu thay đổi thuộc tính khuôn mặt khác.

4.1. Tăng cường dữ liệu bằng hình ảnh tổng hợp

Bộ dữ liệu huấn luyện của chúng ta được xây dựng dựa trên bộ dữ liệu FFHQ [16], một bộ dữ liệu độ phân giải cao chứa 70.000 hình ảnh khuôn mặt với độ phân giải 1024x1024. Bộ dữ liệu này bao gồm sự biến đổi lớn về tuổi tác, dân tộc, tư thế, ánh sáng và nền ảnh. Tuy nhiên, bộ dữ liệu chỉ chứa các hình ảnh gốc không có nhãn thu thập từ Flickr.

Để thu thập thông tin về tuổi tác, chúng ta sử dụng một bộ phân loại tuổi được tiền huấn luyện trên bộ dữ liệu IMDB-WIKI [31]. Chúng ta quan sát thấy rằng FFHQ chứa nhiều hình ảnh khuôn mặt trẻ hơn là các khuôn mặt già. Sự mất cân bằng dữ liệu này là một thách thức vì các nhiệm vụ về lão hóa và trẻ hóa sẽ không được đối xử một cách công bằng trong quá trình huấn luyện: vì phần lớn các khuôn mặt đều là trẻ, mô hình biến đổi tuổi sẽ được huấn luyện để thực hiện lão hóa nhiều hơn là trẻ hóa, dẫn đến kết quả trẻ hóa không thỏa mãn. Để cân bằng mất cân bằng này trong phân phối tuổi, chúng ta đề xuất thực hiện tăng cường dữ liệu bằng cách sử dụng StyleGAN - một mô hình tạo ảnh độ phân giải cao hàng đầu [16]. Chúng ta sử dụng mô hình StyleGAN đã được tiền huấn luyện trên FFHQ để tạo ra 300.000 hình ảnh tổng hợp. Kiểm tra hình ảnh một cách nhanh chóng cho thấy hầu hết các hình ảnh được tạo ra không có lỗi nổi bật và gần như không thể phân biệt với hình ảnh thực bằng mắt thường. Do đó, chúng ta sử dụng chúng để tăng cường dữ liệu để có được một phân phối tuổi gần như đồng đều: đối với mọi khoảng tuổi có ít hơn 1.000 mẫu trong bộ dữ liệu gốc FFHQ, chúng ta hoàn thiện khoảng tuổi này bằng một số hình ảnh khuôn mặt tổng hợp được tạo ra; đối với mọi khoảng tuổi có nhiều hơn 1.000 mẫu, chúng ta ngẫu nhiên chọn 1.000 hình ảnh khuôn mặt từ bộ dữ liệu gốc FFHQ. Bộ dữ liệu được cân bằng về tuổi chứa 47.990 hình ảnh trong khoảng Q = {20, ..., 69}.

4.2. Chi tiết thực hiện

Mô hình của chúng ta được triển khai bằng PyTorch [28]. Chúng ta lấy 95% của bộ dữ liệu được cân bằng làm tập huấn luyện và phần còn lại làm tập kiểm tra. Đối với mô hình biến đổi tuổi và bộ phân biệt, chúng ta áp dụng phổ chứng thực (spectral normalisation) [25] trên tất cả các lớp tích chập, trừ lớp cuối cùng của mô hình biến đổi tuổi. Tất cả các lớp kích hoạt sử dụng Leaky ReLU [21] với độ dốc âm là 0.2.

Chúng ta chỉ xem xét biến đổi tuổi trong khoảng Q = {20, ..., 69}. Hằng số α^* được đặt là 25. Chúng ta đã quan sát thấy rằng các lỗi nổi bật nhất xuất hiện khi khoảng cách giữa tuổi nguồn và tuổi đích là lớn. Bằng cách chọn α^* đủ lớn, chúng ta buộc bộ phân biệt D kiềm chế những lỗi nổi bật này trong quá trình huấn luyện đối kháng. Các trọng số λ_{recon} và λ_{class} được đặt là 10 và 0.1, tương ứng. Chúng ta sử dụng bộ tối ưu hóa Adam với tỷ lệ học là 10^-4. Mô hình biến đổi tuổi G được cập nhật một lần sau mỗi lần cập nhật bộ phân biệt. Mô hình của chúng ta được huấn luyện trong 20 epochs để đạt được việc chỉnh sửa tuổi khuôn mặt trên hình ảnh có độ phân giải cao. 10 epochs đầu tiên được huấn luyện trên hình ảnh có kích thước 512 × 512 với kích thước lô

là 4. 10 epochs tiếp theo được huấn luyện trên hình ảnh có kích thước 1024 × 1024, trong đó chúng ta giảm kích thước lô xuống còn 2, tỷ lệ học xuống 10^-5 và λrecon xuống 1.

4.3. Đánh giá chất lượng

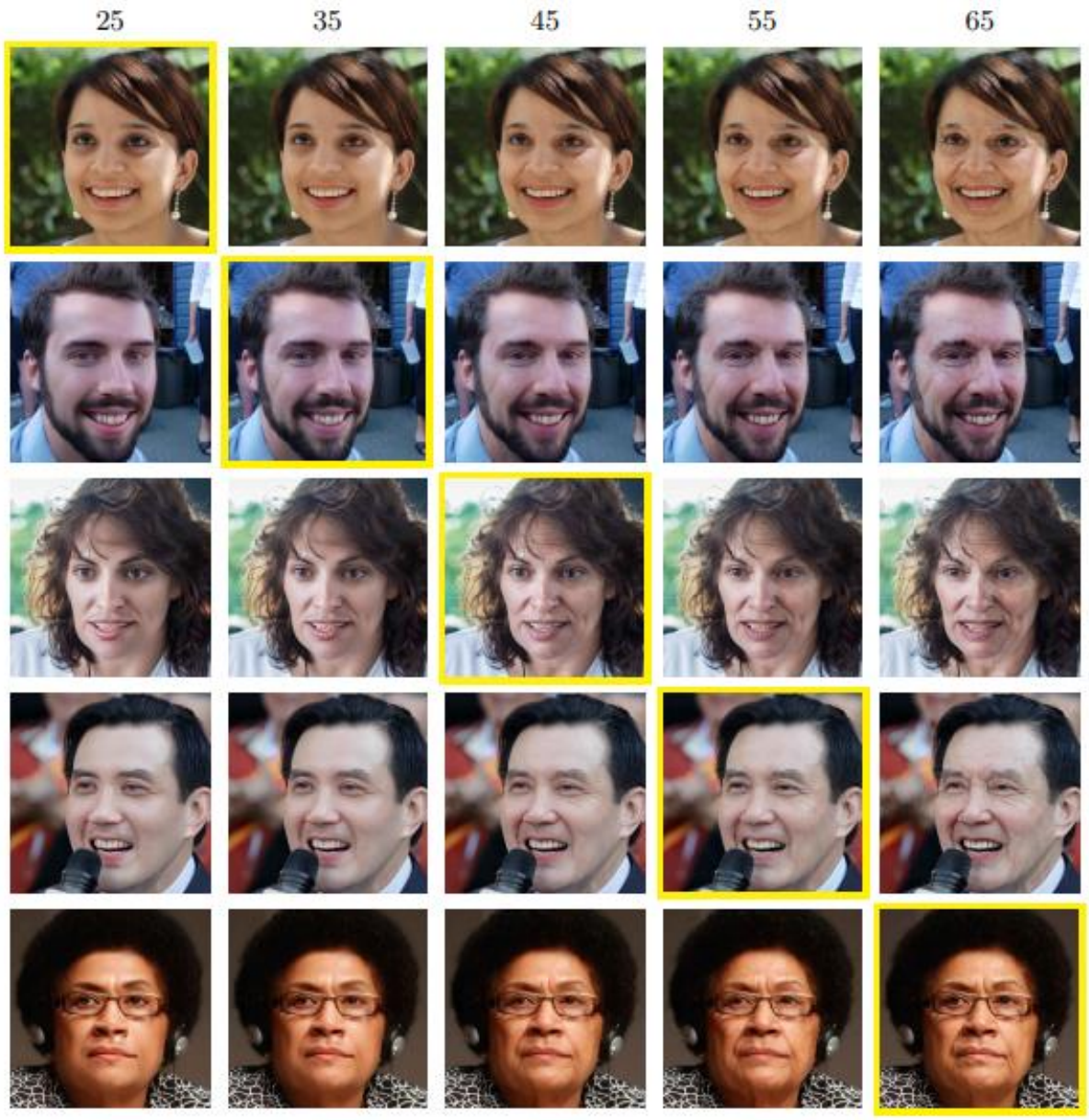
Hình ảnh 9 trình bày các kết quả chỉnh sửa tuổi trên các hình ảnh đầu vào 1024 × 1024 thuộc các nhóm tuổi khác nhau. Phương pháp của chúng ta cho kết quả hài lòng mắt với các chi tiết sắc nét (tốt nhất khi phóng to các kết quả) và không gây ra lỗi nổi bật đáng kể. Chỉ có các đặc điểm liên quan đến tuổi được sửa đổi, trong khi nhận dạng, kiểu tóc, cảm xúc và nền được bảo toàn tốt. Điều này càng tốt hơn khi không sử dụng bất kỳ mặt nạ nào để cô lập khuôn mặt khỏi phần còn lại của hình ảnh. Hình ảnh 4 trình bày các kết quả chỉnh sửa tuổi với một quá trình tuổi đích mượt mà. Sự khác biệt giữa hai kết quả liền kề gần như không nhìn thấy được, điều này minh họa sự mượt mà của quá trình làm già.

Chúng ta so sánh phương pháp của mình với hai phương pháp tiên tiến nhất hiện nay về chỉnh sửa tuổi khuôn mặt, trong đó các mã nguồn chính thức đã được công bố - IPCGAN [36] và PAGGAN [38]. Chúng ta cũng so sánh kết quả của chúng ta với kết quả thu được bằng FaderNet [18], cho phép điều chỉnh nhiều thuộc tính khuôn mặt bao gồm cả tuổi.

Hình ảnh 5 trình bày các kết quả làm già của IPCGAN, PAGGAN và phương pháp của chúng ta trên bộ dữ liệu CACD [2]. Kích thước đầu ra của mỗi phương pháp là: 128 × 128 cho IPCGAN, 224 × 224 cho PAGGAN, 256 × 256 cho phương pháp của chúng ta. Chúng ta chỉ so sánh các phương pháp trên một nhóm tuổi cụ thể để thể hiện tác động của tuổi lên kết quả chỉnh sửa tuổi. Các kết quả cho thấy phương pháp của chúng ta đạt được các kết quả tốt nhất về cả mặt mô phỏng và đồng thời giữ được nhiều chi tiết cần thiết của khuôn mặt ban đầu.

Khả năng tổng quát hóa cho hình ảnh trong tập dữ liệu chưa được nhìn thấy trước đó: Để có sự so sánh công bằng và giảm thiểu tác động của việc overfitting trên dữ liệu huấn luyện, chúng ta đánh giá tất cả các phương pháp trên một tập dữ liệu không được xem trong quá trình huấn luyện bởi bất kỳ phương pháp nào. Chúng ta đã chọn tập dữ liệu CelebA-HQ [15], một phiên bản có độ phân giải cao của bộ dữ liệu CelebA. Các hình ảnh đầu vào có độ phân giải 1024 × 1024, và được giảm mẫu xuống độ phân giải mà mỗi phương pháp được huấn luyện bằng cách sử dụng các mã nguồn chính thức của chúng. Kích thước đầu ra của mỗi phương pháp là: 224 × 224 cho PAGGAN, 128 × 128 cho IPCGAN, 256 × 256 cho FaderNet, và 1024 × 1024 cho phương pháp

của chúng ta. Chúng ta chỉ so sánh kết quả chỉnh sửa tuổi từ nhóm tuổi trẻ đến nhóm tuổi cao, vì PAGGAN và IPCGAN chỉ được huấn luyện cho việc làm già. Hình ảnh 6 thể hiện các kết quả thu được bằng các phương pháp khác nhau. FaderNet [18] chỉ tạo ra các sửa đổi nhỏ. PAGGAN [38] tạo ra hiệu ứng tiến hóa tuổi thú vị, tuy nhiên, lỗi rõ rệt xuất hiện trên cạnh mặt và tóc. IPCGAN [36] bị giới hạn ở độ phân giải thấp và gây giảm chất lượng hình ảnh mạnh.



Hình 3. Kết quả chỉnh sửa tuổi trên hình ảnh 1024 × 1024 trên FFHQ [16]. Trên mỗi hàng, khung màu vàng chỉ ra hình ảnh gốc. Mỗi cột tương ứng với một độ tuổi mục tiêu: 25, 35, 45, 55, 65. Phương pháp của chúng ta mang lại kết quả hài lòng về mặt hình ảnh mà không gây ra lỗi đáng kể. Chỉ các đặc điểm liên quan đến tuổi được chỉnh sửa, trong khi danh tính, kiểu tóc, cảm xúc và nền ảnh được bảo toàn hoàn hảo.

Hình 4 Kết quả liên tục của việc chỉnh sửa tuổi trên khuôn mặt trên FFHQ [16]. Như có thể quan sát, sự khác biệt giữa hai kết quả liền kề gần như không thể nhìn thấy, điều này cho thấy tính mượt mà của quá trình làm già đi.

Bảng 1. ***Đánh giá định lượng bằng cách sử dụng API nhận dạng khuôn mặt trực tuyến*** *[12]. Chúng ta so sánh phương pháp của chúng ta với ba phương pháp khác: Fader Network [18], PAGGAN [38] và IPCGAN [36]. Các hình ảnh được chuyển đổi thành nhóm tuổi cao nhất (50+) cho tất cả các phương pháp. Cột thứ hai hiển thị tuổi được dự đoán trung bình. Cột thứ ba chỉ số độ mờ của kết quả (giá trị thấp hơn có nghĩa là ít mờ hơn). Cột thứ tư là tỷ lệ bảo toàn giới tính, có nghĩa là tỷ lệ phần trăm giới tính gốc được bảo toàn. Cột thứ năm liên quan đến việc bảo toàn biểu hiện - tỷ lệ bảo toàn nụ cười. Hai cột cuối cùng chỉ số tỷ lệ bảo toàn cảm xúc.*

Method	Predicted Age	Blur	Gender Preservation(%)	Smiling Preservation(%)	Emotion Preservation(%) Neutral	Happiness
FaderNet [18]	44.34 ± 11.40	9.15	**97.60**	95.20	90.60	92.40
PAGGAN [38]	49.07 ± 11.22	3.68	95.10	93.10	90.20	91.70
IPCGAN [36]	49.72 ± 10.95	9.73	96.70	93.60	89.50	91.10
Ours	54.77 ± 8.40	**2.15**	97.10	**96.30**	**91.30**	**92.70**

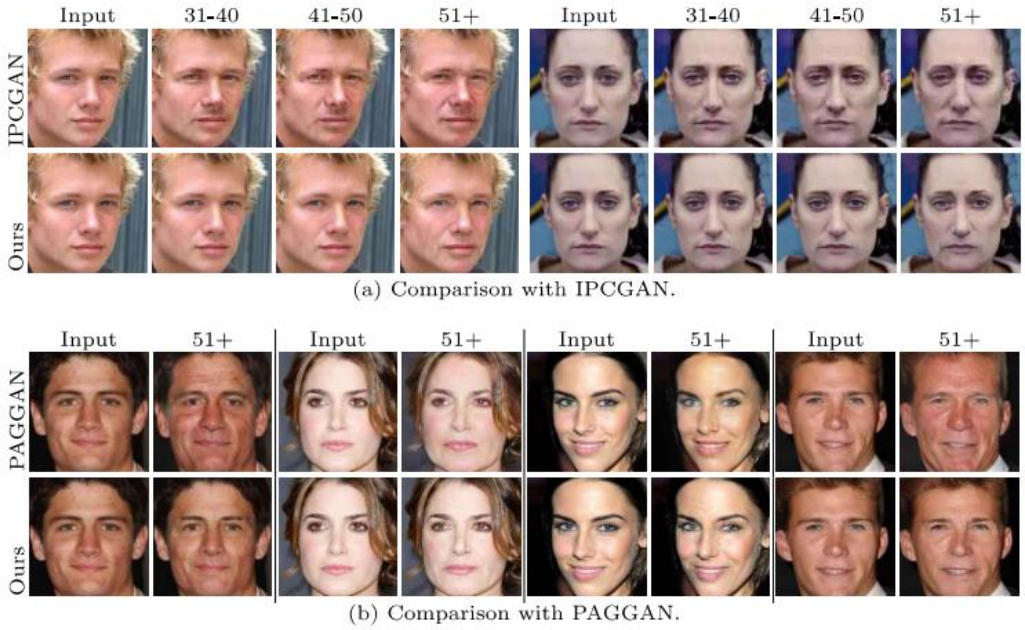
So với các kết quả này, phương pháp của chúng ta tạo ra ít lỗi hơn và bảo toàn tốt các chi tiết tinh tế của khuôn mặt và nền ảnh.

4.4. Đánh giá định lượng

Việc đánh giá định lượng của việc chuyển đổi hình ảnh thành hình ảnh mới vẫn là một vấn đề chưa được giải quyết và chưa có một phương pháp đo lường chung để đo đạc tính thực tế và định lượng các đặc tính nhân tạo được thêm vào trong một hình ảnh. Các nghiên cứu gần đây [9, 20, 38] về việc tăng độ tuổi của khuôn mặt đã sử dụng một API nhận diện khuôn mặt trực tuyến để ước tính tuổi và độ chính xác trong việc giữ đúng đặc điểm nhận dạng của các hình ảnh đã chỉnh sửa. Vì vậy, chúng ta áp dụng một quy trình đánh giá tương tự.

Trong quá trình đánh giá, chúng ta lựa chọn 1.000 hình ảnh đầu tiên được gắn nhãn "Young" trong tập dữ liệu CelebA-HQ để sử dụng làm hình ảnh kiểm tra. Bằng việc sử dụng tập kiểm tra này, chúng ta tiến hành một so sánh định lượng với FaderNet (18), IPCGAN (36) và PAGGAN (38). Mỗi hình ảnh được chuyển đổi sang nhóm tuổi cao nhất bằng cách sử dụng các mô hình đã được phát hành chính thức của từng phương pháp. Đối với IPCGAN và PAGGAN, nhóm tuổi cao nhất

tương ứng với 50+ và [51, 60]. Đối với FaderNet, thuộc tính "old" được thiết lập là giá trị lớn nhất mặc định trong quá trình tăng độ tuổi trong mã nguồn chính thức của phương pháp này. Để có một so sánh công bằng với các phương pháp nhóm, và vì 50+ được coi là nhóm tuổi cao nhất, chúng ta chọn một độ tuổi mục tiêu là 60 (trung bình của phạm vi tuổi $\{51, 69\} \subset \mathcal{Q}$) cho bộ biến đổi tuổi của chúng ta.



Hình 5. So sánh với IPCGAN [36] và PAGGAN [38] trên tập dữ liệu CACD. Trong phần (a), hàng trên hiển thị kết quả tăng độ tuổi của IPCGAN, trong khi hàng thứ hai hiển thị hình ảnh được tạo ra bởi phương pháp của chúng ta. Tương tự, trong phần (b), hàng trên hiển thị kết quả tăng độ tuổi của PAGGAN và hàng thứ hai hiển thị hình ảnh được tạo ra bởi phương pháp của chúng ta.

Do đó, chúng ta thu được 1000 hình ảnh đã chỉnh sửa cho mỗi phương pháp. Chúng ta tiếp tục đo lường các hình ảnh kết quả này bằng cách sử dụng API nhận diện khuôn mặt trực tuyến từ Face++ [12]. Từ API nhận diện, chúng ta thu được một số chỉ số thú vị bao gồm: tuổi, giới tính, độ mờ (xem mặt có mờ hay không, giá trị lớn hơn tức là mờ hơn), cười và ước tính cảm xúc. Các cảm xúc được ước tính bao gồm: buồn, bình thường, ghê tởm, giận dữ, ngạc nhiên, sợ hãi và hạnh phúc. Sau một phân tích sơ bộ của kết quả, chúng ta nhận thấy 94,20% hình ảnh đầu vào được phân loại là bình thường hoặc hạnh phúc. Vì vậy, chúng ta chỉ tập trung so sánh việc bảo toàn cảm xúc này bằng cách giữ lại hai thuộc tính này. Chúng ta cũng đã so sánh tỷ lệ bảo toàn danh tính bằng cách so sánh các hình ảnh đã chỉnh sửa với đầu vào ban đầu sử dụng API. Tuy nhiên, vì tất

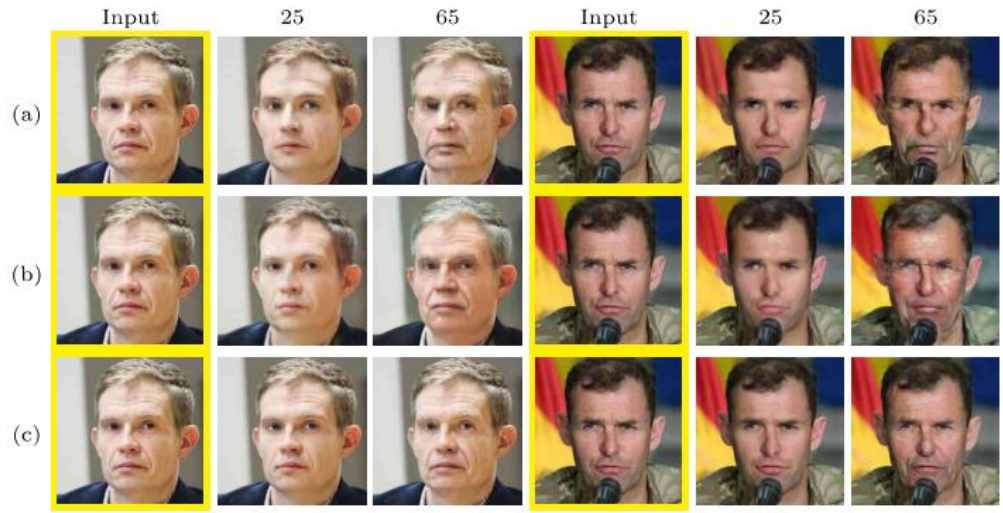
cả các phương pháp đều đạt độ chính xác gần như 100%, nên sẽ không có số liệu cụ thể được báo cáo ở đây.

Bảng 1 hiển thị các kết quả đánh giá định lượng. Tất cả các phương pháp đều được chọn với nhóm tuổi cao nhất làm mục tiêu, và chúng ta nhận thấy phương pháp của chúng ta có độ tuổi dự đoán trung bình cao nhất. Để đánh giá sự bảo toàn giới tính, chúng ta so sánh giới tính ước tính với các giới tính gốc từ bộ dữ liệu CelebA. Kết quả này cho thấy FaderNet đạt hiệu suất tốt nhất, tiếp theo là phương pháp của chúng ta. Đối với bảo toàn biểu cảm (như mỉm cười) và bảo toàn cảm xúc (như trạng thái trung lập và hạnh phúc), phương pháp của chúng ta mang lại kết quả tốt nhất. Sự thật là tất cả các phương pháp đều cho kết quả tương tự nhau. Khi đánh giá mức độ mờ, kết quả không đồng nhất hơn nhiều. Phương pháp của chúng ta thực hiện tốt hơn trong việc tạo ra hình ảnh sắc nét, nhất quán với so sánh hình ảnh trực quan.

Hình 6. So sánh kết quả lão hóa khuôn mặt trên CelebA-HQ [15]. Cột đầu tiên chứa hình ảnh đầu vào. Các cột thứ hai đến thứ năm lần lượt là kết quả của Fader Network [18], PAG-GAN [38], IPC-GAN [36] và phương pháp của chúng ta. Kết quả của chúng ta đạt được độ phân giải cao nhất mà không gây ra bất kỳ hiện tượng nhiễu đáng kể nào. Phương pháp của chúng ta bảo toàn phông nền tốt hơn so với các kỹ thuật khác. Ví dụ, nhìn vào các chữ cái, phương pháp của chúng ta không xuất hiện hiện tượng nhiễu hoặc mờ.

4.5. Thảo luận

Về nghiên cứu về phép loại bỏ trên bộ phân biệt, chúng ta đã khám phá ba loại bộ phân biệt khác nhau để huấn luyện Age Transformer. Hình 7 trình bày kết quả chỉnh sửa tuổi mặt tương ứng với các cài đặt khác nhau của bộ phân biệt.



Hình 7. Trình bày kết quả chỉnh sửa tuổi mặt sử dụng ba loại bộ phân biệt khác nhau. (a) Sử dụng bộ phân biệt có điều kiện. (b) Sử dụng hai bộ phân biệt riêng biệt, một cho nhóm tuổi cao và một cho nhóm tuổi trẻ. (c) Phương pháp đề xuất của chúng ta sử dụng một bộ phân biệt duy nhất.

- Bộ phân biệt có điều kiện. Chúng ta áp dụng một bộ phân biệt theo mảnh [13] với phép chiếu nhãn được áp dụng lên các đặc trưng trước lớp tích chập cuối cùng, tương tự như cài đặt trong [26]. Bộ phân biệt được điều kiện theo bốn nhóm tuổi: 20-35, 35-45, 45-55, 55-70. Trong quá trình huấn luyện, chúng ta nhận thấy rằng việc cung cấp số lượng ảnh thật và ảnh giả từ mỗi lớp cho bộ phân biệt có vai trò quan trọng để đạt được sự thành công. Khi chúng ta chọn một độ tuổi mục tiêu a từ tập hợp $\mathcal{Q}_{\alpha_0} = \{\alpha \in \mathcal{Q}: |\alpha - \alpha_0| \geq \alpha_*\}$ trong quá trình huấn luyện, bộ phân biệt sẽ nhận được nhiều hình ảnh đã được chỉnh sửa trong nhóm tuổi trẻ nhất và già nhất. Điều này dẫn đến xu hướng phân loại tất cả các hình ảnh trong hai nhóm này là giả. Bộ phân biệt có điều kiện rất nhạy cảm đối với phân phối dữ liệu ban đầu và yêu cầu điều chỉnh siêu tham số nhiều hơn để đạt được sự hội tụ. Hình 7(a) trình bày kết quả chỉnh sửa tuổi với bộ phân biệt có điều kiện. Các hiện tượng nhiễu mạnh có thể được quan sát trong kết quả lão hóa.

- **Hai bộ phân biệt riêng biệt**. Một bộ phân biệt nhận cả hình ảnh đã được chỉnh sửa và hình ảnh thật với độ tuổi mong muốn thuộc nhóm tuổi cao (45-70), trong khi bộ phân biệt khác nhận cả hình ảnh đã được chỉnh sửa và hình ảnh thật thuộc nhóm tuổi trẻ (20-44). Với cài đặt này, nhiệm vụ tạo ra hiệu ứng lão hóa / trẻ hóa được chia sẻ giữa bộ phân loại và các bộ phân biệt. Mặc dù kết quả trong hình 7(b) tốt hơn so với 7(a), nhưng ta có thể nhận thấy các hiện tượng làm mờ quá mức trong kết quả trẻ hóa và các hiện tượng nhiễu màu xuất hiện trong kết quả lão hóa.
- **Một bộ phân biệt duy nhất**. Đây là phương pháp mà chúng ta đề xuất. Bộ phân biệt này có thể được coi là một yếu tố điều chỉnh thông thường, đảm bảo tính chân thực của hình ảnh, vì nó nhận cả hình ảnh đã được chỉnh sửa và hình ảnh thật làm đầu vào. Việc tạo ra hiệu ứng lão hóa / trẻ hóa chỉ được điều chỉnh bởi bộ phân loại tuổi. Chúng ta chỉ có thể đạt được kết quả với độ phân giải cao khi sử dụng cài đặt cuối cùng này.A Kiến trúc mạng
- **Reconstruction từ mã tiềm ẩn được tối ưu hóa**. Như đã đề cập trong Phần 2, công trình gần đây của Shen và đồng nghiệp [32] đề xuất một cách hiệu quả để điều chỉnh mã tiềm ẩn của bộ tạo hình ảnh để đạt được sự thay đổi chất lượng hình ảnh tổng hợp cao. Do đó, việc điều chỉnh trực tiếp mã tiềm ẩn để tạo ra sự thay đổi khuôn mặt (và do đó là chỉnh sửa tuổi) trên hình ảnh tự nhiên có vẻ hấp dẫn với phương pháp này. Tuy nhiên, việc tìm mã tiềm ẩn như vậy cho một hình ảnh khuôn mặt tùy ý vẫn là một vấn đề đáng thách thức. Theo các thử nghiệm của chúng ta sử dụng StyleGAN [16], chỉ có một phần nhỏ các hình ảnh khuôn mặt tự nhiên có thể được tái tạo chính xác từ mã tiềm ẩn bởi [27]. Do đó, loại phương pháp này là không thực tế cho đến khi có sẵn một bộ mã hóa StyleGAN tốt hơn. Hình 8 được thiết kế để chứng minh điều này, trong đó ta có thể đánh giá kết quả tái tạo của các hình ảnh khuôn mặt tự nhiên. Chúng ta nhận thấy rằng các hình ảnh tái tạo có hiện tượng giống hình vẽ, nền mờ và đôi khi không thể duy trì đúng danh tính của người trong hình ảnh ban đầu. Thực tế, StyleGAN hiệu quả hơn trong việc lấy mẫu các khuôn mặt ngẫu nhiên từ không gian mã tiềm ẩn so với việc xấp xỉ hình ảnh khuôn mặt đã cho. Điều này là do một GAN không nhất thiết phải là một phép nghịch đảo. Do đó, một phương pháp chỉnh sửa dựa trên việc tái tạo mã tiềm ẩn này sẽ gặp khó khăn trong việc xử lý đúng các hình ảnh tự nhiên và đạt được chất lượng hình ảnh cao như phương pháp của chúng ta.

Hình 8. Hình ảnh tái tạo từ tối ưu mã tiềm ẩn. Chúng ta phân tích khả năng mã hóa hình ảnh tự nhiên vào không gian mã tiềm ẩn của StyleGAN [16], thông qua tối ưu trong không gian mã tiềm ẩn để giảm khoảng cách giữa hình ảnh được tạo ra và hình ảnh đầu vào. Sau đó, mỗi hình ảnh được tái tạo từ mã tiềm ẩn đã được tối ưu này. Chất lượng tương đối thấp của việc tái tạo này mạnh mẽ cho thấy việc chỉnh sửa được thực hiện trong không gian mã tiềm ẩn không thể dẫn đến kết quả sắc nét và không có hiện tượng nhiễu.

- **Huấn luyện theo hình thức giám sát yếu**. Theo hiểu biết của chúng ta, công trình của chúng ta là công trình đầu tiên sử dụng dữ liệu không có nhãn để huấn luyện trong số các nghiên cứu gần đây về lão hóa khuôn mặt [9,20,33,36,39]. Một bộ phân loại được huấn luyện trước trên IMDB-WIKI [31], một tập dữ liệu khuôn mặt độ phân giải thấp, được sử dụng để cung cấp thông tin về tuổi. Hơn nữa, bộ phân biệt trong phương pháp của chúng ta chỉ được sử dụng để phân biệt hình ảnh thật và hình ảnh đã được chỉnh sửa. Dựa chỉ duy nhất vào bộ phân loại, chúng ta thành công trong việc trích xuất các đặc trưng cụ thể về tuổi và thực hiện biến đổi tuổi trên các hình ảnh độ phân giải cao. Điều này cho thấy khả năng của bộ phân loại, ngay cả khi được huấn luyện trên các hình ảnh chất lượng thấp. Phương pháp của chúng ta có thể được áp dụng rộng rãi cho các nhiệm vụ chỉnh sửa thuộc tính khuôn mặt khác nhau bằng cách sử dụng một cặp mạng điều tiết và bộ phân loại riêng biệt cho mỗi thuộc tính.

5. Kết luận

Trong báo cáo này, ta đã đề xuất một kiến trúc age transformer, cho phép chỉnh sửa tuổi của khuôn mặt một cách liên tục với một mạng duy nhất, cố gắng giữ cho mô hình đơn giản nhất có thể. Với phương pháp này, kết hợp với một kiến trúc mã hóa-giải mã, thay vì dựa vào một mạng GAN phức tạp, là con đường tốt nhất để đạt được kết quả chỉnh sửa khuôn mặt với chất lượng và độ phân giải cao. Ta đã chứng minh được khả năng của mô hình khi tạo ra kết quả chân thực và sắc nét, mà không gây ra các hiện tượng nhiễu, trên ảnh có độ phân giải 1024 × 1024. Khối feature modulation làm việc hiệu quả trong phân tách thông tin về tuổi và thông tin về định danh. Với hiệu suất đạt được, thiết kế này có thể hữu ích cho các bài tập hay yêu cầu thay đổi thuộc tính khuôn mặt khác.

References

1. Antipov, G., Baccouche, M., Dugelay, J.L.. Face aging with conditional generative adversarial networks. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 2089-2093. IEEE (2017)

2. Chen, B.C., Chen, C.S., Hsu, W.H: Cross-age reference coding for age-invariant face recognition and retrieval. In: European conference on computer vision. pp. 768-783. Springer (2014)

3. Chen, Y.C., Shen, X., Lin, Z., Lu, X., Pao, L., Jia, J., et al: Semantic compo- nent decomposition for face attribute manipulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9859-9867 (2019)

4. Choi, Y., Choi, M., Kim, M., Ha. J.W., Kim, S., Choo, J.: Stargan: Unified gener- ative adversarial networks for multi-domain image-to-image translation. In: Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8789-8797 (2018)

5. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. Proc of ICLR (2017)

6. Fu, Y., Guo, G., Huang, T.S. Age synthesis and estimation via faces: A survey. IEEE transactions on pattern analysis and machine intelligence 32(11), 1955-1976 (2010)

7. Goodfellow, L., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672-2680 (2014)

8. He, K., Zhang, X., Ren, S., Sun, J: Deep residual learning for image recognition. In: CVPR (2016)

9. He, Z., Kan, M., Shan, S., Chen, X.: S2gan: Share aging factors across ages and share aging trends among individuals. In: Proceedings of the IEEE International Conference on Computer Vision: pp. 9440-9449 (2019)

10. Huang, X., Belongie, S.J.: Arbitrary style transfer in real-time with adaptive in-stance normalization. In: ICCV (2017)

11. Huang, X., Liu, M.Y., Belongie, S., Kautz, J: Multimodal unsupervised image- to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 172-189 (2018)

12. Inc, M.: Face+- research toolkit. http://www.faceplusplus.com. (2013)

13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with condi- tional adversarial networks. In: CVPR (2017)

14. Johnson, J. Alahi, A., Li, FF. Perceptual losses for real-time style transfer and super-resolution. In: ECCV. Springer (2016)

15. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANS for improved quality, stability, and variation. In: International Conference on Learning Representations (2018), https://openrevies.net/forum?id=Hk99zCeAb

16. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4401-4410 (2019)

17. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. arXiv preprint arXiv:1912.04958 (2019)

18. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., DENOYER, L., et al.: Fader networks: Manipulating images by sliding attributes. In: Advances in Neural In- formation Processing Systems (2017)

19. Li, P., Hu, Y., He, R., Sun, Z.: Global and local consistent wavelet-domain age synthesis IEEE Transactions on Information Forensics and Security (2019)

20. Liu, Y., Li, Q., Sun, Z.: Attribute-aware face aging with wavelet-based generative adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

21. Maas, A.L., Hannun, A.Y., Ng, A.Y: Rectifier nonlinearities improve neural net- work acoustic models. In: in ICML Workshop on Deep Learning for Audio, Speech and Language Processing. Citeseer (2013)

22. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares gen-erative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision pp. 2794 2802 (2017)

23. Mescheder. L., Nowozin, S., Geiger, A.: Which training methods for gans do actu- ally converge? In: International Conference on Machine Learning (ICML) (2018)

24. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)) 25. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y: Spectral normalization for generative adversarial networks. In: International Conference on Learning Repre-sentations (2018), https://openreview.net/forum?id=B1QRgziT-

26. Miyato, T., Koyama, M. cgans with projection discriminator arXiv preprint arXiv:1802.05637 (2018)

27. Nikitko, D.: Stylegan encoder for official tensorflow implementation https://github.com/Puzer/stylegan-encoder (2019)

28. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: NIPS Autodiff Workshop (2017) 29. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F: Gani- mation: Anatomically-aware facial animation from a single image. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 818-833 (2018)

30. Qian, S., Lin, K. Y., Wu, W., Liu, Y., Wang, Q. Shen, F., Qian, C., He, R.: Make a face: Towards arbitrary high fidelity face manipulation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10033-10042 (2019)

31. Rothe, R., Timofte, R., Van Gool, L.: Dex: Deep expectation of apparent age from a single image. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 10-15 (2015)

32. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. arXiv preprint arXiv:1907.10786 (2019)

33. Song, J., Zhang, J., Gao, L., Liu, X., Shen, H.T.; Dual conditional gans for face aging and rejuvenation. In: IJCAI. pp. 899-905 (2018)

34. Upchurch, P., Gardner, J., Pleiss, G., Pless, R., Snavely, N., Bala, K., Weinberger, K.: Deep feature interpolation for image content changes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7064-7073 (2017)

35. Wang, T.C., Lin, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In:CVPR (2018)

36. Wang, Z., Tang, X., Luo, W., Gao, S.: Face aging with identity-preserved condi- tional generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7939-7947 (2018)

37. Xiao, T., Hong, J., Ma, J. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 168-184 (2018)

38. Yang, H., Huang, D., Wang, Y., Jain, A.K.: Learning face age progression: A pyramid architecture of gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 31-39 (2018)

39. Zhang, Z. Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

40. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE interna- tional conference on computer vision, pp. 2223-2232 (2017)

A. Kiến trúc mạng

Bảng 2 trình bày các siêu tham số của kiến trúc mạng được đề xuất. Mạng phân biệt là với kích thước patch là 142 × 142. Mỗi phần tử của feature map đầu ra tương ứng với một vùng nhận thức có kích thước 142 × 142 trên ảnh đầu vào gốc.

B. Phân loại tuổi

Để thu thập thông tin về tuổi của bộ dữ liệu FFHQ [16], ta sử dụng bộ phân loại tuổi [31] đã được huấn luyện trước trên IMDB-WIKI. Bộ dữ liệu này bao gồm 523,051 hình ảnh khuôn mặt của 20,284 người nổi tiếng được thu thập từ các trang web IMDB và Wikipedia. Bộ dữ liệu chủ yếu bao gồm khoảng tuổi [20, 65], và chỉ có rất ít mẫu cho các khoảng tuổi trẻ hơn và cao tuổi hơn. Do đó, bộ phân loại tuổi có thể đưa ra ước lượng tuổi không chính xác hơn đối với khuôn mặt của những người dưới 20 tuổi hoặc lớn hơn 65 tuổi. Do đó, ta chọn sử dụng hình ảnh trong khoảng tuổi Q = {20, . . . , 69} cho quá trình huấn luyện. Truyền các hình ảnh của bộ dữ liệu FFHQ qua bộ phân loại tuổi và nhận thấy FFHQ chứa nhiều mẫu khuôn mặt trẻ hơn so với khuôn mặt già. Sau đó, mở rộng bộ dữ liệu bằng cách thêm các hình ảnh tổng hợp được tạo ra bởi StyleGAN [16] để đạt được phân phối tuổi gần như đồng đều trên khoảng tuổi Q, như được mô tả trong phần 4.1 của bài báo.

C. Kết quả bổ sung

Trong phần này, ta trình bày các kết quả bổ sung trên các hình ảnh có độ phân giải 1024 × 1024.

C.1 Kết quả trên bộ dữ liệu

FFHQ Các kết quả chuyển đổi tuổi trên hình ảnh có độ phân giải 1024 × 1024 của bộ dữ liệu FFHQ được trình bày trong Hình 9 và 10.

C.2 So sánh với các phương pháp khác

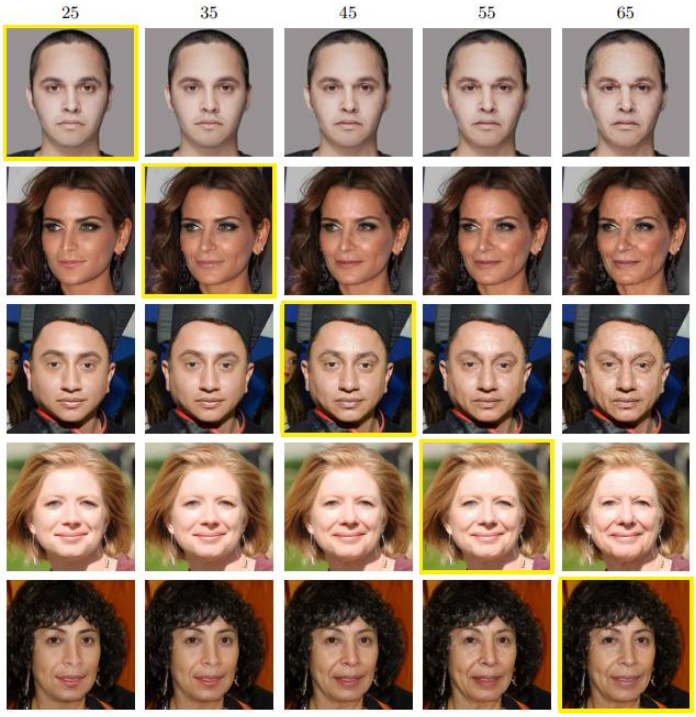
Trong Hình 13, ta trình bày thêm các kết quả so sánh về chuyển đổi tuổi trên bộ dữ liệu Celeba-HQ [15]. Như đã đề cập trong bài báo, Ta so sánh phương pháp với hai phương pháp tiên tiến nhất về chuyển đổi tuổi khuôn mặt có mã nguồn được công bố - PAGGAN [38] và IPCGAN [36]. Ta cũng so sánh kết quả của Ta với kết quả thu được bằng Fader Network [18], cho phép thay đổi nhiều thuộc tính khuôn mặt bao gồm tuổi. Mỗi hình ảnh đầu vào được chuyển đổi thành nhóm tuổi

lớn nhất bằng cách sử dụng các mô hình chính thức được công bố của từng phương pháp. Đối với IPCGAN và PAGGAN, nhóm tuổi lớn nhất tương ứng với 50+ và [51, 60]. Đối với Fader Network, thuộc tính tuổi được đặt thành giá trị mặc định lớn nhất cho tuổi trong mã nguồn chính thức của họ. Để có một cuộc so sánh công bằng với các phương pháp theo nhóm, và vì 50+ được coi là nhóm tuổi cao nhất, Ta chọn một tuổi mục tiêu là 60 (trung bình của khoảng tuổi $\{51, \ldots, 69\} \subset Q$) cho age transformer.

Bảng 2. Siêu tham số của kiến trúc mạng được đề xuất. Kích thước đầu vào là 1024 × 1024 × 3. Đối với age transformer, trừ lớp cuối cùng, mỗi phép tích chập được theo sau bởi một quá trình chuẩn hóa theo mẫu và một hàm kích hoạt LeakyReLU. Đối với bộ phân biệt, trừ lớp đầu tiên và lớp cuối cùng, mỗi phép tích chập được theo sau bởi một quá trình chuẩn hóa batch và một hàm kích hoạt LeakyReLU.

Operation	Kernel size	Stride	Channel
Age transformer			
Encoder			
Convolution	9×9	1	32
Convolution	3×3	2	64
Skip connection 1			
Convolution	3×3	2	128
Skip connection 2			
Residual block	3×3	1	128
Residual block	3×3	1	128
Residual block	3×3	1	128
Residual block	3×3	1	128
Modulation layer			
Decoder			
Concatenation with skip connection 2			
Upsampling			
Convolution	3×3	1	64
Concatenation with skip connection 1			
Upsampling			
Convolution	3×3	1	32
Convolution	9×9	1	3
Discriminator			
Convolution	4×4	2	32
Convolution	4×4	2	64
Convolution	4×4	2	128
Convolution	4×4	2	256
Convolution	4×4	1	512
Convolution	4×4	1	1
Upsampling mode	Nearest (scale factor = 2)		
Padding mode	Reflection		
Normalization	InstanceNorm for age transformer BatchNorm for discriminator		
Activation	LeakyReLU (negative slope = 0.2)		

- **Convolution**: Lớp tích chập là một phép biến đổi dữ liệu đầu vào bằng cách áp dụng một kernel (bộ lọc) lên các vùng nhỏ của dữ liệu để tạo ra các đặc trưng mới. Quá trình tích chập giúp trích xuất thông tin từ dữ liệu và tạo ra các đặc trưng cấp cao.
- **Residual block**: Khối residual là một cấu trúc trong mạng nơ-ron sâu để tạo ra các mạng học sâu dễ huấn luyện và có khả năng học được các đặc trưng phức tạp. Khối residual nhận đầu vào và tạo ra đầu ra bằng cách thực hiện một chuỗi các phép tích chập và kết hợp kết quả với đầu vào ban đầu thông qua kết nối trực tiếp. Điều này cho phép mạng học các biến thể học tương đối nhỏ từ đầu vào và chỉ tập trung vào việc học các sai số hoặc đặc trưng mới.
- **Upsampling**: Upsampling là quá trình tăng kích thước của hình ảnh hoặc dữ liệu bằng cách tạo ra các giá trị trung gian giữa các điểm dữ liệu đã có sẵn. Trong kiến trúc mạng, upsampling được sử dụng để tăng kích thước của đầu ra từ một lớp trước đó để đạt được độ phân giải cao hơn. Phép upsampling có thể được thực hiện bằng các phương pháp như nâng tỷ lệ, chèn giá trị trung bình hoặc chèn giá trị lân cận gần nhất giữa các điểm dữ liệu đã có sẵn.

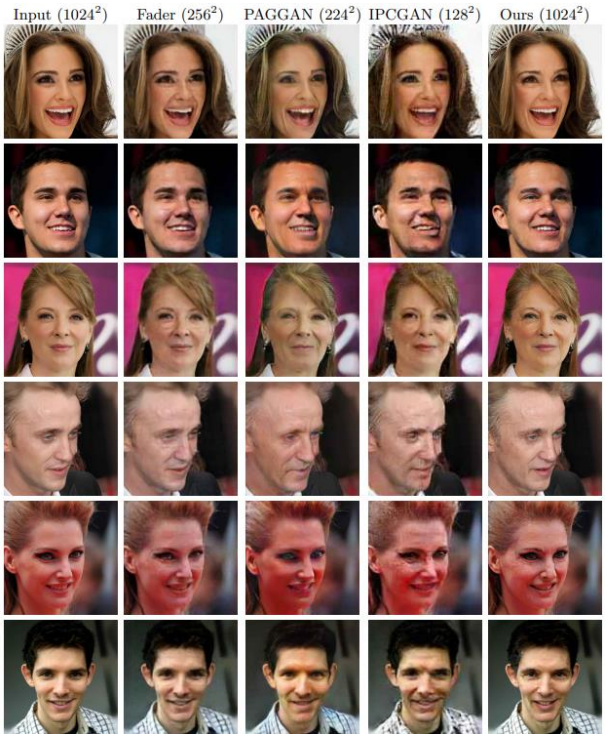


Hình 9. Biến đổi tuổi trên hình ảnh kích thước 1024x1024. Trên mỗi hàng, khung màu vàng chỉ ra hình ảnh gốc. Mỗi cột tương ứng với một tuổi mục tiêu là: 25, 35, 45, 55, 65. Phương pháp của ta mang lại kết quả đáng mừng từ mặt hình ảnh mà không gây ra các hiện tượng nhiễu đáng kể. Chỉ có các đặc trưng liên quan đến tuổi được thay đổi, trong khi danh tính, kiểu tóc, cảm xúc và nền được lưu lại trong quá trình biến đổi. (Xem thêm các hình 10,11,12)

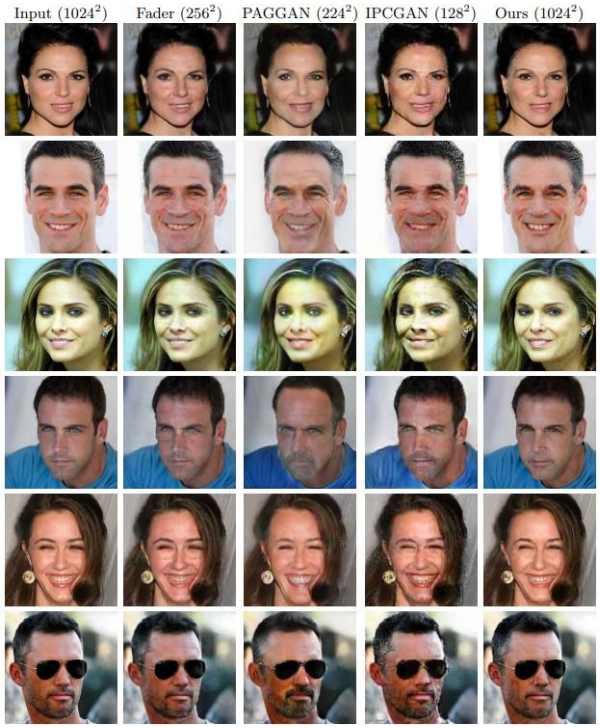
Hình 10

Hình 11

Hình 12



Hình 13



Hình 14

Hình 13, 14: So sánh kết quả biến đổi tuổi trên tập dữ liệu CelebA HQ [15]. Cột đầu tiên là các hình ảnh đầu vào. Cột thứ hai đến thứ năm là kết quả từ Fader Network [18], PAG-GAN [38], IPC-GAN [36] và phương pháp của ta. Kết quả này đạt độ phân giải cao nhất mà không gây ra các hiện tượng nhiễu đáng kể. Phương pháp bảo tồn nền tốt hơn so với các kỹ thuật khác, ví dụ như

các chữ cái trên hàng thứ ba. Ngoài ra, so với các kỹ thuật khác, phương pháp của chúng ta mang lại kết quả không có nhiễu và mờ mờ.